# National Assessment Governing Board
## Assessment Development Committee
### August 4-5, 2016

## AGENDA

### Thursday, August 4

| Noon – 4:00 pm | **Closed Session**  
Review of NAEP Items for Reading, Mathematics, and Writing  
*Committee Discussion* | Secure material provided under separate cover |

### Friday, August 5

| 10:30 – 10:35 am | Welcome, Introductions, and Agenda Overview  
*Shannon Garrison, Chair* |
| 10:35 – 11:00 am | Studies Involving the NAEP Mathematics Framework and State Standards  
*Dan McGrath, NCES* |
| 11:00 – 10:25 am | Update on Technology and Engineering Literacy (TEL) Outreach Activities  
*Shannon Garrison  
Cary Sneider, ADC Vice Chair  
Mary Crovo, Deputy Executive Director*  
Attachment A |
| 11:25 am – 12:15 pm | **Closed Session**  
Update on NAEP Digital-Based Assessments  
*Eunice Greer, NCES*  
Attachment B |
| | Information Item:  
NAEP Item Review Schedule  
Attachment C |
Update on NAEP Technology and Engineering Literacy (TEL)

Outreach Activities

Since the May 17, 2016 release of the NAEP Technology and Engineering Literacy (TEL) results in Detroit, a number of Board members and staff have participated in major events to further disseminate TEL data and related information to a wide range of audiences. Recent conference presentations are listed below.

**U.S. News STEM Solutions National Leadership Conference**  
*May 20, 2016 in Baltimore, Maryland*

*What Testing Technology and Engineering Literacy Means for Education*

The National Assessment of Educational Progress (NAEP)—the country’s largest nationally representative measurement of student achievement—released findings from the first-ever assessment of Technology and Engineering Literacy (TEL), putting NAEP on course to measure all the subjects in STEM. TEL goes beyond testing the ability to “do” engineering or use technology to measuring a myriad of skills today’s 8th graders will need for any career path. A panel will discuss what the implications are for testing and instruction and what the lessons learned from this new assessment can mean for educators, students, and policymakers nationwide.

Moderator:  Tonya Miles, Maryland-National Capital Park and Planning Commission; and National Assessment Governing Board

Panelists:
- Claus von Zastrow, Chief Operating Officer and Director of Research, Change the Equation
- Mary Crovo, Deputy Executive Director, National Assessment Governing Board
- Steve Barbado, Executive Director and CEO, International Technology and Engineering Educators Association
Exchanging Access to Technology and Engineering in K-12

Are American youth getting enough hands-on experience in technology and engineering? Join Change the Equation (CTEq) for an exclusive overview of new data that sheds light on this critical question and many others. CTEq’s analysis of student survey data from the Nation’s Report Card’s first Technology and Engineering Literacy (TEL) assessment will reveal important insights about students’ opportunities to gain skills most Americans will need, whether or not they go into engineering or technology careers.

Do American youth have equal opportunities for exposure to engineering and technology, regardless of race, gender, or income? Are they gaining these opportunities in or out of schools? How can STEM advocates improve opportunity? Hear answers to these questions, discuss strategies for improving technology and engineering literacy, and network with like-minded professionals from education and industry.

Speakers:
- Linda Rosen, CEO, Change the Equation; and Member, National Assessment Governing Board
- Claus Von Zastrow, Chief Operating Officer and Director of Research, Change the Equation

Are You Smarter Than an 8th Grader?
Briefing on the First-Ever Nation’s Report Card for Technology and Engineering Literacy

The National Assessment Governing Board and the Museum of Science, Boston’s National Center for Technological Literacy®, in conjunction with the STEM Education Caucus, invite you to learn more about the first-ever Nation’s Report Card for Technology and Engineering Literacy. At this dynamic briefing, participants learned more about this new assessment and what these results mean for the nation.
Event for the STEM Education Caucus (continued)

Moderator: Bill Bushaw, Executive Director, National Assessment Governing Board

Panelists:

- Rep. Joseph Kennedy III (D-MA)
- Rep. Paul Tonko (D-NY)
- Nate Ball, mechanical engineer, TV Host, entrepreneur, and author
- Peggy Carr, Acting Commissioner, National Center for Education Statistics
- Ioannis Miaoulis, President and Director, Museum of Science, Boston and Founder, National Center for Technological Literacy®
- Tonya Miles, Member, National Assessment Governing Board, and Chief Departmental Administrator, Office of the General Counsel, Maryland-National Capital Park and Planning Commission
- Claus von Zastrow, Chief Operating Officer and Director of Research, Change the Equation
Measuring Technology and Engineering Literacy on the Nation's Report Card

In the spring of 2016, the National Assessment of Educational Progress (NAEP) (also known as The Nation’s Report Card) will release results of the first Technology and Engineering Literacy (TEL) assessment, which was administered to a national sample of grade eight students in 2014. In this session, TEL content and assessment specialists from the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) will share:

- The key elements of the framework, and how the framework was developed
- How the 8th-grade assessment was developed and delivered, including sample items
- How the achievement levels setting process helps shape the interpretation of results
- The results of the 2014 assessment at grade eight, including scale scores, achievement levels, and comparison of student groups

Moderator: Pamela Lottero-Perdue, Towson University (Maryland)

Panelists:
- Sharyn Rosenberg, National Assessment Governing Board
- Cary Sneider, Portland State University and National Assessment Governing Board
- Senay Purzer, Purdue University
- Greg Pearson, National Academy of Engineering
- Laura Bottomley, North Carolina State University

Technology and Engineering Literacy: An Interactive Exploration

Using their own devices, attendees will explore the tasks and findings of the groundbreaking NAEP Technology and Engineering Literacy assessment to enhance instruction and learning.

Panelists:
- Cary Sneider, Portland State University and National Assessment Governing Board
- Shannon Garrison. Solano Avenue Elementary School: Los Angeles, CA and National Assessment Governing Board
Update on Digital-Based Assessment Development

NAEP’s transition from a paper and pencil assessment to a tabled-administered, digital-based assessment (DBA) continues to move forward. Work is progressing in all subject areas with an eye toward DBA operational and pilot administrations in 2017 for mathematics, reading, writing, civics, U.S. history and geography, followed by 2018 for mathematics, reading, science, civics, U.S. history, geography and technology and engineering literacy. However, some of our most recent work has been focused on civics, U.S. history and geography. For these three subjects, at grade 8, the DBA start-up and DBA pilot will be administered in 2017. Following on the heels of 2017, the first operational assessment of DBA civics, U.S. history and geography will take place, along with a paper and pencil bridge study in 2018. In addition to these processes of trans-adapting existing paper and pencil items to DBA and conducting bridging studies to evaluate the strength of the relationship between the old and new assessments, a more innovative line of development that more fully exploits the technology environment is also underway.

This session will provide the Assessment Development Committee with its first look at some of the more innovative Interactive Item Component (IIC) tools and items for the assessment of civics, U.S. history and geography. Many of these assessment elements are constructed around two to three stimuli which will serve as the focus for a set of items. Several tools are featured. In U.S. history, students will use a “select-in-stimulus” tool to identify content they wish to include in their response. In civics, we are building a constrained web-search tool that functions much like Google© but limits the content students see and can select from. And in geography, students’ geographic skills and knowledge will be assessed using Geographic Information Systems (GIS). GIS is one of the leading classroom tools used for teaching geographic skills and technology to students in the United States.*

*Note added by Governing Board staff: The NAEP Geography Framework, first developed for the 1994 geography assessment, addresses the importance of Geographic Information Systems in instruction and assessment. GIS can now be incorporated into the 2018 NAEP geography assessment in the DBA environment!
### Assessment Development Committee

#### Item Review Schedule

**July 2016 – January 2017**

**July 22, 2016**

<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>7/20/16</td>
<td>8/12/16</td>
<td>Cognitive</td>
<td>2017 Reading (4, 8) Operational (DI)</td>
<td>20 (2 blocks)</td>
<td>For review at August Board meeting</td>
</tr>
<tr>
<td>7/20/16</td>
<td>8/12/16</td>
<td>Cognitive</td>
<td>2017 Writing (4) Operational (DI)</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>7/20/16</td>
<td>8/12/16</td>
<td>Cognitive</td>
<td>2017 Writing (8) Operational (DI)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7/20/16</td>
<td>8/12/16</td>
<td>Cognitive</td>
<td>2019 Reading (4) Pilot (SBT)</td>
<td>2 tasks</td>
<td></td>
</tr>
<tr>
<td>7/21/16</td>
<td>8/12/16</td>
<td>Cognitive</td>
<td>2017 Math (4, 8) Operational (DI)</td>
<td>216</td>
<td>✓</td>
</tr>
<tr>
<td>11/9/16</td>
<td>11/29/16</td>
<td>Survey</td>
<td>2019 Math (12) Pilot</td>
<td>20*</td>
<td></td>
</tr>
<tr>
<td>11/9/16</td>
<td>11/29/16</td>
<td>Survey</td>
<td>2019 Reading (12) Pilot</td>
<td>10**</td>
<td></td>
</tr>
<tr>
<td>1/5/17</td>
<td>1/26/17</td>
<td>Cognitive</td>
<td>2019 Reading (12) Pilot (SBT) Draft Builds</td>
<td>2 tasks</td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** “SBT” indicates Scenario-Based Task

“DI” indicates Discrete Item
**Thursday, August 4**

<table>
<thead>
<tr>
<th>Time</th>
<th>Agenda Item</th>
<th>Speaker(s)</th>
<th>Attachments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:00 – 1:05 pm</td>
<td>Welcome, Introductions, and Agenda Overview</td>
<td><em>Andrew Ho, Chair</em></td>
<td></td>
</tr>
<tr>
<td>1:05 – 2:20 pm</td>
<td>Participation and Engagement in NAEP</td>
<td><em>Ariel Jacobs, AnLar Incorporated</em></td>
<td>Attachment A</td>
</tr>
<tr>
<td></td>
<td>• Synthesis of Secondary Research on Motivation and Engagement in NAEP</td>
<td><em>Allison LaFave, AnLar Incorporated</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Indicators from Operational Administration of NAEP</td>
<td><em>Joe Taylor, Abt Associates</em></td>
<td>Attachment B</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Holly Spurlock, NCES</em></td>
<td></td>
</tr>
<tr>
<td>2:20 – 2:30 pm</td>
<td>Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2:30 pm – 3:30 pm</td>
<td>Update on Research on Academic Preparedness for College</td>
<td><em>Andreas Oranje, ETS</em></td>
<td>Attachment C</td>
</tr>
<tr>
<td></td>
<td>• Initial Findings from 2013 NAEP Grade 12 Linking Studies in MA, MI, and TN</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Planned Additional Analyses from 2013 NAEP Grade 8 and 2013 NAEP Grade 12</td>
<td><em>Sharyn Rosenberg, Assistant Director for Psychometrics</em></td>
<td>Attachment D</td>
</tr>
<tr>
<td>3:30 – 4:00 pm</td>
<td>Overview of 2017 Writing Grade 4 Achievement Levels Setting Contract</td>
<td><em>Sharyn Rosenberg, Assistant Director for Psychometrics</em></td>
<td>Attachment E</td>
</tr>
</tbody>
</table>
# AGENDA

**Friday, August 5**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30 – 10:35 am | Welcome, Introductions, and Agenda Overview  
*Andrew Ho, Chair* |
| 10:35 – 11:35 am | Exploring the Use of NAEP as an Indicator of Academic Preparedness for Job Training  
- Overview of Lessons Learned from Previous Job Training Research  
  *Michelle Blair, Senior Research Associate*  
- Next Steps  
  *COSDAM Members*  
  [Attachment F](#) |
| 11:35 – 11:55 am | Follow Up on Uses of NAEP  
*Andrew Ho, Chair* |
| 11:55 am – 12:15 pm | Updates on Various Topics  
*Pat Etienne, NCES*  
*Sharyn Rosenberg, Assistant Director for Psychometrics* |
STUDENT ENGAGEMENT IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP): CRITICAL REVIEW & SYNTHESIS OF RESEARCH

July 2016
# TABLE OF CONTENTS

I. Executive Summary .................................................................................................................. 2

II. Background and Context ......................................................................................................... 6
   - Motivation and NAEP ............................................................................................................ 6
   - The Current Review ............................................................................................................ 8
   - A Theoretical Framework for Motivation ........................................................................... 8

III. Methods .................................................................................................................................. 9
   - Phase 1: Relevance Screening ............................................................................................ 9
   - Phase 2: Methodological Rigor Screening ......................................................................... 10
   - Phase 3: Full Coding ............................................................................................................ 11
   - Phase 4: Comprehensive Critical Analysis .......................................................................... 11
   - Eligible Studies from Phases 1-4 ....................................................................................... 11

IV. Findings .................................................................................................................................. 15
   - Synthesis of Eligible Study Characteristics ....................................................................... 15
   - Synthesis Through Critical Review of Individual Studies .................................................. 15
   - Statistical Methods for the Comprehensive Meta-Analysis ............................................... 19
   - Synthesis Through Comprehensive Meta-Analysis ............................................................. 21

V. Limitations and Threats to Validity in Synthesis ................................................................... 36
   - Process ................................................................................................................................. 36
   - Findings ............................................................................................................................... 36

VI. Discussion ................................................................................................................................ 37
   - Motivation Matters .............................................................................................................. 37
   - Not All Students Are Motivated ......................................................................................... 37
   - Interventions May Increase Students’ Motivation .............................................................. 37

VII. Recommendations .................................................................................................................. 38
   - Revisit “Motivation” Questions on NAEP Contextual Questionnaire ................................ 38
   - Commit to a Strong Set of Motivation-Related Questions .................................................. 38
   - Share NAEP Studies with a Broader Audience ................................................................. 38
   - Encourage Future Analyses of More Recent NAEP Data .................................................. 38
   - Encourage Additional Intervention Studies ....................................................................... 39

VIII. References ............................................................................................................................. 40

IX. Endnotes ................................................................................................................................... 41

Appendix A. Design Document ................................................................................................... 42

Appendix B. Code Book .............................................................................................................. 43

Appendix C. Study Eligibility Status After Phase 2 .................................................................. 52

Appendix D. Systematic Review Table ....................................................................................... 67

Appendix E. Annotated Bibliography and Technical Review ....................................................... 69
I. EXECUTIVE SUMMARY

Background and Context

In September 2015, AnLar Incorporated (Project Team) was awarded a contract to conduct a systematic literature review documented via an annotated bibliography and synthesis summary. The goal of this review was to capture what the field knows about the extent to which sub-optimal engagement and/or test administration may affect students’ performance on the National Assessment of Educational Progress (NAEP).

This report provides a systematic examination of empirical research about students’ motivation for NAEP in grades 4, 8, and 12. It answers three critical questions:

1. To what extent is test-taker motivation related to students’ performance on NAEP?
2. To what extent are students motivated to take NAEP?
3. Can test-taker motivation be influenced by incentives and/or other interventions?

A Theoretical Framework for Motivation. The Expectancy-Value Theory, developed by John William Atkinson and applied to the education field by Jacquelynne Eccles, serves as a theoretical framework for this review. On NAEP, contextual questions asking students to report how “good” they are at a subject attempt to capture the expectancy aspect of motivation, while contextual questions asking students to report the degree to which they “like” a subject or find a subject “useful” attempt to capture the value aspect of motivation. Most of the research discussed in the findings section assesses student expectancy and value separately, while a few studies conflate the two. The Project Team tracked the motivation constructs used throughout the studies. Across eligible studies, the most commonly-used constructs, aside from motivation itself, were “effort,” “self-concept,” “perception,” and “attitude.”

Methods

The Project Team research associates used a four-phase process to identify and analyze the extant literature. All resources were duplicate-coded by two Project Team research associates through Phase 3. Any discrepancies were resolved by the principal researcher. Phase 4 coding was completed by the principal researcher.

Phase 1: Relevance Screening

Resource Selection. The Project Team research associates searched several research databases, including Web of Science, Education Resources Information Center (ERIC), the Institution of Education Sciences (IES), and Teachers College Record for studies about students’ motivation on low-stakes assessments, specifically NAEP, Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme for International Student Assessment (PISA). Other resources were identified through reference harvesting. Additional resources from internal databases of the Governing Board and the National Center for Education Statistics were also provided to the Project Team. Duplicate studies or studies for which there were no available abstracts or full text were eliminated. Ultimately, this process yielded a net total of 1,018 studies.

Eligibility Screening. The Project Team created a Code Book and a corresponding online coding tool to determine each study’s eligibility. After reviewing each study abstract, identifying information and answers to the following three eligibility questions were entered into the online coding tool:

1. Does this resource address student motivation and/or engagement in NAEP?
2. Is this resource an empirical study?
3. Is this study eligible for inclusion based on abstract screening? (To be deemed eligible, studies must have received affirmative responses to questions 1 and 2, addressed a K-12 student sample, and been published in 1990 or later.)

This process yielded 140 resources, which studied a variety of assessments, both NAEP and non-NAEP (e.g., TIMSS, PIRLS, PISA, or other low-stakes assessments). While the non-NAEP assessments are also low-stakes, they rely on different measures of motivation and are administered in several countries and to varied student populations. Moreover, the research questions are specific to motivation on the NAEP. Thus, the Project Team concluded that only NAEP-specific studies would be eligible for inclusion in the subsequent phases of the literature review. As a result, 27 NAEP-specific studies were advanced to Phase 2.

Phase 2: Methodological Rigor Screening. The Project Team evaluated the methodological rigor of the 27 studies that advanced through Phase 1. The Osborne Framework was used to assess observational, psychometric, and descriptive studies. Intervention studies were assessed using What Works Clearinghouse standards. However, the information needed to apply WWC criteria was only available for one of the studies. Thus, the Project Team relaxed its application of WWC standards per the Design Document protocol. Studies that satisfied these methodological rigour requirements (n = 15) were advanced to Phases 3 and 4.

Phase 3: Full Coding. The Project Team coded all eligible studies (n = 15) for additional study information, findings, limitations, and descriptive statistics. A selection of pertinent information for each eligible study was captured in the Systematic Review Table.

Phase 4: Comprehensive Critical Analysis. During Phase 4, the principal researcher provided critical analyses of each study’s methodology, findings, inferences, and methodology and coded for cluster data, outcomes comparisons, and comparison adjustments.

Comprehensive Meta-Analysis (CMA, Version 2.2) software was used to conduct random effects meta-analyses of the correlation and intervention studies. The expression for the sample-size weighted mean of percentages and other observational statistics was developed by the authors specifically for this synthesis report.

Eligible Studies

Of the 15 studies deemed eligible for this literature review, seven are observational, four are descriptive, and four are intervention studies. Data were collected from tests administered during a 17-year span (1990 to 2007); of these, seven used data collected in 1990 or 1992. All assessments analyzed were administered via paper and pencil. These assessments spanned multiple grades (i.e., 4, 8, 12) and academic subjects (e.g., mathematics, science, civics). Several studies assessed the value students place on a subject and their perceived ability in that subject. These studies were included in the literature review because they capture the value and expectancy aspects of motivation. Several eligible studies had methodological or statistical limitations, which were taken into account during the Project Team’s analysis.

Findings

Test-Taker Motivation and Student Performance. A random effects meta-analysis of bivariate correlations between scores on student motivation measures and NAEP achievement scores was conducted across the six eligible observational studies. In studies with multiple motivation questions or grade levels, separate correlations were provided for each (e.g., liking math, thinking math is useful). Correlations ranged from .22 to .50. There was a statistically significant summary correlation of .30, suggesting that, across eligible studies, test-taker motivation is related to NAEP achievement. Disaggregating studies by motivation construct (i.e., expectancy vs. value) revealed that both correlations are statistically significant, but achievement is more strongly associated with expectancy (.38) than with value (.19).
One additional correlational study (Craig, 2013) also reported associations between motivation constructs and achievement. However, the study used a multivariate regression and did not report Pearson $r$; thus, its statistical information could not be converted into a bivariate correlation.

**Student Motivation Levels.** Descriptive statistics on students’ self-reported test effort, expectancy motivation, and value motivation revealed that some students—particularly older students—are less motivated to take NAEP. Among fourth graders, the weighted mean of students reporting that they did not try as hard on NAEP as on other tests is 9 percent. For eighth and twelfth graders, it is 17 and 42 percent, respectively.

Disaggregating motivation questions into expectancy (e.g., “I am good at math”) and value (e.g., ”I like science” and “It is important for me to do well on this test”) constructs provided additional insights:

**Expectancy.** Across grade levels, approximately half of students report feeling that they are good at academics. The weighted mean was 54 percent, with minimal variation across grade levels (55, 55, and 48 percent for fourth, eighth, and twelfth graders, respectively).

**Value.** Across grade levels, approximately half of students report that they value academics. However, there was significant variation across grade levels. Eighty-nine percent of fourth graders reported that it was “important” or “very important” to do well on the test, compared to just 34 percent of twelfth grade students. Across all data addressing the value aspect of motivation, 67 percent of fourth graders saw value in the tested subjects or in doing well on NAEP, compared to 54 percent of eighth grade students and 46 percent of twelfth grade students.

**Intervention Effects on Motivation.** The Project Team conducted random effects meta-analyses of treatment effects in intervention studies. The first meta-analysis examined the effect of interventions on students’ self-reported motivation. The results suggest that interventions (e.g., financial incentives, alternative instructions) do not have a statistically significant impact on students’ self-reported effort (summary effect = .04). However, considered alone, the financial incentives in two studies did yield a statistically significant summary effect of .20. A second meta-analysis examined the effects of interventions on students’ NAEP achievement. This meta-analysis yielded a statistically significant summary effect of .10.

**Limitations**

Both the process and findings of this review are subject to limitations.

**Process.** Although the Project Team was intentional about its systematic review process, it is possible that their search strings failed to capture every eligible study. Moreover, the Project Team was unable to locate an abstract or full text for 21 of the studies found in its initial search. Thus, these studies could not be coded. Finally, while the Project Team deliberately restricted its search to studies from 1990 or later, nearly half ($n = 7$) of the 15 eligible studies relied on test data from 1990 or 1992.

**Findings.** This review is most limited by the motivation questions asked in each study. For example, students were asked to self-report their effort and/or motivation (as measured by various proxies), yielding inherently subjective results. The questions also varied across studies, e.g., asking how important it was to do well on NAEP versus asking about the extent to which students liked or saw usefulness in a particular subject. The Project Team addressed this issue, in part, by using random effects meta-analysis, which does not assume that all correlations or intervention effects have the same true value. Perhaps most importantly, the questions—as currently worded—may not be a reliable proxy for students’ NAEP motivation.
Additionally, results from studies of different NAEP tests were combined with an untested assumption that the relationships between motivation/effort and achievement are consistent across grades and disciplines. To address this, study data were disaggregated by grade level and motivation construct, i.e., expectancy and value. These disaggregated data could then be compared to the aggregated analyses. There were not enough intervention studies to disaggregate by incentive type; thus, the goal of those analyses was simply to show whether incentives of any type can increase achievement and/or student motivation.

Finally, none of the four intervention studies established students’ baseline equivalence. Thus, the authors may have attributed differences in NAEP achievement to differences in intervention-induced motivation, rather than to extant differences in student ability.

Discussion

• Data from the eligible studies suggest that motivation levels are related to NAEP achievement. However, as noted in the “Limitations,” this review and its findings are limited by the motivation questions used in each study, which—as currently worded—may not accurately capture students’ NAEP motivation.

• The meta-analysis of descriptive statistics suggests that test effort is lower among older students (i.e., grades 8 and 12), and one in four students reports trying less hard on NAEP than on other tests. Older students (i.e., grades 8 and 12) are less likely to report confidence in their academic abilities or to place value on NAEP and/or academics. This suggests that incentives and growth mindset interventions (Dweck, 2006) should be introduced early, and that the intensity of these incentives and interventions should increase with students’ age.

• Some interventions may have a modest positive effect on the achievement of student test-takers. However, researchers and practitioners must consider whether their interventions could be plausibly scaled up for use among thousands of students.

Recommendations

In light of their findings, the Project Team compiled several recommendations. Notably, several of these recommendations echo those of the National Commission on NAEP 12th Grade Assessment and Reporting (2004) and the Governing Board’s Ad Hoc Committee on 12th Grade Participation and Motivation (2005).

1. To ensure that students’ answers to “motivation” questions are truly a proxy for their motivation levels, the NAEP Program should adopt more NAEP-specific motivation questions.

2. To ensure that researchers are able to track motivation fluctuations over time, the NAEP Program should commit to using a strong and consistent set of motivation-related questions—new or revised—for the foreseeable future.

3. The NAEP Program should improve dissemination of extant studies on how, if at all, student motivation affects NAEP performance. This could create an impetus for much-needed future studies.

4. The NAEP Program should encourage future studies of student motivation to incorporate more recent test data. The majority of eligible studies in this review relied on NAEP data from the 1990s. This is especially significant given NAEP’s gradual transition to digital-based administration.

5. The NAEP Program should support future intervention studies, particularly those that occur during normal NAEP administrations. Intervention studies provide critical insights into how to mitigate issues of low motivation; yet, the Project Team’s review of the literature yielded just four intervention studies, two of which relied on the same data.
II. BACKGROUND AND CONTEXT

Motivation and NAEP

The National Assessment Governing Board (Governing Board) is tasked with setting policy for the National Assessment of Educational Progress (NAEP), otherwise known as the Nation's Report Card, which informs the public about the academic achievement of elementary and secondary school students in the United States. Since 1969, NAEP has been administered in various subjects, including mathematics, reading, writing, science, geography, U.S. history, civics, economics, the arts, and technology and engineering literacy to students in grades 4, 8, and 12. Results from NAEP enable comparisons of student achievement among states, several large urban districts, public and private schools, and student demographic groups. NAEP results not only enable current comparisons among these groups but also allow for the analysis of trends over time. NAEP does not report results for individual students or schools, so it has lower performance stakes than most state accountability systems. For example, scores for individual students are not produced so participating students do not receive test scores, and teachers are not evaluated based on student results.

In 2003, the Governing Board established the National Commission on NAEP 12th Grade Assessment and Reporting to “review the current purpose, strengths, and weaknesses of 12th grade assessment and reporting by [NAEP] and set forth recommendations to the National Assessment Governing Board” (National Commission, 2004, p. 1). One of the Commission’s recommendations was to study the motivation of twelfth graders taking NAEP. Additional recommendations included the following:

- Developing observable indicators of student engagement in taking NAEP and measuring student engagement against those indicators;
- Evaluating the effectiveness of different incentives for participation; and
- Determining whether low completion rates on open-response questions signal low student motivation.

The Governing Board’s Ad Hoc Committee on 12th Grade Participation and Motivation, which elaborated on the Commission’s recommendations in a 2005 report to the Governing Board (Governing Board Ad Hoc Committee, 2005). In this report, the Ad Hoc Committee advised the Governing Board to recommend research in the following areas:

- Developing and evaluating the efficacy of objective indicators of student engagement; and
- Evaluating the efficacy of various material incentives on participation and student engagement.

In order to implement the recommendations from the National Commission on NAEP 12th Grade Assessment and Reporting and the Ad Hoc Committee on 12th Grade Participation and Motivation, and to further its understanding of student motivation on NAEP, the Governing Board commissioned a paper to guide its decision-making about the 12th grade NAEP. In their 2005 report, Jere Brophy and Carole Ames drew upon three areas of motivational theory: expectancy-value theory, self-determination theory, and goal theory—and applied these constructs to the NAEP assessment. Based on their analysis of these theories, the authors concluded that, because NAEP does not offer students any value for participation, students do not have an incentive to participate. The authors also noted that there may be some drawbacks to student participation, especially for students with histories of low achievement, test anxiety, or stereotype threat.
Brophy and Ames (2005) recommended that the Governing Board drop the twelfth grade assessment from the NAEP Program or incorporate the authors’ suggested principles and strategies. These included the following:

- Creating utility by offering incentives, including financial incentives and training in test-taking skills;
- Appealing to students’ social and civic identities (e.g., emphasizing the opportunity to help the test developers and shape future tests; appealing to students’ identification with peers, school, and community; emphasizing the opportunity for students to show what they know);
- Enhancing the interest value of participation in NAEP;
- Reducing the perceived cost of participation to students (e.g., time, effort, fear of psychological costs);
- Fostering perceptions of self-determination;
- Encouraging mastery rather than performance orientations; and
- Improving testing conditions.

Other researchers have examined the impact of motivation on low-stakes assessments, including well-known international assessments, such as the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Math and Science Study (TIMSS). These studies provide useful context and background for understanding how motivation may affect students’ NAEP performance. However, differences in test content, testing conditions, and context limit their generalizability to NAEP.

For example, Wise and DeMars (2005) conducted an influential meta-analysis of 11 empirical studies to examine participant motivation on various low-stakes assessments. They concluded that, on average, students who are motivated to take an assessment perform more than one-half standard deviation higher than unmotivated students. In light of their findings, the authors offer several recommendations for enhancing student motivation, several of which echo recommendations from Brophy and Ames (2005): raising the stakes, providing incentives, choosing tests that are not too mentally taxing, and making the assessments more intrinsically motivating.

Though influential, this study may have limited implications for NAEP. Just two of the 11 studies included in its meta-analysis addressed NAEP, specifically: Kiplinger and Linn (1993) and O’Neil, Sugrue, and Baker (1995). Moreover, Kiplinger and Linn (1993) was the only study in the meta-analysis that yielded a small effect. In this study, students who answered NAEP mathematics questions embedded in a statewide achievement test performed the same—or only slightly better—than a random sample of students in the same state who were administered the same questions on the 1990 NAEP Trial State Assessment. Wise and DeMars (2005) note that this may be, in part, because the state test had stakes for schools but not for individual students.

More than ten years have passed since these reports and studies; still, there are lingering questions and ongoing debate over students’ motivation on NAEP. A recently published study by the Urban Institute, “Varsity Blues: Are High School Students Being Left Behind?” (2016) examines high school achievement over time using student-level achievement data from NAEP. The study concludes that “stagnant achievement among high school students is a real phenomenon” (p. v) and analyzes four hypotheses as to why this is occurring, including the possibility that scores are affected by “senioritis.” Specifically, the study speculates that today’s twelfth grade students take NAEP tests less seriously than have previous high school students and do not make an effort on the test (Blagg & Chinos, 2016, p. 3). Having analyzed the average proportion of test items skipped and students’ self-reported effort on the twelfth grade NAEP, the authors conclude that “the available evidence provides no reason to believe that effort has declined” and that “more research is needed to better understand possible changes in student effort on low-stakes tests such as the NAEP” (p. 15).
The Current Review

To reconcile conflicting reports and recommendations, in June 2015, the Governing Board sought a third party contractor to conduct a systematic examination of empirical research about students’ motivation for NAEP in grades 4, 8, and 12. The solicitation specifically requested “a comprehensive technical review and critical synthesis of research on student engagement on NAEP to learn the extent to which motivation may play a role in student performance on NAEP” with the ultimate goal of “centraliz[ing] what the field knows about the extent to which sub-optimal engagement may affect student performance on NAEP.” AnLar Incorporated (Project Team) was awarded the contract in September 2015.

The Project Team was then tasked with answering three critical research questions:
1. To what extent is test-taker motivation related to students’ performance on NAEP?
2. To what extent are students motivated to take NAEP?
3. Can test-taker motivation be influenced by incentives and/or other interventions?

To conduct the literature review and synthesis, the Project Team produced the following:
- A Design Document to outline its process for selecting and reviewing research (Appendix A);
- A List of Sources comprising all search results (categorized by their relevance to the three research questions);
- A Systematic Review Table, which enables readers to observe trends or patterns across eligible studies (Appendix D);
- An Annotated Bibliography and Technical Review that summarizes and critiques all eligible studies (Appendix E); and
- This Synthesis Report.

A Theoretical Framework for Motivation

The Expectancy-Value Theory, developed by John William Atkinson and applied to the education field by Jacquelynne Eccles, serves as a theoretical framework for this review. According to this theory, students’ achievement and achievement-related choices are primarily determined by two factors: expectancies for success and values for a subject/task. Expectancies refer to how confident a student is in his or her ability to succeed in a task—for example, believing that he or she is “good” at a certain subject. Task values refer to how important, useful, or enjoyable the individual perceives the task—for example, saying that he or she “likes” a certain subject. Research indicates that expectancies and values interact to predict student engagement, effort, continuing interest, and academic achievement (Wigfield & Eccles, 2000).

The majority of the research discussed in the findings section assesses student expectancy and value separately; however, a few studies conflate the two. Contextual questions asking students to report the degree to which they “like” a subject or find a subject “useful” attempt to capture the value aspect of motivation, while contextual questions asking students to report how “good” they are at a subject attempt to capture the expectancy aspect of motivation.

The Project Team tracked motivation constructs used throughout the studies. Across eligible studies, the most commonly-used constructs, aside from motivation itself, were “effort,” “self-concept,” “perception,” and “attitude.”
III. METHODS

The following section describes the Project Team’s four-phase process for reviewing and synthesizing literature relating to students’ motivation on NAEP: Phase 1—Relevance Screening, Phase 2—Methodological Rigor Screening, Phase 3—Full Coding, and Phase 4—Comprehensive Critical Analysis.

All resources were duplicate-coded by two Project Team research associates through Phase 3. Any discrepancies were resolved by the Project Team’s principal researcher. Phase 4 coding was completed by the principal researcher.

Phase 1: Relevance Screening

Resource selection. The Project Team completed a four-phase literature review on the relationship between students’ test motivation and their NAEP performance. In the first stage of Phase 1, the Project Team identified relevant articles using a select set of keywords linked with Boolean operators. The Team’s primary goal was to identify articles that document students’ motivation to take NAEP and/or the relationship between motivation and NAEP performance, i.e., research questions 1 and 2.

Search strings. The Project Team used the connector “OR” to be inclusive and the connector “AND” to be exclusive. After choosing key search terms, the Team constructed the following search strings: Subject=NAEP AND (motivation OR engagement OR incentive OR grit OR expectancy OR mindset OR perseverance OR value OR academic tenacity OR character strength OR effort OR guessing). When possible, the Project Team filtered the searches to eliminate all studies published before 1990.

This search string was repeated for three other assessments the Project Team deemed analogous to NAEP: TIMSS, PIRLS, and PISA. These assessments were included in the search because, like NAEP, they are low-stakes, meaning that students do not receive individual score results and that test scores do not have any impact on their academic performance; are taken in a traditional test-taking environment, in which students work independently and are allotted a specific time to work on sections of the test; are administered by either pencil and paper or digital-based programs; provide national or international student performance results; and test proficiency on at least language arts or math. The Project Team included these search strings to be as comprehensive as possible during this early phase of coding.

Databases searched. The Project Team conducted preliminary searches using Google Scholar. These searches helped them understand the breadth of available research and informed their decisions about which databases to search going forward. Ultimately, the Project Team searched Web of Science, Education Resources Information Center (ERIC), the Institution of Education Sciences (IES), and Teachers College Record. The Team also identified additional studies by reviewing the reference sections of eligible documents, i.e., reference harvesting. Studies in the grey literature (i.e., unpublished studies such as dissertations and conference papers) were captured within the previously mentioned databases. Later in the Phase 1 coding process, reference materials from internal databases of the Governing Board and the National Center for Education Statistics (NCES) provided the Project Team with additional grey literature search hits for consideration.

Results of database searches (number of resources returned). The Project Team’s initial searches and reference harvesting yielded 1,015 resources. The Team was unable to locate an abstract and/or full text for 21 of the studies; thus, these studies were eliminated. The remaining 994 resources captured from database searches were advanced to Phase 1 coding.
Results of studies from internal NCES searches. NCES searched its internal databases using the same search strings. This yielded 24 resources that had not already been captured through the Project Team’s database searches. These 24 resources were added to the 994 resources captured through the database search, yielding a net total of 1,018 studies for Phase 1 coding.

Eligibility Screening. The Project Team created a Code Book (Appendix B) and a corresponding online coding tool to record key information about each study. This Code Book and online coding tool guided the Project Team research associates as they evaluated the abstract for each search result in order to determine whether the study should advance to Phase 2. At this phase, the research associates were concerned with identifying studies that contained original research and were relevant to the research question, e.g., studies that specifically addressed motivation on NAEP or a similar low-stakes assessment.

Specifically, the research associates entered identifying information for all 1,018 studies and answered three eligibility questions based on the resources’ abstracts. Studies that received an affirmative response to all three screening questions were eligible for advancement to Phase 2:

1. Does this resource address student motivation and/or engagement in NAEP as specified in the performance work statement?
2. Is this resource an empirical study?
3. Is this study eligible for inclusion based on abstract screening? (To check “yes,” the researchers must have responded “yes” to questions 1 and 2. Additionally, the study must have been published in or after 1990 and used a K-12 sample population.)

This process yielded 140 resources, which studied a variety of assessments, both NAEP and non-NAEP (e.g., TIMSS, PIRLS, PISA, or other low-stakes assessments). While the non-NAEP assessments are somewhat analogous to NAEP (e.g., standardized, low-stakes), they rely on different measures of motivation and are administered in several countries and to varied student populations. Moreover, the research questions for this review are specific to motivation on NAEP. Thus, the Project Team concluded that only NAEP-specific studies would be eligible for inclusion in the subsequent phases of the literature review. As a result, 27 NAEP-specific studies were advanced to Phase 2.

Breakdown of Phase 1 Results: Any NAEP-specific source with "yes" responses to all three screening questions (n = 27) advanced to Phase 2: Methodology Screening. Of the 991 studies eliminated after Phase 1, 113 were deemed relevant but not specific to NAEP, and 878 were deemed irrelevant to the research questions.

Phase 2: Methodological Rigor Screening

In Phase 2, the Project Team applied two separate standards: one for observational or descriptive studies (Osborne Framework, 2010) and one for intervention studies (Institute of Education Sciences, What Works Clearinghouse (WWC) Standards, 2014).

Observational and Descriptive Studies. The Osborne Framework includes several evaluative criteria. These include the appropriate treatment of hierarchical (nested) data; sufficient measurement validity for all correlated variables; the testing of statistical assumptions of correlational analyses; the appropriate handling of missing data and outliers; and adjustments to the significance level of statistical tests for multiple comparisons. For each study, the Project Team research associates were asked to determine how many of the 11 framework items applied. Observational studies that did not satisfy at least 50 percent of the applicable criteria were eliminated. For the full list of Osborne Framework criteria considered, see Question 2.5a of the Code Book in Appendix B.
**Intervention Studies.** For intervention studies, the Project Team employed the What Works Clearinghouse (WWC) standards. Under these standards, intervention studies can be eliminated for a variety of reasons—for example, if the combination of overall and differential attrition rates exceed liberal values provided in the relevant WWC protocol or if the baseline effect size could not be determined. However, the information needed to apply WWC Framework criteria was only available for one of the studies. As the research progressed, the Project Team realized that there were a limited number of intervention studies. Thus, the Project Team relaxed its application of WWC standards per the Design Document protocol. For the full list of intervention study elimination coding variables, see Question 2.10a of the Code Book in Appendix B.

A list of the 27 studies and their status after Phase 2 coding are summarized in Appendix C: Study Eligibility Status After Phase 2. Studies were eliminated not only for methodological issues but also for failing to address the research questions. Upon closer review, a few were found to be unempirical.

Ultimately, 15 studies were deemed methodologically rigorous enough to advance to Phases 3: Full Coding and Phase 4: Comprehensive Critical Analysis.

**Phase 3: Full Coding**

The Project Team coded all eligible studies (n = 15) for additional study information, findings, limitations, and descriptive statistics. A selection of pertinent information for each eligible study was captured in the Systematic Review Table (see Appendix D).

**Phase 4: Comprehensive Critical Analysis**

During Phase 4, the principal researcher provided critical analyses of each study’s methodology, findings, and inferences. The lead researcher also coded for cluster data, outcomes comparisons, and comparison adjustments.

The full sample size of the studies that were reviewed through Phases 1-4 are summarized in the table below.

<table>
<thead>
<tr>
<th>Results of Eligibility Screening</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Database Search</td>
<td>1,015</td>
</tr>
<tr>
<td>NCES Search</td>
<td>24</td>
</tr>
<tr>
<td>Total Sources Identified by Initial Searches</td>
<td>1,039</td>
</tr>
<tr>
<td>Studies that could not be located</td>
<td>21</td>
</tr>
<tr>
<td>Phase 1: Relevance</td>
<td>1,018</td>
</tr>
<tr>
<td>Phase 2: Methodology</td>
<td>27</td>
</tr>
<tr>
<td>Phase 3/4: Full Coding and Comprehensive Critical Analysis</td>
<td>15</td>
</tr>
</tbody>
</table>

**Eligible Studies from Phases 1-4**

Phases 1-4 yielded a total of 15 eligible studies. These studies were analyzed and synthesized by the Project Team. For detailed study information and technical analyses, see Appendix D: Systematic Review Table and Appendix E: Annotated Bibliography and Technical Review.
## Eligible Study Details

<table>
<thead>
<tr>
<th>Results of Eligibility Screening</th>
<th>Identifying Information</th>
<th>Descriptive Characteristics</th>
<th>Alignment with Research Question(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results of Eligibility Screening</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Sample Size</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Stokes, L., &amp; Cao, J. (2009). Examination of low motivation in the 12th grade NAEP Secondary Analysis Grant from Institute of Educational Sciences. Southern Methodist University, Dallas, TX.</td>
<td>2009</td>
<td>Technical Report</td>
<td>11,642</td>
</tr>
<tr>
<td>Results of Eligibility Screening</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Sample Size</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>

* The "Motivation Construct" column refers to the primary terminology that the author(s) used to refer to motivation in the study. For example, if the author(s) referred to students’ "motivation level," the study was coded as containing a motivation construct. If the authors referred to students’ "level of effort," the study was coded as containing an effort construct. If a study used multiple terminologies, it was coded as using multiple constructs.

** The "Motivation Construct Categorization" column refers to the motivation framework under which the study’s motivation measures were categorized for purposes of meta-analysis. Measures were categorized as "expectancy," "value," "effort," or a composite of expectancy and effort.

IV. FINDINGS

This section provides findings from the Project Team’s review of the 15 individual studies, as well as their subsequent meta-analysis. It begins with a synthesis of eligible study characteristics followed by a critical review of the individual studies. Next is an overview of the statistical methods used to conduct the comprehensive meta-analysis. The section concludes with an overview of the meta-analysis findings and how they address each of the three research questions.

Synthesis of Eligible Study Characteristics

The 15 eligible studies were compiled into a Systematic Review Table (see Eligible Study Details table or Appendix D), which allows readers to quickly and easily scan and compare information from all eligible studies. The Systematic Review Table reveals the following:

- **Timeframe:** Data were collected from assessments administered during a 17-year span (1990 to 2007); of these, seven used data collected in 1990 or 1992.
- **Subject matter:** Data were collected from assessments on a variety of academic subjects, including science, mathematics, English Language Arts, reading, civics; half (seven) collected data from mathematics assessments.
- **Age of participants:** Data were only collected on fourth grade students in three of the technical reports. Most studies collected data from eighth and/or twelfth grade students, with the majority of studies (ten) collecting some data from twelfth grade students.
- **Administration mode:** All assessments analyzed were administered via paper and pencil.

The Systematic Review Table also reveals the following about the design and methodology of the eligible studies:

- **Sample size:** Sample sizes varied widely, from a small sample of 16 students to a large sample of 160,486. The average sample was 22,709 students, and the median was 6,279 students.
- **Half of the studies (seven) were observational studies; the remaining eight were descriptive studies (four) or intervention studies (four).**
- **The intervention studies were affected by higher rates of attrition, an inability to establish baseline equivalence, or both.**
- **The majority of the studies (ten) used “motivation” as their motivation construct, i.e., they used the term ‘motivation’ throughout their research, specifically, rather than an alternative proxy for motivation. Other motivation constructs used included engagement, effort, self-concept, perception, and attitude.**

Finally, the Systematic Review Table reveals the following about the results of the eligible studies:

- **The correlation or direction of the treatment effect for all observational and intervention studies was positive, with the exception of one study which found no correlation.**
- **All of the reported positive correlations or treatment effects were found to be statistically significant on at least one of the measures, with the exception of the study that found no correlation.**

Synthesis Through Critical Review of Individual Studies

This section summarizes how the 15 eligible studies address each of the three research questions. It also shares information from the Project Team’s critical review of each study, focusing particularly on methodological and statistical limitations. (For a more extensive discussion of limitations, see “Limitations and Threats to Validity in Syntheses.”)
Research Question 1: To What Extent is Test-Taker Motivation Related to Students’ Performance on NAEP?

Of the seven eligible studies that reported bivariate correlations between student motivation and achievement (Byrnes, 2003; Craig, 2013; Kim, 1992; Lee, 2013; Stokes & Cao, 2009; Walberg & Ethington, 1991; and Yepes-Baraya, 1996), all but one (Walberg & Ethington, 1991) found a positive and statistically significant correlation, i.e., greater student motivation tended to result in higher performance on NAEP (and vice versa).

Jakwerth, Stancavage, and Reed (2003) took a slightly different approach to exploring the relationship between motivation and performance, focusing specifically on omission of responses. Their small-scale study (n = 65) gauged student motivation on the 1998 eighth grade NAEP reading and civics assessments through interviews with and observations of students. The authors did not find motivation to be a significant factor in students’ omission of responses. The interview methods used in this study were appropriate given the study’s goals and research questions, and the authors’ conclusions appear to be valid based on their interview excerpts.

General Critiques. These studies were conducted in a variety of subjects with either eighth grade students (Kim, 1992; Lee, 2013; Yepes-Baraya, 1996) or twelfth grade students (Byrnes, 2003; Craig, 2013; Stokes & Cao, 2009; Walberg & Ethington, 1991). Thus, the results of one study may not be generalizable to other grades and subjects. Additionally, the contextual questions were phrased differently across studies and asked about varying aspects of motivation. For example, Craig (2013) and Stokes and Cao (2009) analyzed various contextual questions specific to motivation on the NAEP assessment, while Byrnes (2003), Craig (2013), Kim (1992), Lee (2013), Walberg and Ethington (1991), and Yepes-Baraya (1996) analyzed various contextual questions related more generally to motivation for a given subject. This variation in questions makes it difficult to draw broad conclusions. Moreover, the questions—as currently worded—may not be a reliable proxy for students’ NAEP motivation.

Individual Study Critiques. When reviewing these studies and their conclusions, the Project Team considered the following limitations of their designs and methodologies:

- Byrnes (2003) made limited inferences about motivation on NAEP, although the effect of motivation on performance was just one aspect of this much larger study. Methodologically, the author should have corrected the significance level of statistical tests to account for the increased type I error rate (i.e., multiple comparison correction) and reported p-value thresholds (e.g., p < .001) instead of exact p-values; it is impossible for readers to make the correction themselves. In general, this study was well-conceived and carefully conducted.

- Craig (2013) recognized that the general effort scale’s reliability was too low to trust in the regression analysis and therefore discarded the scale prior to analysis. However, multiple statistical tests were conducted on the same sample within the same outcome domain, and no corrections were made. Specifically, there should have been corrections (e.g., Bonferroni or Benjamini-Hochberg) to the reported p-values of the regression analyses. Despite this shortcoming, the author transparently and systematically reported on all tested hypotheses.

- Kim (1992) utilized a perception scale of limited reliability; the reliability of the two-item scale used was not reported, and the original four-item scale had a reliability coefficient (Cronbach alpha) of 0.63. Furthermore, the study lacked interpretation beyond reporting the statistical significance of the perception-achievement relationship. As the statistical significance tests of these relationships were highly powered (due to a large sample size), it is impossible to know whether the relationships are truly noteworthy.

- Lee (2013) occasionally deemed certain effect sizes “small” or “medium,” using cutoff values that are not necessarily well-established for this unique field of study. Except for this minor critique, the study employed a strong observational design and methodology with an astute de-emphasis of statistical significance and a focus on effect sizes.
• In Stokes and Cao (2009), the study groups were formed on the basis of just one questionnaire item about motivation for taking the NAEP test. Further, because these groups were formed by extant student characteristics and not experimentally, the inferences are more akin to correlation than causation. Overall, however, the design and analyses of this study are logical, rigorous, and sophisticated.

• Walberg and Ethington (1991) dropped likely non-significant predictors from the regression calculation for the motivation-reading achievement relationship, which can result in biased estimates of the remaining factors, place too much emphasis on arbitrary cutoffs for statistical significance, and withhold important information from the field. Furthermore, the authors did not attempt to speculate on the implications of the non-significant motivation-reading achievement relationship.

• Yepes-Baraya's (1996) small sample size (n = 16) precludes readers from generalizing its findings to a broader student population. However, the author was appropriately cautious and transparent about the generalizability and ambiguity of the findings with regard to the relationship between motivation and achievement.

• Jakwerth, Stancavage, and Reed (2003) made their study sample diverse rather than representative, making it impossible to draw statistically significant conclusions about the demographic characteristics of students likely to omit questions.

Research Question 2: To What Extent are Students Motivated to Take NAEP?

Descriptive studies—and descriptive statistics extracted from correlational studies—provided insights into this question. Among the 15 eligible studies, 10 included such statistics: Braun, Kirsch, and Yamamoto (2011), Data Compendium (1993), Jakwerth, Stancavage, and Reed (2003), Kim (1992), Lee (2013), NCES (1991), O'Neil, Sugrue, and Baker (1995), O'Neil et al. (1997), O'Sullivan and Weiss (1999), and Stokes and Cao (2009). Of these, several suggested that test effort and academic confidence is higher among younger students (i.e., grade 4) than among older students (i.e., grades 8 and 12).

Three of these studies reported similar data on fourth, eighth, and twelfth grade student achievement and responses to contextual questions pertaining to motivation, but did not conduct tests of statistical significance for differences in groups: Data Compendium (1993); NCES (1991); and O'Sullivan and Weiss (1999). The Data Compendium (1993) reported responses to questions that were specific to motivation on the NAEP assessment itself, while NCES (1991) reported responses to questions related more generally to motivation on the subject at issue (e.g., whether students “liked” math). O'Sullivan and Weiss (1999) addressed both types of motivation questions.

Though none of these three studies conducted effect sizes or statistical significance tests for differences in groups, the authors of the Data Compendium (1993) conclude that student motivation is not related to achievement, while NCES (1991) and O'Sullivan and Weiss (1999) conclude that motivation and achievement are generally related. O'Sullivan and Weiss (1999) also suggest that motivation’s relationship to achievement differs by age, with older students reporting less NAEP motivation than younger students.

General Critiques. As was true for the correlational studies, these studies’ variation in contextual questions makes it difficult to draw general conclusions. Additionally, while these studies were well-designed and executed, they could have provided more information. For example, it would have been easier to interpret relationships between affective variables and proficiency if these studies had provided effect sizes, correlations, and tests of statistical significance for key relationships or comparisons.
Research Question 3: Can Test-Taker Motivation Be Influenced by Incentives and/or Other Interventions?


Braun, Kirsch, and Yamamoto (2011) conducted a randomized controlled field trial to investigate the effects of monetary incentives on twelfth graders’ performance on a reading assessment closely modeled after the NAEP reading test. The study used a convenience sample of 2,600 students from 59 schools across seven states. Students were either assigned to a control group or one of two incentive interventions: a “fixed” incentive, which offered students $20 at the start of the session or a “contingent” incentive, which offered students $5 in advance and $15 for correct responses to each of two randomly chosen questions for a maximum payout of $35.

The primary study in O’Neil, Sugrue, and Baker (1995) was also a randomized controlled trial. It examined the effects of various reward and instruction treatment conditions on 749 eighth grade students (four treatment conditions) and 719 twelfth grade students (five treatment conditions) from Southern California on two blocks of released items from the 1990 NAEP mathematics test. Students were either assigned to the control group (in which standard NAEP instructions were read) or to one of four interventions: a monetary incentive of $1 for every item answered correctly; ego-involved instructions read at the beginning of the test; task-involved instructions read at the beginning of the test; or a certificate of accomplishment for performing in the top 10 percent of one’s class (grade 12 only). In addition to the test, a self-assessment questionnaire was administered to measure self-reported effort and associated metacognitive variables.

O’Neil et al. (1997) reports on the same study as O’Neil, Sugrue, and Baker (1995), but with some additional information on the self-reported effort of eighth grade treatment groups.

Two of these studies (Braun, Kirsch, & Yamamoto, 2011; O’Neil, Sugrue, & Baker, 1995) tested whether various incentives would have an effect on students’ self-reported test effort. Braun, Kirsch, and Yamamoto (2011) focused solely on financial incentives, while O’Neil, Sugrue, and Baker (1995) tested instructional incentives, a non-financial award incentive, and a financial incentive. These studies yielded a statistically insignificant summary effect of .04. However, in both studies, considering only the financial incentives did yield a statistically significant summary effect of .20.

Three of these studies (Braun, Kirsch, & Yamamoto, 2011; O’Neil, Sugre, & Baker, 1995; O’Neil et al., 1997) tested whether these same incentives influenced students’ achievement. None of these studies was conducted on an actual NAEP administration, but rather on simulated NAEP administrations. All three studies found that interventions, to some extent, improved achievement in certain circumstances. The summary effect is modest, though statistically significant (.10).

A fourth study (Kiplinger & Linn, 1993) did not directly measure student motivation, but rather, compared eighth grade student achievement on the same NAEP questions under two different testing conditions—one that was presumed to be “high-stakes” and one that was presumed to be “low-stakes.” Two subsets of NAEP Block 7 mathematics items were embedded in the 1992 Georgia Curriculum-Based Assessments (CBA)—the “high-stakes” environment. The responses to these items were compared to students’ responses to the same questions on Georgia’s 1990 NAEP Trial State Assessment (TSA)—the “low-stakes” environment. The mean scores of the first subset of NAEP items were significantly higher in the 1992 CBA administration than in the 1990 TSA administration, while the CBA and TSA mean scores were not significantly different for the second subset of NAEP items.
**Individual Study Critiques.** Because these studies have varied research designs, it is difficult to make generalizations about their shared limitations. However, during their critical review, the Project Team identified a number of limitations and/or threats to validity in each individual study.

- The Project Team identified three potential threats to validity in Braun, Kirsch, and Yamamoto (2011). First, its high level of attrition increases the likelihood that the group of students who were actually tested differs from the original group of students who were randomly assigned. Similarly, the authors do not demonstrate that the groups were baseline equivalent on reading-related outcomes prior to the interventions. Finally, the authors ignored clustering in analysis (i.e., students within schools), which could have led them to underestimate standard errors for statistical significance tests and, in turn, underestimate the likelihood that a Type I error had been made.

- O’Neil, Sugre, and Baker (1995) and O’Neil et al. (1997) also have several limitations. Because only statistically significant effects were reported, there is no way to know whether sample attrition could have biased the treatment effects, and—due to the small sample size—some non-significant differences may have been large enough to be noteworthy. Additionally, the authors did not interpret the magnitude or importance of the treatment effects; use baseline measures to adjust treatment effects for extant differences in mathematics achievement; or acknowledge that students’ achievement data were nested within schools.

- Though Kiplinger and Linn (1993) employed a clever design, their study design hinges on the debatable assumption that students will be motivated to try on state tests that have stakes for schools and teachers but not for them. Additionally, while a worthy endeavor, the comparison of NAEP scores between the 1990 NAEP administration and the 1992 state test administration with embedded NAEP items is confounded by other factors: potential differences in student populations across those years, differences in test difficulty and duration, differences in study context, differences in timing of the tests (e.g., the 1990 test was administered in February while the 1992 was administered in May, allowing students an additional two to three months of instruction), and placement of questions near the end of the test (questions might not receive full energy and effort of students).

**Statistical Methods for the Comprehensive Meta-Analysis**

**Authority for Statistical Approach.** The Comprehensive Meta-Analysis (CMA, Version 2.2) software used to calculate the random effects meta-analysis of correlations and intervention effects followed the statistical approach suggested by Borenstein, Hedges, Higgins, and Rothstein (2009). Their recommendations have been adopted widely by synthesis researchers, with over 4,300 citations in the literature. The expression for the sample-size weighted mean of percentages and other descriptive statistics was developed by the Project Team specifically for this synthesis report.

**Meta-Analysis of Correlations.** When possible, Pearson’s $r$ correlations were extracted from eligible studies that used observational designs. All correlations were converted to the Fisher’s $z$ scale, and all analyses were performed using the transformed values. The summary $z$-scale correlation and its confidence interval were then converted back to Pearson’s $r$ for presentation. The transformation from Pearson’s $r$ to Fisher’s $z$ was performed using the expression below.

\[ z = 0.5 \times \ln \left( \frac{1 + r}{1 - r} \right) \]
As per the recommendations of Borenstein et al. (2009, p. 42), the Project Team used the variance of $z$ as an approximation of $z$. The variance of $z$ was computed as below.

$$V_z = \frac{1}{n - 3}$$

The Fisher’s $z$ score and its variance were used to compute the summary correlation and its confidence limits, then each of these was converted back into correlation units using the expression below.

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

**Meta-Analysis of Intervention Effects.** When possible, standardized mean difference-type effect sizes were extracted from all eligible intervention studies. When not provided directly by study authors, the standardized mean difference ($d$) was estimated using the expression below, where $X_1$ and $X_2$ are the sample means in the treatment and comparison groups and $S_{\text{within}}$ is the within-groups standard deviation, pooled across groups.

$$d = \frac{X_1 - X_2}{S_{\text{within}}}$$

The variance of $d$ was calculated using the expression below where $n_1$ and $n_2$ are the treatment and comparison group sample sizes, respectively.

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

**Estimating the Summary Correlation/Effect Size and Confidence Interval Using a Random Effects Model.** Using CMA software, each correlation/effect size was weighted by the inverse of its variance. In the random effects model, this variance included both the within-study and between-study variance. In the random-effects model, the weight $W^*$, assigned to each correlation or treatment effect was computed as below, where $V^*_i$ is the within-study variance for the $i$-th study, plus the between-study variance.

$$W^*_i = \frac{1}{V^*_i}$$

The summary correlation/effect size, $M^*$, was then computed as below (the sum of the products of each correlation or treatment effect multiplied by its weight divided by the sum of the weights).

$$M^* = \frac{\sum_{i=1}^{k} W^*_i Y_i}{\sum_{i=1}^{k} W^*_i}$$
The variance and standard error of the summary correlation/effect size was calculated as the reciprocal of the sum of the weights, as below.

\[ V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*} \]

\[ SE_{M^*} = \sqrt{V_{M^*}} \]

The 95 percent lower and upper confidence limits for the summary correlation/effect size was calculated using the expression below.

\[ UL_{M^*} = M^* + 1.96 \times SE_{M^*} \]

\[ LL_{M^*} = M^* - 1.96 \times SE_{M^*} \]

The statistical significance test for the summary correlation/effect size tested the significance of a z statistic, where:

\[ Z^* = \frac{M^*}{SE(M^*)} \]

with the two-tailed probability (\( p^* \)) of Type I error:

\[ p^* = 2[1 - \Phi(1 Z^*)] \]

Note: \( \Phi \) is the standard normal cumulative distribution function.

**Synthesis of Descriptive Statistics.** The majority of descriptive statistics extracted from studies were percentages of students that agreed with various statements related to their effort and/or motivation when taking NAEP or for a subject area more broadly. When possible, these percentages were synthesized using a sample size-weighted mean (\( M \)) of these percentages, using the expression below. In this expression, \( p_i \) is the percentage of students from study \( i \) who responded in a certain way to an effort or motivation survey item, and \( n_i \) is the study sample size.

\[ M = \frac{\sum_{j=1}^{k} p_i n_i}{\sum_{j=1}^{k} n_i} \]

**Synthesis Through Comprehensive Meta-Analysis**

The Project Team’s findings provide answers to the three critical research questions:

1. To what extent is test-taker motivation related to students’ performance on NAEP?
2. To what extent are students motivated to take NAEP?
3. Can test-taker motivation be influenced by incentives and/or other interventions?
The Project Team answered these research questions by synthesizing data captured from the 15 eligible studies during the Comprehensive Critical Analysis (Phase 4).

Each question was answered through a different type of analysis. For example, correlational analysis was used to determine the relationship between students’ motivation and their performance on NAEP. Descriptive analysis was employed to answer the question, “To what extent are students motivated to take NAEP?” Finally, intervention studies were analyzed to determine whether interventions affected students’ achievement results or self-reported test effort.

**Question 1: To What Extent Is Test-Taker Motivation Related to Students’ Performance on NAEP?**

This question was answered through random effects meta-analyses of bivariate correlations between scores on student motivation measures (as measured by various motivation proxies) and NAEP achievement scores.

First, the Project Team analyzed the correlation between motivation and achievement across six relevant correlational studies: Kim (1992), Yepes-Baraya (1996), Lee (2013), Byrnes (2003), Walberg and Ethington (1991), and O’Neil et al. (1997).

In studies with multiple motivation questions or grade levels, separate correlations were provided for each (e.g., liking math, thinking math is useful). As illustrated in Figure 1, the strongest correlations were found in Yepes-Baraya (8th grade, perceived block difficulty = .50), O’Neil et al. (12th grade, general effort = .48; 8th grade, general effort = .42), and Lee (8th grade, “good at writing” = .41; 8th grade, “like to write” = .37). Lower correlations were found in Kim (8th grade, perception of math = .22) and Lee (8th grade, “writing is my favorite activity” = .24).

The Project Team’s meta-analysis yielded a statistically significant summary correlation of 0.30. This relationship is noteworthy, as it is comparable to other policy-relevant correlations in the literature, e.g., a summary correlation of .27 (p < .001) for socioeconomic status and NAEP achievement (Sirin, 2005). While this would seem to suggest that motivation is related to NAEP achievement, other confounding variables may exist.

Note that in Figures 1, 2, and 3 the area of each rectangle corresponds to the weight of each study in the synthesis, the width of the horizontal line passing through each rectangle represents each study’s 95 percent confidence interval, and the summary correlation is depicted by a diamond.

**Figure 1: Correlations Between Motivation (Expectancy and Value Constructs) and Achievement**
Disaggregating studies by motivation construct (i.e., expectancy vs. value) revealed that both summary correlations are statistically significant, but achievement is more strongly associated with expectancy (.38) than with value (.19).

As illustrated in Figure 2, the strongest correlations between expectancy and achievement were found in Yepes-Baraya (8th grade, perceived ability = .50), O’Neil et al. (12th grade, general effort = .48; 8th grade, general effort = .42), and Lee (8th grade, “good at writing” = .41). Weaker correlations were found in Kim (8th grade, perception of math = .22), Yepes-Baraya (8th grade, perceived block difficulty = .28), and Byrnes (12th grade, ability/liking math = .37).

Figure 2: Correlations Between Motivation (Expectancy Construct) and Achievement

<table>
<thead>
<tr>
<th>STUDY</th>
<th>Correlation</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>z-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim (1992) – Perception of math</td>
<td>0.22</td>
<td>0.21</td>
<td>0.23</td>
<td>68.64</td>
<td>0.00</td>
</tr>
<tr>
<td>Yepes-Baraya (1996) – Perceived ability</td>
<td>0.50</td>
<td>0.01</td>
<td>0.80</td>
<td>1.97</td>
<td>0.06</td>
</tr>
<tr>
<td>Yepes-Baraya (1996) – Perceived block difficulty</td>
<td>0.28</td>
<td>-0.25</td>
<td>0.68</td>
<td>1.04</td>
<td>0.30</td>
</tr>
<tr>
<td>Lee (2013) – Good at writing</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>174.51</td>
<td>0.00</td>
</tr>
<tr>
<td>Byrnes (2003) – Ability/liking math</td>
<td>0.37</td>
<td>0.35</td>
<td>0.39</td>
<td>37.85</td>
<td>0.00</td>
</tr>
<tr>
<td>O’Neil et al. (1997) – Effort (8th)</td>
<td>0.42</td>
<td>0.36</td>
<td>0.48</td>
<td>11.25</td>
<td>0.00</td>
</tr>
<tr>
<td>O’Neil et al. (1997) – Effort (12th)</td>
<td>0.48</td>
<td>0.42</td>
<td>0.54</td>
<td>13.51</td>
<td>0.00</td>
</tr>
<tr>
<td>Summary Correlation</td>
<td>0.38</td>
<td>0.28</td>
<td>0.48</td>
<td>6.54</td>
<td>0.00</td>
</tr>
</tbody>
</table>

By comparison, correlations between value and achievement were quite weak. For example, Walberg and Ethington (12th grade, math utility/relevance) found zero correlation, while Byrnes (12th grade, utility/relevance of math) reported a correlation of just .03. Kim (8th grade, perception of math = .22) and Lee (8th grade, “like to write” = .37; 8th grade, “writing is my favorite activity” = .24) reported moderately higher correlations.

Figure 3: Correlations Between Motivation (Value Construct) and Achievement

<table>
<thead>
<tr>
<th>STUDY</th>
<th>Correlation</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>z-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim (1992) – Perception of math</td>
<td>0.22</td>
<td>0.21</td>
<td>0.23</td>
<td>68.64</td>
<td>0.00</td>
</tr>
<tr>
<td>Lee (2013) – Like to write</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>155.60</td>
<td>0.00</td>
</tr>
<tr>
<td>Lee (2013) – Writing is favorite</td>
<td>0.24</td>
<td>0.24</td>
<td>0.25</td>
<td>98.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Byrnes (2003) – Math utility/el.</td>
<td>0.03</td>
<td>0.01</td>
<td>0.05</td>
<td>2.92</td>
<td>0.06</td>
</tr>
<tr>
<td>Walberg &amp; ... (1991) – Math utility/impt.</td>
<td>0.00</td>
<td>-0.12</td>
<td>0.12</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Summary Correlation</td>
<td>0.19</td>
<td>0.10</td>
<td>0.27</td>
<td>3.98</td>
<td>0.00</td>
</tr>
</tbody>
</table>

One additional correlational study (Craig, 2013) also reported associations between motivation constructs and twelfth grade NAEP achievement. However, the study used a multivariate regression and did not report Pearson’s r; thus, its statistical information could not be converted into a bivariate correlation. That said, Craig (2013) did observe positive associations between self-concept and achievement ($r_{concept} = 5.56, p < .001$). Test effort was also found to be related to achievement. Indeed, students who reported exerting effort were likely to score
seven science scale score points lower than the mean science score ($\beta_{effort} = -7.15$, $p < .001$). These findings further corroborate the synthesis data in Figures 1, 2, and 3, i.e., motivation-achievement = .30, expectancy-achievement = .38, and value-achievement = .19.

**Question 2: To What Extent Are Students Motivated to Take NAEP?**

To address this question, the Project Team compiled descriptive statistics on students' self-reported test effort, expectancy, and value.

Because this project's scope was limited to secondary research, the Project Team only examined data from eligible studies. However, these eligible studies represent a small subset of the grades, academic subjects, and years in which the NAEP was administered. Full data sets are available on the NAEP Data Explorer.

The tables below provide overall and grade-level results. Data from eligible studies whose metrics could not be combined with other descriptive statistics are also included (see "Corroborating Evidence" in Figures 5, 8, and 11). Whereas most of the eligible studies reported the percentage of students who reported a certain amount of effort, these corroborating studies provided means, which could not be converted into percentages.

Figure 4 lists studies that included data on the percentage of students who reported not trying as hard on NAEP as on other tests. It identifies a weighted mean of 25 percent. However, since the data have been disaggregated by grade level, it is easy to recognize significant variation among fourth, eighth, and twelfth grade students. Just 9 to 10 percent of fourth graders reported trying less hard on NAEP than on other tests, compared to 16 to 20 percent of eighth graders and 29 to 49 percent of twelfth graders.

**Figure 4: Descriptive Statistics on Student Motivation (“Effort”)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Percentage of Students indicating that they did not try as hard on NAEP as on other tests</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braun, Kirsch, and Yamamoto (2011)</td>
<td>29%</td>
<td>12</td>
<td>2,612</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>10%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>20%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>45%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>9%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>16%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>39%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>Stokes and Cao (2009)</td>
<td>49%</td>
<td>12</td>
<td>11,642</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5 presents data on eighth and twelfth grade students’ responses to questions about their effort on the NAEP test, based on data from O’Neil, Sugrue, and Baker (1995) and O’Neil et al. (1997). Students were asked to provide their effort levels using five, four-point scale statements (1 = lowest effort; 4 = highest effort). In both studies, eighth grade students reported higher effort levels (mean = 3.41 of 4) than twelfth grade students (mean = 15.10 of 20).

**Figure 5: Corroborating Descriptive Statistics on Student Motivation (“Effort”)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Average Student Rating of Effort on NAEP</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>O’Neil, Sugrue, and Baker (1995)</td>
<td>85% of max score (mean = 3.41/4)</td>
<td>8</td>
<td>749</td>
</tr>
<tr>
<td>O’Neil et al. (1997)</td>
<td>85% of max score (mean = 3.41/4)</td>
<td>8</td>
<td>749</td>
</tr>
<tr>
<td>O’Neil et al. (1997)</td>
<td>76% of max score (mean = 15.10/20)</td>
<td>12</td>
<td>719</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td>81%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 provides weighted mean percentages of fourth, eighth, and twelfth grade students reporting that they did not try as hard on NAEP as on other tests. Among fourth graders, this weighted mean percentage is 9 percent. For eighth and twelfth graders, it is 17 and 42 percent, respectively.

**Figure 6: Descriptive Statistics on Student Motivation (“Effort”) Disaggregated by Grade Level**

<table>
<thead>
<tr>
<th>Study</th>
<th>Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>10%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>9%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td>9%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>20%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>16%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td>17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests</td>
<td>Grade</td>
<td>Sample Size</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>Braun, Kirsch, and Yamamoto (2011)</td>
<td>29%</td>
<td>12</td>
<td>2,612</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>45%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>39%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>Stokes and Cao (2009)</td>
<td>49%</td>
<td>12</td>
<td>11,642</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td></td>
<td>42%</td>
</tr>
</tbody>
</table>

Figure 7 presents descriptive data on the percentage of students who agree with various expectancy construct statements. The percentages, which range from 39 (O’Sullivan, 1999) to 67 (Kim, 1992), have a weighted mean of 54.

**Figure 7: Descriptive Statistics on Student Motivation (Expectancy Construct)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Expectancy Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree with “I am good at mathematics” (1992)</td>
<td>65%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree with “I am good at mathematics” (1990)</td>
<td>64%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1992)</td>
<td>60%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1990)</td>
<td>62%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1992)</td>
<td>51%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1990)</td>
<td>58%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Kim (1992)**</td>
<td>Composite of students who agree with “I like mathematics” and “I am good at mathematics”</td>
<td>67%</td>
<td>8</td>
<td>3,058</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with “I am good at writing”</td>
<td>51%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree with “I am good at mathematics” (1990)</td>
<td>62%</td>
<td>4</td>
<td>8,902</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1990)</td>
<td>63%</td>
<td>8</td>
<td>8,888</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1990 TSA)</td>
<td>62%</td>
<td>8</td>
<td>94,979</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with “I am good at mathematics” (1990)</td>
<td>57%</td>
<td>12</td>
<td>8,862</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with “I am good at science”</td>
<td>45%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with “I am good at science”</td>
<td>47%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with “I am good at science”</td>
<td>39%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td></td>
<td>54%</td>
<td></td>
</tr>
</tbody>
</table>

* For studies that disaggregated expectancy statement data by assessment year, the assessment year has been provided.

** Kim (1992) used a composite statistic for both expectancy and value.
Figure 8 presents expectancy motivation data from O'Neil et al. (1997). Rather than ask students how much they agree with statements like “I am good at mathematics,” O'Neil et al. (1997) asked students to provide a five-point scale response to the statement “Compared to your classmates, your math ability is…” with 1 meaning much lower than most classmates and 5 meaning much higher. Eighth grade students were slightly more likely than twelfth graders to report that their math ability exceeded that of their peers (68 percent of maximum score versus 64 percent; mean = 3.40 versus 3.20). The weighted mean of eighth and twelfth graders was 66 percent of the maximum score.

**Figure 8: Corroborating Descriptive Statistics on Student Motivation (Evidence for Expectancy Construct)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Expectancy Statements</th>
<th>Average Student Rating of Math Ability Relative to Classmates’ Math Ability</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>O'Neil et al. (1997)</td>
<td>“Compared to your classmates, your math ability is…” (1 = much lower than most of my classmates; 5 = much higher than most of my classmates)</td>
<td>64% of max score (mean = 3.40 /5)</td>
<td>12</td>
<td>670</td>
</tr>
<tr>
<td>O'Neil et al. (1997)</td>
<td>“Compared to your classmates, your math ability is…” (1 = much lower than most of my classmates; 5 = much higher than most of my classmates)</td>
<td>68% of max score (mean = 3.40 /5)</td>
<td>8</td>
<td>634</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>66%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Disaggregating Figure 7 data by grade level (4, 8, and 12) reveals minimal variations (see Figure 9). For example, the weighted mean of fourth grade students who reported feeling that they are good at science or math was 55 percent; among eighth graders, the weighted mean of students who reported feeling that they are good at science, math, or writing was also 55 percent. By contrast, the weighted mean for twelfth graders reporting that they are good at science or math was slightly lower: 48 percent.

**Figure 9: Descriptive Statistics on Student Motivation (Expectancy Construct) Disaggregated by Grade Level**

<table>
<thead>
<tr>
<th>Study</th>
<th>Expectancy Statements*</th>
<th>Percentage of Students Who Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree with “I am good at mathematics” (1992)</td>
<td>65%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree with “I am good at mathematics” (1990)</td>
<td>64%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree with “I am good at mathematics” (1990)</td>
<td>62%</td>
<td>4</td>
<td>8,902</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with “I am good at science”</td>
<td>45%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td><strong>55%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For studies that disaggregated expectancy statement data by assessment year, the assessment year has been provided.*
<table>
<thead>
<tr>
<th>Study</th>
<th>Expectancy Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I am good at mathematics&quot; (1992)</td>
<td>60%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I am good at mathematics&quot; (1990)</td>
<td>62%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Kim (1992)</td>
<td>Composite of students who agree with &quot;I like math&quot; and &quot;I am good at math&quot;</td>
<td>67%</td>
<td>8</td>
<td>3,058</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with &quot;I am good at writing&quot;</td>
<td>51%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>63%</td>
<td>8</td>
<td>8,888</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990 TSA)</td>
<td>62%</td>
<td>8</td>
<td>94,979</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I am good at science&quot;</td>
<td>47%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td><strong>Weighted Mean of Percentage</strong></td>
<td></td>
<td><strong>55%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For studies that disaggregated expectancy statement data by assessment year, the assessment year has been provided.*

<table>
<thead>
<tr>
<th>Study</th>
<th>Expectancy Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I am good at mathematics&quot; (1992)</td>
<td>51%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I am good at mathematics&quot; (1990)</td>
<td>58%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I am good at mathematics&quot; (1990)</td>
<td>57%</td>
<td>12</td>
<td>8,862</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I am good at science&quot;</td>
<td>39%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td><strong>Weighted Mean of Percentage</strong></td>
<td></td>
<td><strong>48%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For studies that disaggregated expectancy statement data by assessment year, the assessment year has been provided.*
Figure 10 captures the percentage of students who either enjoyed or saw value in particular subjects (e.g., mathematics, science) or in doing well on NAEP. These percentages varied greatly, from 34 to 89. Interestingly, both the highest and lowest percentages were associated with questions about the value students placed on doing well on NAEP. According to the Data Compendium (1993), 89 percent of fourth graders reported that it was “important” or “very important” to do well on the test. By contrast, both the Data Compendium (1993) and O’Sullivan and Weiss (1999) found that just 34 percent of twelfth grade students thought it was “important” or “very important” to do well on the test.

**Figure 10: Descriptive Statistics on Student Motivation (Value Construct)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Value Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braun, Kirsch, and Yamamoto (2011)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>44%</td>
<td>12</td>
<td>2,612</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>89%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1992)</td>
<td>57%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>57%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>60%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1992)</td>
<td>51%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>54%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>34%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Jakwerth, Stancavage, and Reed (2003)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>73%</td>
<td>8</td>
<td>84</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with &quot;I like to write&quot;</td>
<td>52%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with &quot;Writing helps share ideas&quot;</td>
<td>61%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1990)</td>
<td>67%</td>
<td>4</td>
<td>8,902</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1990)</td>
<td>67%</td>
<td>4</td>
<td>8,902</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>56%</td>
<td>8</td>
<td>8,888</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990 TSA)</td>
<td>57%</td>
<td>8</td>
<td>94,979</td>
</tr>
<tr>
<td>Study</td>
<td>Value Statements*</td>
<td>Percentage of Students Who Agree or Strongly Agree with Statements</td>
<td>Grade</td>
<td>Sample Size</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>54%</td>
<td>12</td>
<td>8,862</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like science&quot;</td>
<td>67%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students indicating that is was important or very important to do well on the test</td>
<td>85%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1992)</td>
<td>71%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1990)</td>
<td>70%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>35%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>40%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree or strongly agree with &quot;I like science&quot;</td>
<td>50%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>58%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>50%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like science&quot;</td>
<td>52%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>O’Sullivan and Weiss (1999)</td>
<td>Students indicating that is was important or very important to do well on the test</td>
<td>34%</td>
<td>12</td>
<td>22,116</td>
</tr>
</tbody>
</table>

Weighted Mean of Percentage 56%

*For studies that disaggregated value statement data by assessment year, the assessment year has been provided.

Figure 11 provides eighth grade students’ composite ratings (1 = low attraction; 2 = neutral or moderate attraction; 3 = high attraction) on their attraction to the assessment subject, to science, and to science as a possible career/occupation. The reported mean is 1.94 (65 percent of the maximum score).

**Figure 11: Corroborating Descriptive Statistics on Student Motivation (Value Construct)**
Finally, Figure 12 disaggregates all value construct data by grade level and presents weighted mean percentages. Again, there was notable variation across grade levels. Sixty-seven percent of fourth grade students saw value in tested subjects or in doing well on NAEP, compared to 54 percent of eighth grade students and 46 percent of twelfth grade students.

**Figure 12: Descriptive Statistics on Student Motivation (Value Construct) Disaggregated by Grade Level**

<table>
<thead>
<tr>
<th>Study</th>
<th>Value Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>89%</td>
<td>4</td>
<td>8,738</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1990)</td>
<td>67%</td>
<td>4</td>
<td>8,902</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like science&quot;</td>
<td>67%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>85%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1992)</td>
<td>71%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like mathematics&quot; (1990)</td>
<td>70%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>35%</td>
<td>4</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td>67%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* For studies that disaggregated value statement data by assessment year, the assessment year has been provided.
<table>
<thead>
<tr>
<th>Study</th>
<th>Value Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1992)</td>
<td>57%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>57%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>60%</td>
<td>8</td>
<td>9,432</td>
</tr>
<tr>
<td>Jakwerth, Stancavage, and Reed (2003)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>73%</td>
<td>8</td>
<td>84</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with &quot;I like to write&quot;</td>
<td>52%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>Lee (2013)</td>
<td>Students who agree or strongly agree with &quot;Writing helps share ideas&quot;</td>
<td>61%</td>
<td>8</td>
<td>160,486</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>56%</td>
<td>8</td>
<td>8,888</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990 TSA)</td>
<td>57%</td>
<td>8</td>
<td>94,979</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>40%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like science&quot;</td>
<td>50%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>58%</td>
<td>8</td>
<td>22,116</td>
</tr>
<tr>
<td><strong>Weighted Mean of Percentage</strong></td>
<td></td>
<td><strong>54%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For studies that disaggregated value statement data by assessment year, the assessment year has been provided.*
<table>
<thead>
<tr>
<th>Study</th>
<th>Value Statements*</th>
<th>Percentage of Students Who Agree or Strongly Agree with Statements</th>
<th>Grade</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braun, Kirsch, and Yamamoto (2011)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>44%</td>
<td>12</td>
<td>2,612</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1992)</td>
<td>51%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>54%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>Data Compendium (1993)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>34%</td>
<td>12</td>
<td>8,499</td>
</tr>
<tr>
<td>National Center for Education Statistics (1991)</td>
<td>Students who agree or strongly agree with &quot;I like mathematics&quot; (1990)</td>
<td>54%</td>
<td>12</td>
<td>8,862</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;Science is useful for solving everyday problems&quot;</td>
<td>50%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students who agree with &quot;I like science&quot;</td>
<td>52%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>O'Sullivan and Weiss (1999)</td>
<td>Students indicating that it was important or very important to do well on the test</td>
<td>34%</td>
<td>12</td>
<td>22,116</td>
</tr>
<tr>
<td>Weighted Mean of Percentage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* For studies that disaggregated value statement data by assessment year, the assessment year has been provided.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Question 3: Can test-taker motivation be influenced by incentives and/or other interventions?**

This question was answered through two random effects meta-analyses of treatment effects in intervention studies. One meta-analysis compared the self-reported effort of treatment groups that received different incentives to the self-reported effort of a control group. The other compared the NAEP achievement of treatment groups that received different incentives to the NAEP achievement of a control group.

Note that, in both figures, the area of each rectangle corresponds to the weight of each study in the synthesis, the width of the horizontal line passing through each rectangle represents each study’s 95 percent confidence interval, and the summary effect is depicted by a diamond. The results of these analyses are provided in Figures 13 and 14.

The meta-analysis of intervention effects on students’ self-reported motivation (summary effect = .04) in Braun, Kirsch, and Yamamoto (2011) and O’Neil, Sugrue, and Baker (1995) suggests that interventions (e.g., certificates, financial incentives, alternative instructions) do not have a statistically significant impact on students’ self-reported effort. However, considered alone, the financial incentives in both studies did yield a statistically significant summary effect of .20.
The meta-analysis of effects of interventions on achievement in three studies (Braun, Kirsch, & Yamamoto, 2011; O’Neil, Sugrue, & Baker, 1995; O’Neil et al., 1997) yielded a modest, though statistically significant, summary effect of .10. This effect size is consistent with empirical benchmarks for intervention effects on high school students when the outcome is a standardized achievement test. For example, Hill et al. (2008) found that the mean effect size for interventions with a standardized test outcome was just .07. This suggests that incentives that presumably result in higher motivation levels may, by extension, lead to slightly higher levels of student achievement.
Summary of Findings

Research Question 1. Of the seven eligible studies that reported bivariate correlations between student motivation and achievement (Byrnes, 2003; Craig, 2013; Kim, 1992; Lee, 2013; Stokes & Cao, 2009; Walberg & Ethington, 1991; and Yepes-Baraya, 1996), all but one (Walberg & Ethington, 1991) found a positive and statistically significant correlation, i.e., greater student motivation tended to result in higher performance on NAEP (and vice versa). Meta-analysis of eligible correlational studies yielded a statistically significant summary correlation of .30, suggesting that motivation may, in fact, be associated with NAEP achievement.

Research Question 2. Descriptive statistics on students' self-reported test effort, expectancy motivation, and value motivation revealed that some students—particularly older students—are less motivated to take NAEP. Among fourth graders, the weighted mean of students reporting that they did not try as hard on NAEP as on other tests is 9 percent. For eighth and twelfth graders, it is 17 and 42 percent, respectively.

Disaggregating motivation questions into expectancy (e.g., “I am good at math”) and value (e.g., “I like science” and “It is important for me to do well on this test”) constructs provided additional insights:

  Expectancy. Across grade levels, approximately half of students report feeling that they are good at academics. The weighted mean was 54 percent, with minimal variation across grade levels (55, 55, and 48 percent for fourth, eighth, and twelfth graders, respectively).

  Value. Across grade levels, approximately half of students report that they value academics. However, there was significant variation across grade levels. Eighty-nine percent of fourth graders reported that it was “important” or “very important” to do well on the test, compared to just 34 percent of twelfth grade students. Across all data addressing the value aspect of motivation, 67 percent of fourth graders saw value in the tested subjects or in doing well on NAEP, compared to 54 percent of eighth grade students and 46 percent of twelfth grade students.

Research Question 3. Interventions were not found to have a statistically significant effect on students' self-reported effort, at least among students in grades 8 and 12 (summary effect = .04). However, some interventions—particularly financial incentives (summary effect = .20)—were found to have a statistically significant effect on students' achievement.
V. LIMITATIONS AND THREATS TO VALIDITY IN SYNTHESIS

Both the process and findings of this review are subject to limitations.

Process
Although the Project Team was intentional about its systematic review process, it is possible that their search strings failed to capture every eligible study. Moreover, the Project Team was unable to locate an abstract or full text for 21 of the studies found in its initial search. Thus, these studies could not be coded. Finally, while the Project Team deliberately restricted its search to studies from 1990 or later, nearly half (n = 7) of the 15 eligible studies relied on test data from 1990 or 1992.

Findings
This review is most limited by the motivation questions asked in each study. For example, students were asked to self-report their effort and/or motivation (as measured by various proxies), yielding inherently subjective results. The questions also varied across studies, e.g., asking how important it was to do well on NAEP versus asking about the extent to which students liked or saw usefulness in a particular subject. The Project Team addressed this issue, in part, by using random effects meta-analysis, which does not assume that all correlations or intervention effects have the same true value. Perhaps most importantly, the questions—as currently worded—may not be a reliable proxy for students’ NAEP motivation.

Additionally, results from studies of different NAEP tests were combined with an untested assumption that the relationships between motivation/effort and achievement are consistent across grades and disciplines. To address this, study data were disaggregated by grade level and motivation construct, i.e., expectancy and value. These disaggregated data could then be compared to the aggregated analyses. There were not enough intervention studies to disaggregate by incentive type; thus, the goal of those analyses was simply to show whether incentives of any type can increase achievement and/or student motivation.

Finally, none of the four intervention studies established students’ baseline equivalence. Thus, the authors may have attributed differences in NAEP achievement to differences in intervention-induced motivation, rather than to extant differences in student ability.

Additionally, limitations in the individual studies that comprise this meta-analysis (as discussed in the “Synthesis Through Critical Review of Individual Studies” section) must be taken into account when interpreting results.
VI. DISCUSSION

The Project Team’s review and synthesis of eligible literature provided critical insights into student motivation and its impact on NAEP achievement.

However, as noted in the “Limitations” section, this review and its findings are limited by the motivation questions used in each study. These questions varied across eligible studies and, as currently worded, may not accurately capture students’ NAEP motivation.

Motivation Matters

Across eligible studies, the Project Team determined that, when it comes to students’ performance on NAEP, motivation does indeed matter. Our meta-analysis resulted in a statistically significant summary correlation of .30 (see Figure 1) with a 95 percent confidence interval (.18, .33). This relationship is noteworthy, as it is comparable to other policy-relevant correlations in the literature. For example, Sirin (2005) conducted a meta-analysis of correlations between socioeconomic status and NAEP achievement and observed from a random effects analysis a summary correlation of .27 (p < .001) with an associated 95 percent confidence interval (.23, .30).

Implications. Achievement data are only reflective of students’ aptitudes when students have performed to the best of their ability. Thus, future research should explore interventions that may enhance students’ motivation for NAEP. These interventions should prioritize expectancy (more so than value), since expectancy has a stronger association with NAEP achievement.

Not All Students Are Motivated

Across studies, the meta-analysis showed that test effort is highest among younger students (i.e., fourth graders), and one in four students reports trying less hard on NAEP than on other tests. Approximately half of students report feeling confident in their academic abilities. Similarly, half of students report valuing the NAEP or academics, generally. Older students (i.e., eighth and twelfth graders) are less likely to report confidence in their academic abilities or to place value on NAEP and/or academics.

Implications. Incentives and growth mindset interventions (Dweck, 2006) should be introduced early. Growth mindset interventions stress the importance of dedication and hard work, rather than innate ability. The intensity of these incentives and interventions should increase with students’ age.

Interventions May Increase Students’ Motivation

The Project Team’s random effects meta-analyses of four intervention studies yielded two summary effects: .04 (effects of interventions on students’ self-reported effort) and .10 (effects of interventions on achievement). While the first meta-analysis did not yield statistically significant results, the second meta-analysis suggests that interventions that presumably increase motivation—in these studies, financial rewards, alternative test instructions, or a certificate—can also increase student achievement. It is worth noting that two of the four intervention studies used data from grades 8 and 12 (O’Neil et al., 1995; O’Neil et al., 1997). One study’s data were limited to grade 8 (Kiplinger & Linn, 1993). Another’s was limited to grade 12 (Braun, Kirsch, & Yamamoto, 2011). The magnitude of this effect is modest but consistent with empirical benchmarks for intervention effects on high school students when the outcome is a standardized achievement test. For example, in their synthesis, Hill et al. (2008) found that the average effect size for high school interventions was 0.27. However, this mean was based on studies that included narrowly focused outcome measures as well as broadly focused, standardized outcome measures. Given that, within the elementary school interventions in their synthesis, the overall mean effect size was larger (.33), but the mean effect size for interventions that used a standardized test was just .07, the Project Team speculates that the .10 effect size from this synthesis is likely consistent with, if not larger than, the Hill et al. (2008) average for high school interventions with a standardized test outcome.

Implications. Some interventions may have a modest positive effect on the achievement of student test-takers. However, researchers and practitioners must consider whether their interventions could be plausibly scaled up for use among thousands of students.
VII. RECOMMENDATIONS

In light of their findings, the Project Team compiled several recommendations for the NAEP Program. Some of these recommendations call for additional research and knowledge sharing. Others encourage the NAEP Program to consider how NAEP itself can generate more reliable, relevant data on students’ test-taking motivation. Notably, several of these recommendations echo those of the National Commission on NAEP 12th Grade Assessment and Reporting (2004) and the Ad Hoc Committee on 12th Grade Participation and Motivation (2005).

Revisit “Motivation” Questions on NAEP Contextual Questionnaires

The Project Team’s meta-analysis confirms that students who want to do well on NAEP are more likely to perform better, emphasizing that motivation matters. However, as has been noted throughout this report, the NAEP contextual questionnaires for students ask few motivation-related questions directly related to NAEP, e.g., how important it is to do well and how hard students tried relative to other tests. Moreover, these questions—as currently worded—do not necessarily capture students’ motivation for taking NAEP.

Additionally, the inconsistency of contextual questions year to year makes it difficult to conduct reliable studies on motivation and NAEP achievement. Recognizing this, some researchers seeking to explore the connection between motivation and NAEP achievement have utilized alternative motivation surveys or developed their own. Since there are several challenges to grouping alternative motivation measures, this poses an additional challenge to synthesizing results. To ensure that students’ answers to “motivation” questions are truly a proxy for their motivation levels, the Project Team recommends that the NAEP Program adopts more NAEP-specific motivation questions.

Commit to a Strong Set of Motivation-Related Questions

Using consistent data points is essential to tracking year-to-year changes in students’ motivation for NAEP. To ensure that researchers are able to track motivation fluctuations over time, the NAEP Program should commit to using a strong and consistent set of motivation-related questions—new or revised—for the foreseeable future. Salvaging old questions would help ensure one-to-one motivation comparisons in future studies. However, new or revised questions could focus more explicitly on the extent to which students are motivated to take the test itself and the extent to which their motivation is influenced by the administration mode, i.e., paper-and-pencil or digital-based.

Share NAEP Studies with a Broader Audience

Few researchers have evaluated students’ motivation for taking the NAEP and how, if at all, their motivation affects performance. The few studies that have been conducted were often commissioned by the Governing Board. The Project Team’s literature review revealed that such studies are rarely, if ever, cited by other scholars. This suggests that, despite their rigor and relevance, few are referenced or acknowledged by other academics. As a result, the research community has little to respond to or challenge (e.g., findings, recommendations for future research). Disseminating these publications to a broader audience could create an impetus for future studies.

Encourage Future Analyses of More Recent NAEP Data

The majority of eligible studies in this review relied on NAEP data from the early 1990s. Students, and what motivates them, may have changed since then, and standardized testing—particularly high-stakes testing—is garnering increased attention.
Notably, none of the eligible studies in this review was conducted on digital-based administrations of NAEP. As NAEP moves away from paper and pencil administration to digital-based administration, these older studies may lose their relevance. To ensure that NAEP decision-making is guided by current data, the NAEP Program should encourage future studies of student motivation to incorporate more recent test data.

Encourage Additional Intervention Studies

Intervention studies provide critical insights into how to mitigate issues of low motivation, e.g., monetary incentives and alternative instructions. However, the Project Team's review of the literature yielded just four intervention studies. Maximizing students' motivation is essential to ensuring that NAEP data accurately reflect students' aptitudes. Thus, the Project Team suggests that the NAEP Program support future intervention studies, particularly those that occur during normal NAEP administrations.
VIII. REFERENCES


National Commission on NAEP 12th Grade Assessment and Reporting (2004). 12th grade student achievement in America: A new vision for NAEP.
National Assessment Governing Board Ad Hoc Committee on NAEP 12th Grade Participation and Motivation. (2005). Preliminary recommendations for discussion with the National Assessment Governing Board.


Stokes, L., & Cao, J. (2009). Examination of low motivation in the 12th grade NAEP. Secondary Analysis Grant from Institute of Educational Sciences. Southern Methodist University, Dallas, TX.


IX. ENDNOTES


Executive Summary

This Design Document sets forth the process for conducting a literature review of research on student motivation for taking the National Assessment of Educational Progress (NAEP) and similar low-stakes standardized tests. The process set forth in this document ensures that the review is well-defined, systematic, and unbiased. Sections Two and Three describe the background and objectives of the research. Section Four describes the methodology of the research in depth, including detailed descriptions of the search strategy, standards for research article selection, training of research associates, and coding process. The research conducted pursuant to this document will culminate in an Annotated Bibliography and Synthesis Report for the National Assessment Governing Board's use to understand research in this area and to inform future policy discussions, as outlined in Section Five. Section Six provides a timeline for the activities outlined in this document.

Background

The National Assessment Governing Board (Governing Board) has a need to conduct a systemic examination of empirical research about the motivation of elementary and secondary school students for taking NAEP in Grades 4, 8, and 12. The research will be a comprehensive technical review and critical synthesis of research on student engagement on NAEP to learn the extent to which motivation may play a role in student performance on NAEP.

Objectives

The main goal of the present research is to systematically examine the available evidence for students' motivation to take NAEP and to centralize what the field knows about the extent to which sub-optimal engagement may affect student performance on NAEP. The researchers will use the findings from this research to make recommendations to the National Assessment Governing Board regarding next steps and useful foci for further research.

The Research Questions for this literature review are:
1. To what extent is test-taker motivation related to students' performance on NAEP?
2. To what extent are students motivated to take NAEP?
3. Can test-taker motivation be influenced by incentives and/or other interventions?
Methodology

4.1 Search strategy

Process

The proposed review will strive to locate and retrieve the most complete collection of studies about the relationship between student motivation and performance on NAEP, including published and unpublished research from a variety of databases. All searches will be conducted using a selected set of keywords linked with Boolean operators. Research will be conducted in four phases:

1. identifying relevant studies based on titles, abstracts, and key words;
2. including or excluding articles based on specific and strict criteria applied through a reading of the full text;
3. full coding of all eligible articles; and
4. conducting a deeper critical analysis coding by senior researchers of a selection of eligible articles.

Two separate searches will be conducted- one primary search and one exploratory searches. The primary search will identify studies that document student motivation to take NAEP and/or the relationship between motivation and NAEP performance (research questions 1 and 2). It is expected that this primary search will retrieve both correlational and intervention studies.1 The exploratory search will examine motivation in digital- or computer-based assessment environments, more broadly.2 This will enrich the discussion section of the synthesis report by providing insight into the motivation levels that might be expected once the NAEP tests have been fully converted to digital-based delivery.

Key Search Terms

Key Search Terms are intended to cover all grades and subjects. Researchers will use the connector “OR” to be inclusive and the connector “AND” to be exclusive. Researchers will utilize a search string that searches for “Subject=NAEP AND (motivation OR engagement OR incentive OR grit OR expectancy OR mindset OR perseverance OR value OR academic tenacity OR character strength OR effort OR guessing). Depending on the search rules of the source being used, the search string may include a date range to only capture studies subsequent to 1990.

Similar search strings will be used to pull studies that examine similar relevant assessments besides NAEP. Similar relevant assessments are those assessments that: are low-stakes, meaning that students do not receive individual score results and the test scores do not have any impact on their academic performance; are taken in a traditional test-taking environment, in which students work independently and are allotted a specific time to work on sections of the test; are administered by either pen and paper or digital-based programs; provide national or international student performance results; and test proficiency on at least language arts or math.

Given the above criteria for determining relevant assessments, the following assessments have been deemed relevant for purposes of the search process: Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme for International Student Assessment (PISA). Researchers will search both the acronym and full names of both NAEP and these other assessments.

---

1 In this literature review, an intervention study is a study in which the outcomes of two or more groups are compared after application of some type of intervention designed to impact motivation. For example, relevant intervention studies will likely include those studies that provide the experimental group with a variable that is predicted to impact motivation, such as monetary incentives or pre-test directions designed to influence motivation. Randomized Control Trials or Quasi-Experimental Design studies are types of intervention studies that will likely be included.

2 There was little research on digital assessment, motivation, and NAEP so the researchers did not conduct an in-depth exploratory search.
The exploratory searches examining intervention studies and motivation in digital- or computer-based assessment environments will rely on separate search strings. Key search terms will likely be expanded upon during Phase 1 of article screening (discussed below), as researchers become more familiar with relevant terms of art.

**Sources**

Initial searches will be conducted on Google Scholar to understand the breadth of research available on a wide variety of databases. The results of these Google Scholar searches will inform which databases will be subscribed to and searched going forward. Researchers will conduct searches on Web of Science, Institute of Education Sciences (IES), Education Resources Information Center (ERIC), and Teachers College Record (TCR). Additionally, researchers will search eligible documents for in-text references of additional studies. Grey literature searches (searches for unpublished studies such as dissertations and papers presented at conferences) will be conducted using the databases named above and by reviewing the in-text references of studies.

### 4.2. Research Article Selection

As discussed in Section 4.1, the process for screening studies for inclusion/exclusion will be organized in four phases. The number of articles screened at each phase will depend on the number of hits from the search strings and the content of the searched articles. If the search string initially retrieves a high number of hits, researchers will pull a random sample of the hits and review these hits for relevance. If a high proportion of the hits are irrelevant, researchers will work to refine the search string so that it produces results with a higher proportion of relevant hits. Refining the search string might involve greater use of the connector “AND” or modification of the terms used, depending on what the researchers determine is causing the high proportion of irrelevant hits.

An online survey tool will be used to code studies during all three phases. The results of this online survey tool can be exported into an Excel spreadsheet, which will be converted into a Systematic Review Table for useful visual representation of the similarities and differences across those studies meeting the threshold of the full text review phase.

**Phase 1: Relevance**

During the first phase of selection, in which researchers are screening studies based on titles, abstracts, and key words, screening criteria will be based on relevance. If a study is captured by a search term and a review of the title, abstract, and key words indicates that the study pertains to student motivation on NAEP or another specified analogous assessment, the study will be included for the next phase of review. Studies that are captured by a search term but are not related to student motivation on NAEP or another specified analogous assessment will be excluded and reasons for exclusion will be documented in the online survey tool.

Specific information that will be documented and/or coded at this phase includes:

- Name of coder
- Study ID
- Date the article was initially published
- Article title
- Study author
- Publishing organization
- Does this study address student motivation and/or engagement in NAEP as specified in the PWS?
• Is this study eligible for inclusion based on abstract screening?
  • If no, the coder will be required to select a reason for exclusion or describe the reason for exclusion if none of the options are applicable
• Supporting information, concerns, or questions

Any additional criteria for inclusion or exclusion of articles during Phase 1 will be outlined in detail in the code book, which will be completed in its initial form by December 15, 2015 and updated throughout the life of the project.

**Phase 2: Methodological Rigor Screening**

In the second phase of selection, researchers will apply more rigorous criteria for inclusion of articles. Researchers will apply separate methodological standards for observational studies (Osborne Framework) versus intervention studies (What Works Clearinghouse Standards). Only studies that involve NAEP, as opposed to international assessments or miscellaneous domestic assessments, will be coded in Phases 2-4.

AnLar will adapt the framework of Osborne (2010)³ for assessing the quality of observational designs. In the Osborne framework, key evaluative criteria includes: appropriate treatment of hierarchical (nested) data; sufficient measurement validity for all correlated variables; the statistical assumptions of correlational analyses are tested and compliance confirmed; appropriate handling of missing data and outliers; and the significance level of statistical tests is adjusted for multiple comparisons.

The standards in the Osborne framework may need to be relaxed if researchers find that strident application of the framework is disqualifying too many studies for minor considerations. Researchers will continually reassess throughout Phase 2 whether strict application the Osborne framework is needlessly disqualifying studies. The exact level of relaxation of these standards will be determined during the Phase 2 screening process and will be based on the number of studies being excluded and the specific criteria that are disqualifying these studies.

For intervention studies, AnLar will employ What Works Clearinghouse (WWC) Standards (2014). As with the Osborne framework, researchers will continually reassess throughout Phase 2 screening whether the WWC Standards need to be relaxed due to disqualification of too many studies for minor considerations. If researchers find that standards need to be relaxed, researchers will relax baseline equivalence standards, use liberal attrition standards, and relax measurement reliability cutoffs. The exact level of relaxation of these standards will be determined during the Phase 2 screening process and will be primarily based on the specific criteria that are disqualifying these studies. For example, if researchers find that the baseline equivalence standard is disqualifying an unreasonable number of studies, researchers will relax baseline equivalence in order to allow a reasonable number of these studies to qualify.

Specific information that will be documented and/or coded at this phase include:

- Sample size range
- Statistical methodology (e.g., correlation, regression/multilevel modeling, ANCOVA, etc.)
- Data reliability: Do the study’s measures (motivation and/or achievement) have reliability of 0.50 or higher? (if reported⁴)
- Study design (i.e., observational design, intervention study)


⁴ If the reliability of measures is not reported but the study meets other rigor criteria, the study will be included in Phase 3 coding. The study will be reviewed by the Principal Researcher, who will have discretion as to whether the study should be included in the final List of Resources, Annotated Bibliography, and/or Synthesis Report.
• Study characteristics- Observational design rigor
  ◦ Coders will be directed to select whether specific characteristics of the Osborne Framework do or do not apply to the study
• Intervention study type
• Baseline equivalence effect size for QED intervention studies
• Use of statistical adjustment for baseline differences in QED studies
• If the study does not meet Osborne or WWC standards, what is the reason?
• Elimination justification
  ◦ Coders will be prompted to capture additional supporting information for why the study does not meet the level of rigor within the Osborne or WWC frameworks

Any additional criteria for inclusion or exclusion of Articles during Phase 2 will be outlined in detail in the code book, which will be completed in its initial form by December 15, 2015 and updated throughout the life of the project.

**Phase 3: Full Coding of Eligible Studies**

During the third phase, all studies that pass the methodological review will be coded on technical and critical criteria. Specific information that will be documented and/or coded at this phase include:

• Intervention studies
  ◦ Intervention name
  ◦ Intervention description
  ◦ Level of independence between the intervention developer and the researcher ("Bias Firewall")

• Observational designs
  ◦ What research question does this Observational study ask?

• All Designs
  ◦ Administration mode
  ◦ Participant age group
  ◦ Assessment type (e.g. NAEP, TIMSS, etc.)
  ◦ Achievement assessment subject area
  ◦ Motivation construct references (e.g. motivation, engagement, grit, etc.)
  ◦ Independent variables
  ◦ Dependent variables
  ◦ Number of citations
  ◦ Date of assessment implementation
  ◦ Source of publication
  ◦ Funding entity
  ◦ Author affiliation at time of study
  ◦ Stated limitations of study
º Data findings summary
º Data findings critique
º Inferences/conclusions summary
º Additional information, questions, or concerns
º Additional relevant studies cited within the study

Any additional specific categories that will be included during Phase 3 will be included in the final code book, which will be completed in its initial form by December 15, 2015 and updated throughout the life of the project.

**Phase 4: Comprehensive Critical Analysis**

All eligible studies will go through a deeper critical analysis by senior staff. Specific information that will be documented and/or coded at this phase will include:

- Analysis of subjective Osborne Framework characteristics
- Hierarchical data
  - Presence of hierarchical data
  - Did researchers ignore hierarchical data?
  - Ability to apply cluster (hierarchical) data adjustments
- Comparisons
  - Presence of multiple comparisons with the same sample on the same outcome
  - Ability to make multiple comparisons using the Benjamin-Hochberg procedure
- Strength of study’s inferences/conclusions
- Critiques of study’s inferences/conclusions
- Critique of methodology
- Additional information, questions, or concerns

**4.3 The Coding Process and Training of Research Associates**

Two Research Associates and a Principal Researcher will be in charge of the coding process. The Research Associates will work independently to determine the inclusion/exclusion of articles following the criteria set forth in Section 4.2. The participation of two independent coders is aimed at reducing systematic coder bias and reducing the likelihood of mistakes.

**Independent Coding vs. Duplicate-Coding**

The Research Associates will duplicate-code studies, if reasonable. If there are too many studies to duplicate-code, coders will duplicate-code 15-20% of the articles and independently code the rest. For studies that are duplicate-coded, the Principal Researcher will act as the reconciler. For studies that are independently coded, the Principal Researcher will compute and report inter-reliability coefficients.

The ability to duplicate-code studies will depend on the number of studies that are deemed eligible in Phase 2. In order to determine whether it would be reasonable to duplicate-code studies, researchers will employ the following process: (1) have each Research Associate code two to three sample studies to determine how long

5 The 15-20% number was derived from the Principal Researchers’ experience, and is also loosely based on Mark W. Lipsey, & Wilson, D. B. (2001). Practical meta-analysis (Vol. 49). Thousand Oaks, CA: Sage publications, which states that meta-analyses should duplicate code 20 studies for the best inter-rater reliability estimates.
it is likely to take one Research Associates to code one study; (2) multiply the time it takes to code one study by two to determine how long it will take both Research Associates to code one study; (3) multiply this number by the number of eligible studies; and (4) compare this final time estimate to the budget. If the estimated time to duplicate-code all studies would be greater than the allotted budget, researchers will calculate the estimated time required to duplicate-code the entire 20% of the studies (which is the highest percentage of studies that would be duplicated coded within the 15-20% range) and compare this time to the allotted budget. If duplicate-coding 20% of articles would not fit within the budget, researchers would proceed to calculate the time required to duplicate-code a lower percentage of studies (e.g. 19% of the studies), until researchers have determined the percentage of studies for duplicate-coding that would fit within the budget.

Independent coding will only occur if there are too many studies for researchers to duplicate code all studies within budget. If the number of studies is such that researchers are able to duplicate-code all studies, then researchers will proceed with duplicate coding all studies.

**Research Associate Training and Management**

Research Associates will be trained by the Principal Researcher to use the criteria and frameworks set forth in this document. Research Associates will independently practice-code three to four studies chosen by the Principal Researcher. The Principal Researcher will meet with the Research Associates to discuss coding decisions and the rationales for different coding decisions. The Principal Researcher will utilize shared PDF documents and require the Research Associates to highlight parts of the text that were the basis for certain decisions. The Principal Researcher will calculate inter-rater reliability and ensure it is above 90%. If inter-rater reliability does not meet the 90% threshold, the Principal Researcher will institute a new round of practice coding using different articles.

In order to continually ensure coding reliability throughout the project, the Research Associates will meet bi-weekly with the Principal Researcher to compare notes. Research Associates will also meet weekly with the Principal Researcher to discuss coding issues and reconcile coding as needed.

**Tools**

Researchers will utilize an online survey tool for each study coded. This survey tool will be applicable during all article selection and coding phases. The survey will contain pre-populated categories that researchers are required to fill in. If a study meets the threshold for Phase 1, the survey will proceed to another page focusing on Phase 2 criteria. Finally, if a study meets the threshold for Phase 2, it will proceed to the Phase 3 final coding page. Studies that have been selected for critical analysis by senior staff will proceed from Phase 2 to Phase 4 coding. The purpose of this online tool is to provide a centralized location for all coded studies and to facilitate an organized method of coding. The survey also is useful in compiling the information on studies and codes for export into an Excel spreadsheet. All data collected by the survey will be exported to an Excel spreadsheet at the conclusion of Phase 3 coding.

Once an Excel spreadsheet has been exported using the survey tool, it will be reformatted into a Systematic Review Table. The title/author of the studies will comprise the columns and relevant categories (identified jointly through the AnLar/Abt team and Governing Board staff) of data extraction will comprise the rows. For study summaries and annotated bibliographies, the reader can look vertically at the coded categories for a single study. For synthesis, the reader can look horizontally (across studies) for a specified category of combination of categories. Categories will closely align with the coding categories utilized during Phase 3 of article selection and

---

6 The survey contains a field for each coder to type in his or her name and a series of questions. Each question contains answer options as either multiple choice or pull-down menus. Depending on the question, a coder may be able to type in a response to the question if the options provided do not apply. Once a coder is finished with a page, she hits the “Next” button and the survey will either end or automatically take her to the next appropriate page.
include additional details as deemed necessary during Phase 3 (please see Section 4.2 for specific categories). Systematic Table categories will be detailed in the code book, which will be completed in its initial form by December 15, 2015 and updated throughout the life of the project. A template of the Systematic Review Table with some of the included categories is displayed below.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Title/Author Article #1</th>
<th>Title/Author Article #2</th>
<th>Title/Author Article #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year Published</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source of Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funding Entity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year(s) Data Collected</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant Grade(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment Subject Area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administration Mode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation Construct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Citations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nature of Relationship between motivation and achievement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direction of Treatment Effect on motivation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnitude of Relationship between motivation and achievement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnitude of Treatment Effect/Effect size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value of relationship/effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistically significant relationship/effect at the Five Percent Level (a=0.05)?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met minimum level of criteria for either Osborn or WWC Frameworks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low attrition (RCT Intervention studies only)?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Equivalence Established (QED Intervention studies only)?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alignment with research question(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Reports**

The research process above will culminate in both an Annotated Bibliography and a Synthesis Report. The Annotated Bibliography will contain a technical synopsis of all eligible studies, including the date of the study, the assessment mode, the age of participants, the type of research study, the methodology, findings and conclusions of the study, and stated limitations of each study. The Critical Review will be part of the Annotated Bibliography.
document and will contain additional notes that critically assess the claims the authors make and highlight the particularly relevant studies to NAEP (e.g. the information coded during Phase 4). The Annotated Bibliography document will draw attention to the studies involving digital-based, computer-based, or online administration, as well as those studies that were particularly well-designed.

The Synthesis Report will synthesize and analyze the literature in an organized, straightforward manner, and also explicitly set forth the relevance to NAEP and recommend actions that the Board can take using the findings in the report. This report will delve deeper into the material set forth in the Annotated Bibliography and will also go one step further by providing the Board with a clear understanding of research on student engagement with NAEP. The Report will include a quantitative meta-analysis on eligible studies, including separate meta-analyses for intervention, observational, and descriptive statistics.

Additionally, the Background and Context section of the Report will draw upon the “most influential” studies that were not eligible for full coding as well as relevant Governing Board-sponsored studies in order to provide the reader with a greater understanding of the analysis and issues. These “most influential” studies will be chosen through the following process: (1) sort the studies into the following categories: NAEP search string ineligible, non-NAEP relevant ineligible, and non-NAEP search string ineligible; (2) obtain citation counts for all studies (when the citation counts are available); (3) pull the median number of citations and look at the mean; (4) pull the top 95th percentile of studies cited from each category. Once these top 95th percentile studies have been selected, AnLar will conduct another round of review on these studies to eliminate those studies determined not relevant and select articles that would be appropriate for inclusion and provide important context.

Milestones

Key dates for project milestones and deliverables are set forth below:

<table>
<thead>
<tr>
<th>Task</th>
<th>Activity</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1a</td>
<td>Design Document</td>
<td>12/15/15</td>
</tr>
<tr>
<td>2.1b</td>
<td>Code Book</td>
<td>12/15/15</td>
</tr>
<tr>
<td>2.2b</td>
<td>List of Relevant Sources</td>
<td>3/11/16</td>
</tr>
<tr>
<td>3.1/3.2</td>
<td>Systematic Review Table</td>
<td>3/11/16</td>
</tr>
<tr>
<td>4.1</td>
<td>Annotated Bibliography and Critical Technical Review</td>
<td>6/10/16</td>
</tr>
<tr>
<td>4.1</td>
<td>Discussion draft of Synthesis Report</td>
<td>7/8/16</td>
</tr>
<tr>
<td>4.2</td>
<td>Final Synthesis Report and research documentation</td>
<td>8/31/16</td>
</tr>
<tr>
<td>4.2</td>
<td>Governing Board Quarterly Progress Updates</td>
<td>3/3/16, 4/8/16</td>
</tr>
<tr>
<td>4.2</td>
<td>Present findings at COSDAM annual meeting</td>
<td>8/4-6/16</td>
</tr>
</tbody>
</table>
Appendix B. Code Book

CODE BOOK

CONTRACT # ED-NAG-15-C-0001

STUDENT ENGAGEMENT IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP): CRITICAL REVIEW AND SYNTHESIS OF RESEARCH

Phase 1.0: Relevance

Description: This section collects basic information about the Research Associate and the study under review. Information should be gathered from the study cover pages and the abstract.

1.1 Category: Reviewer Last Name
Options: text box
Description: In text box, enter reviewer’s last name.

1.2 Category: Study ID
Options: text box

1.3 Category: Publication Date
Options: Text box
Description: In the text box, type in the year that this article was originally published.

1.4 Category: Article Title
Options: Text box
Description: In the text box, type in the full title of the article.

1.5 Category: Publishing Organization
Options: Text box
Description: If the Publishing Organization is obvious in the abstract only, type it into the text box during Phase 1. ***If not obvious in the abstract, leave this question blank and return to this question during Phase 3 to enter the name of the organization that published this study. If a presentation, enter the meeting at which the research was presented.
1.6a Category: Does this resource address student motivation and/or engagement in NAEP as specified in the PWS?
Options: Radio buttons (select one)
• No
• Yes
• Other
  • Enter justification for uncertainty

Description: Designate whether the study addresses the Statement of Need in the PWS by selecting one of the options. If you are uncertain, select "other" and provide a brief justification.

1.6b Category: Is this resource an empirical study?
Options: Radio buttons (select one)

Description: Select "yes" if the study is new and unique data and/or research. Select "no" if it is a technical review such as a literature review. If no, flag for principal researcher before proceeding to Phase 2.

1.6c Category: Is this study eligible for inclusion based on abstract screening?
Options: Check boxes (select all that apply)
• Yes
• Uncertain
• No – Not an original, empirical study
• No – Does not discuss student motivation
• No – Does not discuss student motivation on assessments
• No – Study is not relevant to NAEP, TIMSS, PIRLS, or PISA
• No – Study does not meet sample relevance (i.e. the age range of subjects is younger than 3rd grade, or older than 12th grade)
• No – Publication date is pre-1990
• Other
  • Text box: Describe reasons

Description: Designate whether the study is eligible to be considered in Phase 2 of the review by selecting one of the options. If no options capture the reason for ineligibility, select "other" and enter the reason in the text box. If you are uncertain, mark this option, continue to question 1.7 and then flag this study for the principal researcher.

1.7 Category: Supporting information, concerns or questions
Options: Text box

Description: Type any additional thoughts, questions, concerns or notes about the article in the text box. If relevant, type your best guess of study design (intervention, observational, descriptive, psychometric, technical review, etc.). If making edits to a previous submission, add new text. DO NOT WRITE OVER PREVIOUS ENTRIES.

Phase 2.0: Methodological Rigor Screening

Description: This section collects detailed information about the rigor of eligible study methodology, design, and data.
2.1 Category: Sample size range
Options: Radio buttons (select one)
• 0-30
• 31-100
• 101-1,000
• 1,001-10,000
• 10,001-50,000
• 50,001-100,000
• 100,001-500,000
• n>500,000
Description: Designate the sample size of the population on whom this study was conducted.

2.2 Category: Statistical methodology
Options: Check boxes (select all that apply)
• Descriptive Statistics Only
• Bivariate Correlations
• ANOVA
• Multiple Regression/ANCOVA
• Multilevel Modeling
• Structural Equation Modeling
• MANOVA
• Factor Analysis
• Item Response Theory/Rasch Analysis
• Can't Tell/Don't Know
• Other
  o Text box (please specify)
Description: Designate the statistical methodology utilized by the study. If no options capture the statistical methodology, select “other” and manually type the statistical methodology in the text box. (select all that apply)

2.3 Category: Data reliability: Do the study’s measures (motivation and/or achievement) have reliabilities of 0.50 or higher?
Options: Radio button (select one)
• Yes
• Unreported
• No
Description: If “Yes” or “Unreported,” continue to Question 2.4. If “No,” skip to Question 2.10b and eliminate. (Note to Coder: Focus response to this question on the motivation effect, because we maintain an underlying expectation that the NAEP data will be reliable.)

2.4 Category: Study design
Options: Check boxes (select one):
• Observational Design (designs that involve a single group of participants and one or more variables are observed for those participants)
• Intervention Study (designs that entail comparisons between two or more groups on one or more outcomes)
• Other
  ◦ Text box (please specify)

Description: Designate the design of the study. If no options capture the study design, select “other” and manually type the study design in the text box. If “Observational Design” is selected, proceed to Question 2.5. If “Intervention Study” is selected, skip to Question 2.6.

2.5a Category: Study characteristics – Observational Design – Osborne Framework

Options: Radio buttons (all that apply)
• The goals and the correlational nature of the research questions(s) are clearly stated.
• The variables of interest are explicitly identified and operationalized.
• The sampling framework and sampling method(s) are clearly defined and justified.
• Relevant psychometric characteristics are presented and discussed. At minimum this includes reliability and factor structures. Variables with unacceptable reliability are not included in the analyses. Uses a reliability coefficient cutoff of 0.50 or higher.
• Fundamental descriptive statistics of the variables are presented and discussed (e.g., measurement scale, mean, variance/standard deviation, skewness, and kurtosis).
• The testing of assumptions underlying the analyses are presented.
• Discussion of correlational analyses refrains from making causal inferences.
• If data are to be nested, multi-level in nature, or otherwise more appropriate for multi-level modeling, those methods are used.
• Not Applicable (N/A): Nested data
• Author(s) report how outliers were defined, identifies and, if any were present, how they were dealt with.
• N/A: Outliers
• Where variables violate distributional assumptions of Pearson r, alternative correlational coefficients are used.
• N/A: Pearson r
• P values are interpreted correctly.
• N/A: P values

Discussion: Select each option for which the response is “yes” within the Osborne Framework.

2.5b Category: Observational Design - How many of the applicable characteristics were satisfied in Question 2.5a?

Option: text box

Description: Enter the number of applicable options you checked in Question 2.5a. There are a total number of 11 possible options. Four (4) could be N/A.

2.5c Category: Observational Design – Framework Rigor

Options: radio buttons (select one)
• Yes
• No

Description: Were more than 50 % the applicable criteria in Question 2.5a satisfied? If “Yes,” proceed to Additional Information and then Phase 3. If “no,” skip to Question 2.10b and eliminate.
2.6a Category: Intervention Study Type – What Works Clearinghouse (WWC) Framework
Options: Radio buttons (select one)
• Randomized Controlled Trial (RCT)
• QED
Description: If “RCT,” proceed to Question 2.7a, RCT Attrition. If “QED,” skip to Question 2.8a, Baseline Differences.

2.7a Category: Does the study report overall and differential attrition?
Options: Radio buttons (select one)
• Yes
• No
Description: If yes, proceed to question 2.7b. If no skip to 2.7d.

2.7b Category: If yes, type in the overall attrition (OA) as reported in the study.
Text box
Description: Type in the overall attrition (OA) as reported in the study.

2.7c Category: If yes, type in the differential attrition (DA) as reported in the study.
Text box
Description: Type in the differential attrition (DA) as reported in the study.

2.7d Category: If no, does the study provide the information needed to calculate attrition?
Options: Radio buttons (select one)
• Yes
• No
Description: select one.

2.7e Category: If yes to Question 2.7d enter the total number of the following in the text box below: randomly assigned; total number with outcome measure; total number randomly assigned to treatment; total number of treatment group with outcome measure; total randomly assigned to comparison; total number of comparison group with outcome measure.
Option: Text box
Description: Use N/A if certain information is not available. Enter using the following format: total number randomly assigned = , total number with outcome measure = , total number randomly assigned to treatment = , total number of treatment group with outcome measure = , total randomly assigned to comparison = , total number of comparison group with outcome measure = .
Overall attrition = (total assigned – total with outcome)/ total assigned
Treatment group attrition = (total assigned to treatment – total in treatment with an outcome) / total assigned to treatment
Comparison group attrition = (total assigned to comparison – total in comparison with an outcome) / total assigned to comparison

Differential attrition = difference between the treatment group attrition rate and the comparison group attrition rate

In Braun 2011, for the comparison of incentive 1 to control group...

\[
OA = \frac{(3117 - 1719)}{3117} = 45%
\]

\[
TA = \frac{(1565 - 884)}{1565} = 43.5%
\]

\[
CA = \frac{(1552 - 835)}{1552} = 46.2%
\]

\[
DA = 46.2 - 43.5 = 2.7%
\]

2.7f Category: Does the combination of overall and differential attrition rates exceed liberal values provided in the relevant WWC protocol?

Options: radio buttons (select one)
- Yes
- No

Description: If yes, proceed to next question. If no, Skip to Question 2.10a and eliminate. (See Table 111.1 on page 12 of the WWC Procedures and Standards Handbook v.3.0 for attrition categories.)

2.8a Category: Does the study report a baseline equivalence effect size for the main effect of treatment?

Options: radio buttons (select one)
- Yes
- No

Description: Select one option. If yes, continue to Question 2.8b. If no, skip to Question 2.8c. Baseline equivalence effect size will be based on pretest data.

2.8b Category: If yes to Question 2.8a, type the baseline equivalence effect size into the text box.

Options: Text box

Description: Enter the effect size = xx standard deviations.

2.8c Category: If no to Question 2.8a, was sufficient information provided to compute a baseline equivalence effect size?

Options: Radio buttons (select one)

Description If yes, proceed to next question. If no, skip to Question 2.10a and eliminate.

2.8d Category If yes to Question 2.8c, enter the following information into the text box below: mean pretest for
treatment group; mean pretest for comparison group; standard deviation for pretest of treatment group; standard deviation for pretest of comparison group; sample size for treatment group pretest; sample size for comparison group pretest.

Options: Text box

Description: Use N/A if certain information is not available. Enter using the following format: mean pretest for treatment group = , mean pretest for comparison group = , standard deviation for pretest of treatment group = , standard deviation for pretest of comparison group = , sample size for treatment group pretest = , sample size for comparison group pretest = .

The effect size formula we will use for extracting treatment main effects as well as these baseline equivalence effect sizes is…

Cohen’s d is defined as the difference between two means divided by a standard deviation for the data, i.e.

\[
d = \frac{X_1 - X_2}{s}
\]

\[s = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}\]

From Braun…

Turning to the estimation of treatment effects, students in the first incentive condition scored, on average, 3.4 points higher than those in the control condition, whereas students in the second incentive condition scored 5.5 points higher. Because the standard deviation of the scores overall is just under 36 points, the larger effect size is approximately 0.15. That is \(-5.5/36 = 0.15\).

Sometimes instead of using the difference between the two means in the numerator, we can use a regression coefficient for the treatment effect. Just flag instances for the principal researcher.

2.9a Category: Is the baseline equivalence effect size larger than 0.25?

Options: Radio buttons (select one)

• Yes
• No

Description: If the effect size is larger than 0.25, select “Yes” and proceed to Phase 2 Additional Information. If the effect size is smaller than 0.25, select “No” and continue to Question 2.9b.

2.9b Category: Intervention Study –Baseline differences: Did the authors use statistical adjustment to account for baseline differences?

Options: Radio button (select “yes” or “no”)

• Yes
• No

Description: If “yes,” proceed to Phase 3. If “no,” continue to Question 2.10 and eliminate.
2.10a Category: If the study does not meet standards for Intervention Studies, what is the reason?
Options: Check boxes (select all that apply)
• …the measures of effect cannot be attributed solely to the intervention – there was only one unit assigned to one or both conditions.
• …the measures of effect cannot be attributed solely to the intervention – the intervention was combined with another intervention.
• …the measures of effect cannot be attributed solely to the intervention – the effects are not reported separately for the intervention.
• …only includes outcomes that are over-aligned with the intervention or measured in a way that is inconsistent with the minimal level of rigor as defined in the Research Design Document.
• …does not provide adequate information to determine whether it uses an outcome that is valid or reliable.
• …is randomized controlled trial in which the combination of overall and differential attrition exceeds WWC standards for this area, and subsequent analytic intervention and comparison groups are not shown to be equivalent.
• …is a randomized controlled trial that either did not generate groups using a random process of had nonrandom allocations after random assignment, and the subsequent analytic intervention and comparison groups are not shown to be equivalent.
• …uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.
• …other
  ° Text box (please specify)
Description: Select all options that apply to explain why the study does not meet the minimum level of rigor within the Osborne or WWC Frameworks. If no option applies, select “other” and type in the reason.

2.10b Category: Elimination justification
Option: Text box
Description: Enter an additional justification that was not captured in Question 2.10a or supporting information for why this study does not meet the minimum level of rigor within the Osborne or WWC Frameworks. Eliminate study.

2.11 Category: supporting Information, concerns or questions
Option: text box
Description: Type any additional thoughts, questions, concerns or notes about the article in the text box. If making edits to a previous submission, add new text. DO NOT WRITE OVER PREVIOUS ENTRIES.
Phase 3.0: Full Coding of Eligible Studies

Description: This section collects detailed information to inform the Systematic Review Table and the Annotated Bibliography.

- For Intervention Studies answer Question 3.1, then skip to Question 3.3.
- For Observational Design studies, skip to Question 3.2, then proceed to the remaining questions.

3.1a Category: Intervention name
Options: Text box
Description: Type in the name of the intervention.

3.1b Category: Intervention description
Options: Text box
Description: Type in a description of the intervention; include who gathered/elicited information and how, and who used the information and how it was used.

3.1c Category – Bias Firewall (Intervention Studies only)
Options: Radio button (select one)
- Developer totally independent of researcher
- Developer collaborated with researcher (e.g., implemented the intervention)
- Developer conducted the research
Description: Select one option to indicate the level of independence between the intervention developer and the researcher.

3.2 Category: What research question(s) does this Observational Design study ask?
Options: Text box
Description: Type in the research question(s) that this study intended to investigate.

3.3 Category: Administration mode
Options: Radio button (select one)
- Paper & pencil
- Digital-based
- Hybrid
Description: Choose one option that best defines the mode for distributing the assessment.
3.4 Category: Participant age group
Options: Check boxes (select all that apply)
- K
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- Post-secondary
Description: Select the applicable grades of all participants

3.5a Category: Assessment type
Options: Check boxes (select all that apply)
- NAEP
- TIMSS
- PIRLS
- PISA
- Other
  - Text box: (please specify)
Description: Select the assessment option(s) that were included in this study. If no option applies, select “other” and type in the name of the assessment.

3.5b Category: Achievement assessment subject area
Options: Check boxes (select all that apply)
- ELA
- Mathematics
- Science
- Social Studies
- Career/Tech/Vocational
- Special Education
- Foreign Language
- Other
  - Text box (please specify)
Description: Identify the subject/content area(s) of the assessment. If no option applies, select “other” and type in the subject/content area
3.5c Category: Motivation construct references
Options: Check boxes (select all that apply)
• Motivation
• Engagement
• Incentive
• Grit
• Expectancy
• Mindset
• Perseverance
• Value
• Academic tenacity
• Character strength
• Effort
• Guessing
• Other
  ° Text box (please specify)
Description: Identify the subject/content area(s) of the assessment. If no option applies, select “other” and type in the subject/content area.

3.6a Category: Independent variables
Options: Check boxes (select all that apply)
• Time spent per question
• Responses to background questions
• Receipt of intervention
• Other
  ° Text box (please specify)
Description: Select the independent variables measured in this study. If no option applies, select “other” and type in the inputs.

3.6b Category: Dependent variables
Options: Check boxes (select all that apply)
• Motivation (or similar measurement)
• Achievement score
• Other
  ° Text box (please specify)
Description: Select the dependent variables measured in this study. If no option applies, select “other” and type in the outputs.

3.7 Category: Number of citations
Options: Text box
Description: Using Google Scholar/or search engine of choice, enter the number of times this study or article has been cited by another source (i.e. multiple introductions to the field).
3.8 Category: Date of assessment implementation
Options: Text box
Description: Enter the date or range of dates for when the assessment was implemented (i.e. the year(s) that data was collected, rather than the date of the study publication).

3.9 Category: Source of publication (i.e. journal article, dissertation, panel presentation)
Options: Radio button (select one)
• Journal article
• Dissertation
• Conference presentation
• Technical report
• Book or book chapter
• Other
  ° Text box
Description: Select one option that best describes the type of source of the study. Or select “other” and type the source in the text box.

3.10 Category: Funding Entity
Options: Text box
Description: In the text box, type in the name of the entity that funded this study. (Only if funding entity is identified)

3.11 Category: Limitations Identified by the Author(s)
Options: Text box
Description: Type in any limitations, as stated by the study author(s).

3.12a Category: Summary of Primary Findings
Options: Text box
Description: Type in summaries of the study’s data findings. In your summary, describe the main finding in terms of the groups being compared, the outcome measure being used, the grade level of the students, and the size of the effect (include statistical significance where appropriate). For intervention studies extract only the main effect of treatment, using the effect size as the metric. For observational studies, extract only the primary measure of association. These usually include correlation or regression coefficients.

3.12b Category: Ancillary Analyses
Options: Radio Button (select one)
• Yes
• No
Description: Are other findings reported beyond the main effect of treatment or primary measure of association?
3.12c Category: Ancillary Analyses Details
Options: Text Box
Description: If other findings are reported (yes to Question 3.12b) enter these analyses in the text box.

3.13 Category: Inferences/Conclusions summary
Options: Text box
Description: Type in a summary of the study's inferences or conclusions.

3.14 Category: Additional information, questions, or concerns
Options: Text box
Description: Type any additional thoughts, questions, concerns or notes about the article in the text box. If making edits to a previous submission, add new text. DO NOT WRITE OVER PREVIOUS ENTRIES.

3.15 Category: Additional relevant studies
Options: Text box
Description: Type in any additional relevant studies that are cited within this study (please add full citation information (i.e. authors, year of publication) for potentially relevant studies). If making edits to a previous submission, add new text. DO NOT WRITE OVER PREVIOUS ENTRIES. (Note to coder: Please check the Study Identifier Directory to see if all the additional resources cited are included on our resource list. If not, find the document(s) and add them to the list.)

Phase 4.0: Comprehensive Critical Analysis

Description: This section collects critical analysis of study data, design, and inferences (to be conducted by Senior Researcher(s)).

4.1 Category: Study characteristics – Observational Design – Osborne Framework
Options: Radio buttons (select all that apply)
- The substantive theory or rationale that led to the investigated relation(s) is explained.
- Results from power analyses that are in line with the chosen sample are reported.
- If analyses suggest that data on variables of interest are not reasonably normally distributed, appropriate actions are taken to normalize the data or subsequent analytic strategies that accommodate significant deviations from normality are chosen (and justified as appropriate).
- Missing data, if present, are appropriately dealt with.
- Multiple zero-order analyses are not reported unless defensible corrections for increased Type I error rates are employed.
- Authors used semipartial and partial correlations where appropriate, and interpret them correctly.
- Appropriate effect size measures are reported and interpreted.

Discussion: Select all the options for which the response is “yes” based on the Osborne Framework.
4.2a Category: Cluster data: Does the study include cluster data?
Options: Radio Button (select “yes” or “no”)
- No
- Yes
Description: Select “yes” or “no” to determine that the study included cluster data.

4.2b Category: Cluster data: Did the researchers ignore clustering of data?
Options: Radio Button (select “yes” or “no”)
- Yes
- No
Description: If “yes,” proceed to Question 4.2c. If “no,” skip to Question 4.3a Multiple Comparisons.

4.2c Category: Cluster data adjustments: 4.2c If the study ignores clustered data, are you able to apply a cluster adjustment to the reported standard errors (and thus p-values) using a default intra-class correlation coefficient = 0.20?
Options: Radio Button (select “yes” or “no”)
- Yes
- No
Description: If “yes,” proceed to Question 4.3. If “no,” skip to Question 4.4 Data Findings.

4.3a Category: Comparisons: Did the researcher make multiple comparisons with the same sample using outcomes within the same domain?
Options: Check boxes (select “yes” or “no”)
- Yes
- No
Description: If “yes,” proceed to Question 4.3b. If “no,” skip to Question 4.4 Data Findings.

4.3b Category: Comparison adjustments: Are you able to make multiple comparisons adjustments using the Benjamini-Hochberg procedure that uses the reported p-values?
Options: Check boxes (select “yes” or “no”)
- Yes
- No
Description: If “yes” or “no” proceed to Question 4.4 Data Findings.

4.4 Category: Data findings critique
Options: Text box
Description: Type in notes identifying strengths and weaknesses of the findings.
4.5 Category: Inferences/Conclusions Critique
Options: Text box
Description: Type in a justification explaining the strengths and weaknesses of the study’s inferences/conclusions.

4.6 Category: Methodology critique
Options: Text box
Description: Use the text box to identify any weaknesses or critique of the study methodology.

4.7 Category: Additional information, questions, or concerns
Options: Text box
Description: Type in any additional information, questions, or concerns you think should be recorded.
### Appendix C. Study Eligibility Status After Phase 2

<table>
<thead>
<tr>
<th>Study Citation</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Status after Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Eliminate- does not address any of the research questions; focus on “engaged reading”</td>
</tr>
<tr>
<td>Eliminate- did not answer research questions; analyzes non-response on tests</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Proceed to Phase 3</td>
</tr>
<tr>
<td>Eliminate- psychometric study that does not address the research questions</td>
</tr>
<tr>
<td>Study Citation</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mullis, I. V. S., &amp; Stancavage, F. (n.d.). Analyzing the NAEP data on testing conditions in schools.</td>
</tr>
<tr>
<td>Qian, J. (2014). An investigation of position effects in large-scale writing assessments. Applied Psychological Measurement, 38(7), 518-534.</td>
</tr>
<tr>
<td>Stokes, L., &amp; Cao, J. (2009). Examination of low motivation in the 12th grade NAEP. Secondary Analysis Grant from Institute of Educational Sciences. Southern Methodist University, Dallas, TX.</td>
</tr>
</tbody>
</table>
Appendix D. Systematic Review Table
### Figure 2: Systematic Review Table — Descriptive Characteristics

<table>
<thead>
<tr>
<th>Identifying Information</th>
<th>Descriptive Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Sample Size</td>
</tr>
<tr>
<td>Reference</td>
<td>Year Published</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
</tr>
<tr>
<td>The state of mathematics achievement: NAEP’s 1980 assessment of the</td>
<td>1991</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Reference</th>
<th>Year Published</th>
<th>Source of Study</th>
<th>Funding Entity</th>
<th>Year(s) Data Collected</th>
<th>Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</th>
<th>Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</th>
<th>Magnitude of Treatment Effect/Effect Size</th>
<th>p-Value of Relationship/Effect Size</th>
<th>Statistically Significant Relationship/Effect (α=0.05)?</th>
<th>Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?</th>
<th>Low Attrition (RCT Intervention Studies Only)?</th>
<th>Baseline Equivalence Established (QED Intervention Studies Only)?</th>
<th>Alignment with Research Question(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braun, H., Kirsch, I., &amp; Yamamoto, K. (2011).</td>
<td>2011</td>
<td>Journal Article</td>
<td>Princeton University; NCES (U.S. ED); Educational Testing Service</td>
<td>not specified</td>
<td>Positive</td>
<td>N/A</td>
<td>0.15</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Q1, Q2, and Q3</td>
</tr>
<tr>
<td>Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders.</td>
<td>2003</td>
<td>Journal Article</td>
<td>NCES</td>
<td>1992</td>
<td>Positive</td>
<td>N/A</td>
<td><em>correlation (ability-liking math) = .37 correlation (utility-relevance of math) = .03 correlation (math is fact-learning belief) = -.36</em></td>
<td>N/A</td>
<td>p &lt; .001</td>
<td><em>Yes (ability; math is fact) No (utility)</em></td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Craig, M. (2013). Attribution theory in science achievement. (Doctoral dissertation). St. John’s University, New York, NY.</td>
<td>2013</td>
<td>Dissertation</td>
<td>N/A</td>
<td>2009</td>
<td>Positive</td>
<td>N/A</td>
<td>beta= -7.153 (students who reported they did not exert effort were likely to score 7 units lower than the mean science score)</td>
<td>N/A</td>
<td>p &lt; 0.001</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Reference</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Funding Entity</td>
<td>Year(s) Data Collected</td>
<td>Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Direction of Treatment Effect on Motivation</td>
<td>Magnitude of Treatment Effect/Effect Size</td>
<td>p-Value of Relationship/Effect Size</td>
<td>Statistically Significant Relationship/Effect (p&lt;0.05)?</td>
<td>Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?</td>
<td>Low Attribution (RCT Intervention Studies Only)?</td>
<td>Baseline Equivalence Established (QED Intervention Studies Only)?</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Washington, DC: National Center for Education Statistics.</td>
<td>1993</td>
<td>Technical Report (NCES)</td>
<td>N/A</td>
<td>1992</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Jakwerth, P. M., Stancavage, F. B., &amp; Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.</td>
<td>2003</td>
<td>Technical Report-NAEP</td>
<td>NAEP Validity Studies</td>
<td>1999</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Kim, L. Y. (1992). Factors affecting student learning outcomes: A school-level analysis of the 1990 NAEP mathematics trial state assessment. (Doctoral dissertation). University of Southern California, Los Angeles, CA.</td>
<td>1992</td>
<td>Dissertation</td>
<td>N/A</td>
<td>1990</td>
<td>Positive</td>
<td>N/A</td>
<td>0.22</td>
<td>N/A</td>
<td>p &lt; 0.05</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Reference</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Funding Entity</td>
<td>Year(s) Data Collected</td>
<td>Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Treatment Effect/Effect Size</td>
<td>p-Value of Relationship/Effect</td>
<td>Statistically Significant Relationship/Effect (p=0.05)?</td>
<td>Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?</td>
<td>Low Attribution (RCT Intervention Studies Only)?</td>
<td>Baseline Equivalence Established (QED Intervention Studies Only)?</td>
<td>Alignment with Research Question(s)</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Kiplinger, V. L., &amp; Linn, R. L. (1993). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. Educational Assessment, 3(2), 111-133.</td>
<td>1993</td>
<td>Technical Report (NCES)</td>
<td>NCES</td>
<td>1990, 1992</td>
<td>N/A Positive (for items 1-9)</td>
<td>N/A</td>
<td>0.18 (for items 1-9)</td>
<td>Not available</td>
<td>Yes (for items 1-9) No (for items 10-17)</td>
<td>Yes</td>
<td>Cannot tell</td>
<td>No</td>
<td>Q1, Q2, and Q3</td>
</tr>
<tr>
<td>Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.</td>
<td>2013</td>
<td>Journal Article</td>
<td>none reported</td>
<td>1998, 2007</td>
<td>Positive (but significant variation by gender)</td>
<td>N/A</td>
<td>N/A</td>
<td>Not available</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>Q1 and Q2</td>
</tr>
<tr>
<td>Reference</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Funding Entity</td>
<td>Year(s) Data Collected</td>
<td>Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Treatment Effect/Effect Size</td>
<td>p-Value of Relationship/Effect</td>
<td>Statistically Significant Relationship/Effect (p=0.05)?</td>
<td>Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?</td>
<td>Low Attribution (RCT Intervention Studies Only)?</td>
<td>Baseline Equivalence Established (QED Intervention Studies Only)?</td>
<td>Alignment with Research Question(s)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Stokes, L., &amp; Cao, J. (2009). Examination of low motivation in the 12th grade NAEP. Secondary Analysis Grant from Institute of Educational Sciences. Southern Methodist University, Dallas, TX.</td>
<td>2009</td>
<td>Technical Report</td>
<td>U.S. Department of Education</td>
<td>2005</td>
<td>Positive</td>
<td>N/A</td>
<td>N/A</td>
<td>p = 0</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>Q1 and Q2</td>
</tr>
<tr>
<td>Reference</td>
<td>Year Published</td>
<td>Source of Study</td>
<td>Funding Entity</td>
<td>Year(s) Data Collected</td>
<td>Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA</td>
<td>Magnitude of Treatment Effect/Effect Size</td>
<td>p-Value of Relationship/Effect</td>
<td>Statistically Significant Relationship/Effect (α=0.05)?</td>
<td>Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?</td>
<td>Baseline Equivalence Established (QED Intervention Studies Only)?</td>
<td>Alignment with Research Question(s)</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------</td>
<td>--------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>Yepes-Baraya, M. (1996). A cognitive study based on the National Assessment of Educational Progress (NAEP) science assessment. Princeton, NJ: National Assessment of Educational Progress.</td>
<td>1996</td>
<td>Technical Report</td>
<td>NAEP (adminis-tered by the Office of Educational Research and Improvement, U.S. ED)</td>
<td>1995</td>
<td>&quot;Standard block scores and perceived ability: positive and weak; Black scores and perceived block difficulty: positive and weak&quot;</td>
<td>N/A</td>
<td>N/A</td>
<td>Not available</td>
<td>&quot;Yes (perceived ability) No (perceived block difficulty)&quot;</td>
<td>Yes</td>
<td>N/A</td>
<td>Q1 and Q2</td>
<td></td>
</tr>
</tbody>
</table>

Appendix E.  
Annotated Bibliography and Technical Review

PARTICIPANT ENGAGEMENT IN NAEP:  
ANNOTATED BIBLIOGRAPHY AND TECHNICAL REVIEW

CONTRACT # ED-NAG-15-C-0001

STUDENT ENGAGEMENT IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP): CRITICAL REVIEW AND SYNTHESIS OF RESEARCH

Background

In September 2015, AnLar Incorporated (Project Team), along with its subcontractors, Abt Associates and Minds Incorporated, were awarded a contract to conduct a systematic literature review documented via an annotated bibliography and synthesis summary. The goal of this review was to capture what the field knows about the extent to which sub-optimal engagement and/or test administration may affect students' performance on NAEP.

A systematic review of extant literature on students' motivation for taking NAEP yielded 15 eligible studies.

Each study answers one or all of the following questions:

1. To what extent are students motivated to take NAEP assessments?
2. To what extent is test-taker motivation related to administration of and performance on NAEP?
3. Can test-taker motivation be influenced by incentives and/or other interventions?

Introduction

IDENTIFYING STUDIES

After identifying key search terms, the Project Team developed a comprehensive search string ("Subject = NAEP AND (motivation OR engagement OR incentive OR grit OR expectancy OR mindset OR perseverance OR value OR academic tenacity OR character strength OR effort OR guessing)"). Similar search strings were used to identify studies on students' motivation for taking Trends in International Mathematics and Science Study, Progress in International Reading Literacy Study, and Programme for International Student Assessment (other low-stakes assessments with very similar characteristics to NAEP). All search strings were adapted for use with the following resource libraries and databases: Education Resources Information Center (ERIC), Web of Science, Teachers College Record, Institute of Education Sciences (IES), and National Center for Education Statistics (NCES). Across libraries and databases, this initial search yielded 1,039 results. Twenty-one resources were eliminated because researchers were unable to locate either their abstract or full text, resulting in an initial samples size of 1,018.

SCREENING RESULTS

The Project Team then used a four-phase screening process to identify and analyze eligible studies. In Phases 1-3, studies were either excluded or advanced to the next phase. In Phase 4, the Project Team's principal researcher reviewed eligible studies and engaged in a critical examination of methodology, inferences, and conclusions.
1. **Phase 1, Relevance:** The Project Team's research associates duplicate-coded all study abstracts for relevance (e.g., empirical studies that addressed one or both of the guiding research questions) \( n = 1,018 \).

2. **Phase 2, Methodological Rigor:** The Project Team's research associates used key characteristics (e.g., sample size, statistical methodology, study design) and widely-accepted research standards (e.g., Osborne Framework for observational studies; What Works Clearinghouse standards for intervention studies) to screen studies for methodological rigor \( n = 27 \).

3. **Phase 3, Full Coding:** The Project Team's research associates completed a more thorough review of the remaining studies. Each was coded for additional criteria, including participant age group, funding entity, data findings, and stated limitations. The 15 studies deemed eligible after Phase 3 are included in this annotated bibliography \( n = 15 \).

4. **Phase 4, Technical Review:** The Project Team’s principal researcher critically analyzed all 15 studies from Phase 3, critiquing each study’s methodology, inferences, and conclusions. Data from the Phase 4 review are included in this technical review.

**Annotated Bibliography with Technical Review**


This randomized controlled field trial investigated the effects of monetary incentives on twelfth graders’ performance on a reading assessment closely modeled after the NAEP reading test. The authors hypothesized that scores obtained at regular NAEP administrations underestimate student capabilities. The study used a convenience sample of 2,600 students from 59 schools across seven states. Students were either assigned to a control group or one of two incentive interventions: a “fixed” incentive, which offered students $20 at the start of the session or a “contingent” incentive, which offered students $5 in advance and $15 for correct responses to each of two randomly chosen questions for a maximum payout of $35. Students assigned to the fixed incentive group scored, on average, 3.4 points higher than those in the control condition, while students assigned to the contingent incentive group scored an average of 5.5 points higher. The authors note some limitations of the sample selection: the study used a convenience sample, making it difficult to generalize its results to the national population; and student response rates (23.1-78.2 percent) and school participation (2-59 percent) varied widely from state to state, and the overall participation rate was just 56 percent. The authors also cite limitations of the testing atmosphere and treatments, noting that only four blocks of items were employed (significantly fewer than are used in the official NAEP examination) and that the findings depend on whether incentives groups understood the nature of the monetary incentives and were not aware of the study ahead of time.

**Technical Review**

The authors provided useful information about the study including threats to validity and limitations, as well as descriptive statistics that can be used to calculate effect sizes for treatment effects. The latter is fortunate, as an effect size for the comparison of Incentive 2 (contingent incentive) to the control was never reported. In addition, it is peculiar that the Analysis of variances (ANOVAs), arguably the most sophisticated analyses of the study, were conducted by gender groups only. Thus, there is no statistical significance test for the treatment effects on the overall study sample.

The authors offered helpful interpretations of the size of the treatment effects. This included reporting and describing the effects in their original metric (NAEP score points) and comparing the effects to those of education achievement gaps. The authors also made several salient points for policy decisions about NAEP.
This is a carefully conducted intervention study, but three issues present themselves as threats to validity. Threats to internal validity include the high rates of attrition for the two treatment groups (each over 45 percent). This level of attrition increases the likelihood that the analytic sample of students who were actually tested do not have the same properties as those students originally assigned (randomly). One way to partially mitigate this threat is to demonstrate that the groups were baseline (prior to the interventions) equivalent on reading-related outcomes, but the authors do not demonstrate this in enough detail. Vague comments about equivalence are not sufficient. The combination of high attrition and a lack of demonstrated baseline equivalence would disqualify this study from meeting What Works Clearinghouse evidence standards. Finally, the authors chose to ignore clustering (students within schools) in their analysis. This could have led them to underestimate standard errors for statistical significance tests and, in turn, underestimate the likelihood that a Type I error has been made.


This NCES-funded observational study sought to identify which variables accounted for test score variance among White, Black, and Hispanic examinees on the 1992 NAEP Mathematics Assessment. Each of the 318 participating schools contributed approximately 30 twelfth grade students to the assessment (N = 9,499). Across all ethnic groups, liking math was positively correlated with math proficiency (.37, p < .001) and believing that “Math is mostly memorizing facts” was negatively correlated with math proficiency (-.36, p < .001). A composite variable labeled “Utility-Relevance of Math” (comprised of responses to “All people use math in their jobs” and “Math is useful for solving everyday problems”) was not found to have a statistically significant correlation with math proficiency.

Technical Review

The author provided standard statistical output from correlation matrices and regression tables. These included helpful information such as standardized regression coefficients. However, it would have aided interpretation to also have descriptive statistics, exact p-values, and confidence intervals. Further, the effects of predictors are all interpreted in terms of variance explained (R2). A more straightforward approach to interpreting the motivation factors would be to explain the unstandardized regression coefficient, which provides the effect in its raw metric (e.g., for every one category increase in motivation belief, proficiency increases by .22 points).

This study made limited inferences about motivation on NAEP, summarizing that motivational factors were predictive of achievement and that motivation was a malleable factor that can be increased through intervention. In the author’s defense, the effect of motivation on performance was just one aspect of this much larger study, so the conclusions necessarily focused elsewhere.

This study was well-conceived and carefully conducted. The author took appropriate steps for adjusting the standard errors of significance tests to reflect the nested nature of the data. However, there were two analyses presented, one with a subsample of the other. In this situation it would have been appropriate to correct the significance level of statistical tests to account for the increased type I error rate (i.e., multiple comparison correction). Further, the author reported p-value thresholds (e.g., p < .001) and not exact p-values; it is impossible for readers to make the correction. That said, the very low Type I error rates of the individual tests suggest that a multiple comparisons correction would likely not change the decision about the null hypothesis.

This observational study examined several potential malleable factors that may predict the science achievement of twelfth graders (including student-perceived effort) using data from the 2009 NAEP science assessment. The sample group consisted of 11,531 twelfth grade students, representing a national population of 3,214,000 American students in public or private schools enrolled in a science class and in twelfth grade during 2009. Principle component factor analysis was used to determine the specific items that contribute to each overall factor. A series of multiple regressions were then analyzed to determine the predictive value of each of these factors for science achievement. Test effort was found to be significantly correlated with student achievement (p < .001). “Students who reported that they did not exert effort on this NAEP assessment or take the assessment seriously were likely to score seven units lower than the mean science score “ (p. 51). The author cautions that the study is limited to twelfth grade, American science students and that many of the sample questions from which the data is compiled consists of student self-reported data, which could potentially contain several sources of bias such as selective memory, exaggeration, or dishonesty. The author further cautions that the lack of test questions related to items such as task difficulty limits the ability to generalize about attribution theory.

Technical Review

The author transparently and systematically reported on all tested hypotheses. It was prudent to estimate and report the reliability coefficients for all scales, including the effort and self-concept scales. Finally, the author put effort into interpreting the unstandardized regression coefficients, a helpful elaboration for readers with limited statistical background.

That said, it would have been helpful to compute and report descriptive statistics and/or effect sizes for the quasi-experimental groups formed by the survey responses (i.e., for students with self-reported low or high levels of effort). This would have allowed for more direct comparisons of relationships within and outside of this study context. In addition, exact p-values should have been reported for all significance tests, rather than p-value thresholds.

The author’s discussion of each set of findings was comprehensive, and any speculations could be defensibly supported by the analyses. However, there were no connections made between this study’s findings and those of the extant literature. It would have been helpful to know whether these findings support or challenge prior observations.

The author was wise to recognize that the general effort scale’s reliability was too low to trust in the regression analysis. As such, the scale was discarded prior to analysis. However, multiple statistical tests were conducted on the same sample within the same outcome domain, and no corrections were made. Specifically, there should have been corrections (e.g., Bonferroni or Benjamini-Hochberg) to the reported p-values of the regression analyses.

1 This sentence is a direct quotation of the study author on page 51 of the report. AnLar notes that the actual question on this NAEP administration asked students to respond to their “level of effort on this science test as compared to others: “Not as hard as others,” “About as hard as others,” “Harder than others,” and “Much harder than others.”

This technical report, funded by NCES, offers descriptive statistics about all participating fourth, eighth, and twelfth grade students’ performance on the 1992 NAEP mathematics exam. Chapter 12 focuses specifically on students’ responses to the test’s motivation-related background questions. Of particular interest are Tables 12.5 (students’ reports on how hard they tried on the NAEP relative to other mathematics exams) and 12.7 (students’ reports on how important it was for them to perform well on the NAEP). Data from these tables suggest that students’ reported motivation does not directly affect their NAEP performance. In fact, students who reported trying “harder” or “about as hard” on the NAEP outperformed those who reported trying “much harder” across all three grades. Among eighth and twelfth graders, average proficiency scores were actually higher for students who reported trying less hard than for those who reported trying much harder. Similarly, students who reported that it was “important” or “somewhat important” to perform well on NAEP outperformed those who claimed it was “very important.” Notably, eighth and twelfth graders who reported it was “not very important” to perform well still averaged higher proficiency scores than those who said it was “very important.”

Technical Review

As a data compendium, this report does what it was intended to do. That is, it provides means, percentages, and standard errors for various NAEP variables, including those related to student motivation. The tables do not include statistical significance tests for differences in groups. However, the footnote reminds readers that they can use the standard errors in Appendix A to form 95 percent confidence intervals, and that the degree to which these confidence intervals overlap will provide information about the statistical significance of any differences.

As a compendium, this publication was not intended to provide interpretations of its findings. As such, no interpretation was provided.

The methods used were entirely appropriate given the purpose of the report.


This descriptive, mainly qualitative study explored potential reasons behind student omission of responses to assessment questions. The authors visited schools in which the 1998 eighth grade NAEP assessments in reading and civics were being conducted. After the assessment sessions, they interviewed a sample of 84 students about their test taking behaviors and their reasons for not answering particular questions. Sixty-five of the 84 students had at least one unanswered question. Of these, 66 percent of those taking the civics assessment and 73 percent of those taking the reading assessment indicated that doing well on the test was either important or very important; over 80 percent said they tried at least as hard as on other tests; and 63 percent said they would not try harder if the test was graded. The authors concluded that lack of motivation did not seem to be a significant factor in students’ omission of responses; however, eight students did indicate a lack of motivation, and students at two particularly low-income sites seemed generally unmotivated to answer test questions. The authors also noted that lack of motivation was apparent in the behavior of many students (e.g., talking, inattention). The sample in this study was chosen to be diverse rather than representative, so it is not possible to draw statistically significant meaningful conclusions about the demographic characteristics of students likely to omit questions.
Technical Review

This small-scale (n = 65) qualitative study explored reasons why students don’t answer certain questions on the eighth grade reading and civics NAEP tests. Interviews revealed that motivation (i.e., task value via the importance of doing well on the test) was not a major influence on non-response except at a few sites. Often, students’ lack of motivation manifested as rushing and/or not reading questions thoroughly. These conclusions appear to be valid based on the interview data “excerpts.”

The authors suggest that creating more relevant item contexts might improve motivation. This is a reasonable recommendation. However, it is not clear that the motivation issue was widespread enough among the sampled students to warrant a recommendation for item revision.

The interview methods used in this study were appropriate given the study’s goals and research questions.


This observational study sought to identify which school-related factors significantly impact students’ learning outcomes in mathematics. Eighth graders’ mathematics data from the 1990 National Assessment of Educational Progress Trial State Assessment (TSA) were used as a proxy for students’ learning outcomes. The study sample contained data for about 100 schools from 37 states, for a sample size of 3,551 schools. Thirty students selected from each school provided a sample size of approximately 3,000 students per state. The author’s model included five predictor constructs (e.g., student characteristics, school conditions, student behavior) measured by eight variables. The percentages of students who agreed with the statements “I like math” and “I am good in math” were compiled into one such variable: “Students’ Math Perception.” This variable was found to have a positive (0.22) and statistically significant (p < .05) relationship with mathematics achievement. The author warns that these results may not be generalizable to other grades and subjects or to the country at large, since the sample only includes data about eighth graders’ mathematics performance in 37 states.

Technical Review

In this doctoral dissertation, the author estimated the relationship between students’ achievement and their mathematics self-perception. This relationship was expressed appropriately through both simple bivariate correlations, as well as multivariate regression techniques as part of a path analysis framework. The author provided helpful psychometric information for the perception scale used to quantify students’ math perception. However, the study lacked interpretation beyond citing the statistical significance of the estimated relationships. Even these were reported with p-value thresholds, not exact p-values. Further, there were instances in which statistical significance tests were applied to the same sample of students for research questions in the same outcome domain. There should have been multiple comparisons corrections applied to the p-value estimates.

As noted, the study lacked interpretation beyond reporting the statistical significance of the perception-achievement relationship. As the statistical significance tests of these relationships were highly powered (large sample size), it is impossible to know whether the relationships are truly noteworthy. Practically speaking, the perception-achievement relationship, estimated from the multivariate regression, was not interpreted (again, only its statistical significance). Further, given the limited reliability of the perception scale, the author should have cautioned readers against interpreting the estimated perception-achievement relationship. Specifically, the original four-item scale had a reliability coefficient (Cronbach alpha) of 0.63, which is reason for caution. The reliability of the two-item scale ultimately used in the regression analysis was not reported, just the simple correlation between the items (r = 0.55). In either case, the author should have acknowledged this important
limitation in the study’s discussion. Finally, in the author’s defense, the perception-achievement relationship was just a small portion of a large study that examined the influence on achievement of other student characteristics, teacher characteristics, school conditions, and teacher behaviors. Therefore, the author had much to address in the study’s implications beyond the perception-achievement relationship.

As stated above, corrected p-values should have been estimated to adjust for multiple comparisons within an outcome domain.


This observational study assessed the performance of eighth grade students completing two subsets of NAEP Block 7 mathematics items that were administered as part of the 1992 Georgia Curriculum-Based Assessments (CBA) (n = 80,836) as compared to the results from Georgia’s participation in the 1990 NAEP Trial State Assessment (TSA). The purpose of the study was to investigate whether differences in test administration conditions and presumed levels of motivation engendered by the different testing environments affect student performance on NAEP, using the 1992 CBA as a “higher-stakes” environment and the 1990 TSA to simulate the “low-stakes” environment of the current NAEP administration. The mean scores of the first subset of NAEP items were significantly higher in the 1992 CBA administration than in the 1990 TSA administration (effect size = .18), while the CBA and TSA mean scores were not significantly different for the second subset of NAEP items. The authors cautioned that differences in the two test administrations might have affected results. Time of testing and practical time constraints (which forced the split of the block of 17 multiple-choice items into two subsets of the first 9 and the last 8 items for the State Embedded administrations) were cited as factors that might have affected results. The authors also cautioned against over-interpreting small differences in mean scores between 1990 and 1992 because the analysis was based not only on the performances of two different cohorts of eighth grade students (1990 and 1992 cohorts), but also on the performances of students tested in February (1990) and of those tested in May (1992), so small increases in scores may have been due to the additional two to three months of instruction students in the May administration received.

Technical Review

This study employed a clever design, comparing responses to NAEP Block 7 items on a low-stakes assessment (NAEP) to responses to the same questions embedded in a presumably higher-stakes state testing environment. The NAEP items were broken into two clusters for insertion into the state test. Descriptive statistics for both clusters were reported, but only the effect size for the first cluster (d = .18) was reported. The overall effect size would have been helpful to interpretation.

The logic of the study design hinges on the debatable assumption that students care enough about teacher and school ratings from state tests, even when they do not receive individual scores, to do their very best. While the comparison that the authors sought to make was a worthy endeavor, the comparison of NAEP scores between the 1990 NAEP administration and the 1992 state test administration with embedded NAEP items is confounded by other factors, most of which were astutely identified by the authors: potential differences in student populations across those years, differences in test difficulty and duration, differences in study context, and differences in timing of the tests.

There are three additional unstated limitations to the study. First, the fact that the NAEP items were embedded near the end of the state test suggests that they might not have received the full energy and effort of students. While this limitation is not expressly stated, the authors do effectively situate the context in which the .18 effect size could be interpreted. For example, they caution that recent research shows that one might expect an effect
size of .04 just from annual fluctuations in form, difficulty, and context. The authors ultimately conclude that there may not be a real difference in achievement based on the perceived stakes of the testing situation, as the small effect could easily be explained by differences in populations, testing context changes, and differing amounts of instructional time before test administration. This conclusion seems reasonable.

Second, the authors provide no information that suggests the students were equivalent on achievement prior to being exposed to either the low-stakes or higher-stakes testing environment. The observed difference in scores (the effect) could have been there all along, regardless of testing environment.

Finally, the analyses appear to ignore clustering of students, so the stated Type I error probabilities are likely underestimated.


Lee uses eighth grade writing data from the 1998 (N = 20,586) and 2007 (N = 139,900) administrations of the NAEP to explore how writing attitudes and learning behavior affect writing performance across gender. As predicted, students who reported positive attitudes toward writing (“I like to write”; “I am good at writing”; “Writing is one of my favorite activities”) performed better on the exam. The effect sizes associated with these statements were medium to large, ranging from .5 to .9. It is notable, however, that females scored substantially higher than males, even when they reported similar writing attitudes. In the 2007 data, even females with the most negative writing attitudes achieved higher scores than males with the most positive attitudes.

Technical Review

This study employed a strong observational design and methodology with an astute de-emphasis of statistical significance and a focus on effect sizes. This approach was quite effective, allowing the results (effect sizes) of many analyses to be defensibly compared and easily interpreted within the context of this study, as well as in the context of other study results.

The authors put much effort into comparing their results with those of other seminal works, pointing out key instances of consistency and inconsistency with the extant literature. This is quite helpful to practitioners, researchers, and decision makers. One minor critique with regard to interpretation of effect sizes is that the authors occasionally fell into the trap of calling certain effect sizes “small” or “medium,” using cutoff values that are not necessarily well-established for this unique field of study.

There were no methodological problems of note.


This NCES technical report provides results from the 1990 NAEP Mathematics assessment and the 1990 NAEP Trial State Assessment in Mathematics, including students’ responses to background questions. The 1990 NAEP Mathematics assessment included 1,237 participating schools and a total of 26,472 students. The report provides the percentage of students and average proficiency of students in the national assessment who “Strongly Agree”; “Agree”; or who are undecided, “Disagree,” or “Strongly Disagree” with the statement “I Like Mathematics.” The report similarly provides the percentage of students and average proficiency of students who “Strongly Agree”; “Agree”; or who are undecided, “Disagree,” or “Strongly Disagree” with the statement “I Am Good in Mathematics.” The report does not conduct analysis on the descriptive data, but observes that the majority of students appeared to have
positive perceptions toward mathematics and that those with positive perceptions also had higher proficiency levels. However, the report also observes that only two-thirds of the fourth graders reported liking mathematics, and by grade 12, only half reported that they liked the discipline; additionally, fewer than two-thirds at any grade strongly agreed or agreed that they were good in mathematics.

**Technical Review**

This NCES report takes a simple descriptive approach to describing the extent to which students are confident in their mathematical abilities and value mathematics, and whether students at different levels of these affective variables have different proficiency levels. The study draws upon established NAEP sampling and psychometric methods and, as such, is presumed valid. The inferences drawn from this study are quite limited in scope and tied directly to descriptive comparisons of the raw NAEP data. Since the goal of this report was not to infer causes of student confidence in or value for mathematics nor the mechanisms for how these affective variables influence mathematics proficiency, this is not problematic. Although this study does not look at background questions directly related to motivation on the NAEP test, the questions collect information on the value students place on math and their self-perceptions of math ability, which are fundamental components of motivation theory generally. Thus, this study was included in the literature review because motivation on mathematics generally is related to whether students are motivated to take a math test.

The presented means and standard errors by confidence or value are appropriate. However, the reader is given no indication about whether differences are large compared to what might be expected through sampling error or by some practical standard. Tests of statistical significance or practical significance (i.e., effect sizes) would have been very helpful in the presentation of these findings. Further, relationships between affective variables and proficiency are presented descriptively and are hard to interpret without a common metric such as a correlation coefficient.


The main study reported was a randomized controlled trial that examined the effects of various reward and instruction treatment conditions on 749 eighth grade students (four treatment conditions) and 719 twelfth grade students (five treatment conditions) from Southern California on two blocks of released items from the 1990 NAEP mathematics test. Students were either assigned to the control group (in which standard NAEP instructions were read) or to one of four interventions: a monetary incentive of $1 for every item answered correctly; ego-involved instructions read at the beginning of the test; task-involved instructions read at the beginning of the test; or a certificate of accomplishment for performing in the top 10 percent of one’s class (grade twelve only). In addition to the test, a self-assessment questionnaire was administered to measure self-reported effort and associated metacognitive variables. The treatment effect on eighth grade students’ easy mathematics items score was F(3, 717) = 2.7, p = .043. Scheffe post hoc comparisons showed that eighth grade students who were promised $1 for every item they answered correctly scored higher (easy items: M on easy items = 7.8, SD = 1.2, n = 183) than students who were given either task-oriented instructions (M = 7.5, SD = 1.6, n = 199), ego instructions (M = 7.7, SD = 1.3, n = 196), or standard NAEP instructions (M = 7.5, SD = 1.5, n = 171). There were no differences among eighth grade treatment groups on metacognitive or affective variables. The correlation between total mathematics score and self-reported effort for eighth graders was .24 (p < .01). In Grade 12, there were no differences among the test scores of students who received different test instructions. However, the group that received the financial incentive reported more metacognitive activity than the group who got the standard NAEP test instructions. The correlation between total mathematics score and self-reported effort for twelfth graders was .22 (p < .01). The results of the “manipulation check,” in which students were asked to identify the test instructions
they had received, indicated that only 444 Grade 8 students and 473 Grade 12 students remembered their test instructions by the end of the test; therefore, a separate analysis of these subsamples was conducted. Whenever the results for the subsample differ from the results for the full sample, the results for the subsample are reported in addition to the results for the full sample.

Technical Review

The authors focused much of their reporting on tests of statistical significance of effects, which taken alone, has many perils. Further, within this framework, only statistically significant effects were reported. For example, significant effects for the easy items were reported but not for the overall test. No effects were presented for twelfth grade students. This practice limits the contribution of this paper.

The authors were prudent to provide descriptive statistics for the eighth grade sample of students. These descriptive statistics provided the only way to estimate an effect size for the treatment effects. The only effect size reported was for a subsample of eighth grade students, those who understood the test directions. Effects based on this subsample are more suspect, as these students are a non-random subsample and therefore do not retain the properties of the original randomized sample.

In the authors’ approach, “no difference” in achievement actually means a non-significant (statistically) difference. This is a problem. With as few as 276 total students in some of the comparisons (i.e., limited statistical power), some of the non-significant differences could be large enough to be noteworthy. In the absence of effect sizes, this paper lacks interpretation of the magnitude or importance of the treatment effects. Further, other related interpretations could have been offered. For example, it would have been useful to know what portion of the score difference between NAEP proficiency levels is represented by these treatment effects.

Three important features of this study are a threat to its internal and statistical validity. Under internal validity, it is not entirely clear that the 749 students in the analytic sample were the exact students who were randomly assigned to treatments. That is, we cannot assess whether there was sample attrition that could have biased the treatment effects. Further, no baseline measure was used to adjust treatment effects for extant differences in mathematics achievement. Finally, under statistical validity, the fact that the students’ achievement data were nested within schools was ignored. Thus an important source of variation (between-school) was ignored and the standard errors of the statistical significance tests were likely too small. This results in an underestimation of the likelihood of Type I error (i.e., the p-value).

All of this said, the authors’ more general implications for policy makers and for the National Assessment Governing Board were well-conceived, realistic, and potentially actionable. Overall, despite some of the noted limitations in the reporting practices and methodological approach, this paper makes an important contribution to the literature on this topic.


See the summary of O’Neil, Sugrue, and Baker (1995) article in the Technical Review above, as this article reports on the same study. This article additionally reported that the eighth grade treatment groups differed in their reported effort, F(3, 713) = 3.22, p = .02, but the mean effort score of the group who was offered $1 per item

---

was not judged to be significantly higher than the mean effort scores of the other groups when Scheffe post hoc multiple comparisons were conducted.

Technical Review

The authors reported descriptive statistics disaggregated by motivation treatment groups, but only for statistically significant effects. It would have been helpful to have these descriptive statistics for all effects. Similarly, the authors conflated having an effect with having a statistically significant effect. In some cases, they simply reported that the treatment had "no effect" when it is unlikely that the effect was indeed zero. Reporting effect sizes would have illuminated the true size of the effects in a common metric and allowed readers to decide for themselves whether the effects are noteworthy. Further, the authors alluded to some amount of sample attrition (i.e., students randomly assigned but not tested) but did not report the extent of the attrition or whether it differed across treatment groups.

In general, the authors' conclusions were thoughtful, but they missed many opportunities to provide useful information to the reader. For example, the authors noted that their findings are consistent with some extant research but never cited specific research. Further, the discussion is entirely reliant on statistical significance in deciding whether an effect was observed or whether a difference is present. This is a dubious approach. Similarly, the fact that no effect sizes were computed prohibited the comparison of this study's findings to extant studies of motivation and NAEP achievement.

The observed effects would have been more convincing had some baseline achievement measure been used to demonstrate comparability of the treatment groups prior to intervention. Such a baseline measure could have also been used as a covariate to adjust the post-intervention means for pre-intervention differences. Finally, because there were multiple statistical tests on the same student sample in the same outcome domain, there should have been appropriate corrections to reported p-values.


This NCES technical report summarizes the 1996 NAEP science results of students in grades four, eight, and twelve, including students' responses to motivation-related background questions. It provides the average score associated with specific questionnaire responses, as well as the number of students for each response scoring at or above proficient. Relevant questions included "How many questions do you think you got right...?" (responses included "Almost All," "More Than Half," "About Half," and "Less Than Half"); "How hard was this test compared to most other tests...?" (responses included "Much Harder," "Harder," "About as Hard," and "Easier"); and "How hard did you try on this test compared to how hard you tried on most other science tests...?" (responses included "Much Harder," "Harder," "About as Hard," and "Not as Hard") and "How important was it to do well...?" Based on students' responses, the authors concluded that, in general, older students were less motivated to do well on the assessment than were younger students. The authors also noted numerous differences among students at different grade levels related to their motivation and performance on the NAEP science assessment.

Technical Review

This NCES report provides clear descriptive information about students' responses to NAEP questionnaire items about motivation, as well as the students' corresponding scale scores and proficiency levels. However, it stops short of testing whether differences in motivation levels correspond to meaningful differences in performance on NAEP. This could have been done with effect sizes or tests of statistical significance. However, those are interpretations that are helpful to the researcher, not the intended audience of science teachers. With this latter
point in mind, the report meets the goal of providing general (although sometimes mixed) conclusions about the relationship between motivation and performance.

This report attempts to draw clear conclusions about the relationship between motivation and grade level, and motivation and performance on NAEP. Although the grade level differences were clear, drawing conclusions about motivation and performance was difficult as the relationships were inconsistent at times. The authors were clear on this point. This report makes few inferences about the causes of low student motivation (most pronounced at twelfth grade) other than suggesting that students do not receive individual scores. Further, the authors do not make inferences about the mechanism by which student motivation levels might affect performance. Once again, in the authors’ defense, these substantive issues are interesting for researchers but were outside the intended scope of this report.

The methods used to collect and analyze the data were rigorous and appropriate. The strategies used to present and summarize the descriptive data were simplistic but aligned with the report’s goal of being accessible to non-researchers.


In this two-part study, Stokes and Cao evaluated the relationship between students’ performance on the 2005 NAEP Reading Assessment and their responses to motivation-related items on the NAEP background questionnaire. In Part I, findings from a permutation-based Mantel test provide strong evidence that student motivation and performance are related, particularly that low student motivation is associated with worse performance ($p = ~0$). In Part II, the authors construct a Bayesian Item Response Theory model to assess the relationship between students’ motivation and their intention to respond to NAEP reading questions. They then use the model to compare the performance of high- and low-motivation students. Their analysis provides evidence of low student motivation on this particular NAEP administration. It also suggests that low-motivation students may have less intention to respond and perform worse than high-motivation students. Stokes and Cao caution that establishing a causal relationship between low motivation and poor performance would require two assumptions: that the NAEP background questionnaire is a valid indicator of student motivation and that motivation alone (not confounding variables) explain differences in student performance.

Technical Review

The design and analyses of this study are logical, rigorous, and sophisticated. The study tackles very important questions about student motivation for NAEP and how declines in scores should be interpreted in light of motivation levels. The authors present and effectively interpret results from statistical tests of the difference between the performance levels of high and low motivation students. In general, descriptive statistics would have made the findings accessible to a broader audience of readers, even if the statistical parameters do not map directly to the unadjusted means. This is especially true in the case of analyses that use the Mantel test, a somewhat uncommon test with less intuitive parameters.

The authors were careful to couch the results of this study with some important caveats. They remind the reader that the study groups were formed on the basis of just one questionnaire item about motivation for taking the NAEP test. Further, the authors caution that because the groups were formed by extant student characteristics and not experimentally, the inferences are more akin to correlation than causation. The authors reasonably conclude that the limitations of correlational studies such as these suggest that the field engage in more true intervention studies that study the effect of incentives on student motivation and performance.

It is uncertain whether the authors adjusted for the nesting of student data. Beyond this ambiguity, the main criticism from a methodological perspective is that the student groupings for comparison are based on student
responses to a single, albeit straightforward, questionnaire item. If responses to this item are not a reliable indicator of motivation, the entirety of the study findings is in question.

Futures studies of this type should use an index of both effort and efficacy questionnaire items to form student groups.


This observational study sought to determine whether nine educational and environmental factors previously linked to standardized test achievement also promote writing and students' valuing of writing. These factors were categorized into three groups: aptitude (e.g., motivation, prior achievement), instruction (e.g., quantity of instruction, quality of instruction), and psychological environment (e.g., classroom climate, television watching). Data on these factors were derived from the responses of 288 nationally representative twelfth graders on the NAEP Writing Assessment. Averaged ratings of student essays served as a proxy for students' writing performance. The mean correlation between motivation and achievement from 40 previous studies was .34; however, in this study, motivation was found to have zero (.00) correlation with performance. The authors speculate that this may have been, in part, due to the limited sample of writing; each averaged rating was treated as one item, whereas multiple-choice tests yield many items. Additionally, while their sample was nationally representative, it was considerably smaller than sample sizes from similar National Assessment studies, which generally used data from 1,000-3,000 students.

Technical Review

The authors provided helpful and encouraging reliability information (alpha = 0.85) for the motivation scale used in estimating the motivation-reading achievement relationship, giving readers more confidence in their estimate. It is unfortunate that the non-significant bivariate correlation estimated for this relationship prompted the researchers not to test the relationship in a multivariate regression framework. Dropping likely non-significant predictors from the regression can result in biased estimates of the remaining factors, places too much emphasis on arbitrary cutoffs for statistical significance, and withholds important information from the field.

It is unfortunate that the authors did not attempt to speculate on the implications of the non-significant motivation-reading achievement relationship, aside from noting that a positive relationship was observed in other studies. In the authors' defense, this study was attempting to estimate many relationships, so they had much to cover in their discussion section.

As noted, the primary methodological critique is the authors' removal on non-significant predictors.


For this observational study, Yepes-Baraya administered test blocks from the 1993 NAEP science field test to 16 eighth grade students in a suburban New Jersey middle school. Each student completed two test blocks: a hands-on task and either a conceptual/problem solving block or a theme block. In the weeks that followed, each student met with the study investigator, who asked them to read each test item aloud and talk about their thought process for producing an answer. The students' science instructors were also asked about their teaching and testing practices and the extent to which the assessment content had been covered in their respective classes. As the author had predicted, students' performance improved with increased perceived ability (p = .014, Pearson's; p = .048, Spearman). The correlation between standard block scores and perceived ability was .601 (Pearson's) and .280 (Spearman), respectively. As was also expected, performance decreased with increased perceived difficulty of the test blocks (p = .615, Pearson's; p = .293, Spearman) with correlations of .136 (Pearson's) and .280 (Spearman).
As the author acknowledges, the study’s small sample size (n = 16) precludes readers from generalizing its findings to a broader student population.

**Technical Review**

This small-scale validation study attempted to provide validation information about the cognitive processes used by respondents as they worked through the NAEP assessment, and to pilot the assessment of cognitive components beyond those in the 1993 NAEP science framework, including metacognitive skills and motivation. The author provided helpful descriptive statistics for all measured variables, as well as the correlation between scores on the motivation items and achievement scores. Although a correlation coefficient computed on such a small sample (n = 16), is suspect, the author was clear that the findings are not generalizable and should be interpreted with caution and with the study’s purpose in mind.

With regard to the correlation between motivation (expectancy) and achievement score, the author notes that the positive relationship observed in this study is consistent with a well-established literature base but cited no studies to support this claim. While this may be true, formally corroborating this finding would have been an easy way to bring credibility to the study. Otherwise, the author was appropriately cautious and transparent about the generalizability and ambiguity of the findings with regard to the relationship between motivation and achievement.

The methodology was appropriate for the limited scope and purpose of the study. Although this study does not assess student motivation on the NAEP test, the study collected information on science value and expectancy, which are fundamental components of motivation theory generally. Thus, this study was included in the literature review because motivation on science generally is related to whether students are motivated to take a science test.
Participation and Engagement of 12th-Graders Taking the Nation’s Report Card

The National Assessment of Educational Progress (NAEP) obtains an accurate portrait of student academic performance while taking relatively little time from students and schools, because NAEP assessments are administered to samples of schools and students, rather than to every school and student in the country. However, this approach requires the participation of those schools and students sampled in order for NAEP to accurately reflect the diversity of our nation's student population. It is also important that students selected as part of the NAEP sample, not only participate, but do their best on the assessment to demonstrate what they know and can do.

The National Center for Education Statistics (NCES) ensures that samples participating in NAEP assessments accurately represent the country by following statistical standards for participation rates. NCES also conducts research studies on the engagement of students taking NAEP. In this update for the Committee on Standards, Design and Methodology (COSDAM), a short review of previous studies as well as new evidence from the 2015 assessment relating to student participation and engagement will be presented.

This presentation will include trends in school participation rates and various measures of student engagement (for example, item response rates by item types, omit and nonresponse rates). It will also describe NCES’ efforts to increase 12th grade participation and engagement rates.
NAEP Academic Preparedness Research

Update on State Statistical Linking Studies with ACT and SAT

In this presentation, we will update the Committee on Standards, Design and Methodology (COSDAM) on the most recent statistical linking work, which is part of a second phase of academic preparedness research. The first phase of the National Assessment Governing Board’s statistical linking research, part of a broader academic preparedness research agenda, was based on 2009 data and included a national NAEP-SAT linking as well as in-depth linking and analysis of Florida’s longitudinal database. The second phase is based on 2013 data and includes several statistical linking studies at the state level that were performed via data sharing agreements.

At the August 2015 COSDAM meeting we discussed three state-level studies that focused on the extent to which 8th graders are on track for being academically prepared for college once they reach the end of high school. To that end, statistical linking studies between 8th grade NAEP (Reading and Mathematics) and EXPLORE®, a test developed and administered by ACT, Inc., were conducted. The EXPLORE® assessment was linked to performance on the ACT, and on-track preparedness benchmarks were established. The study was conducted in three states (Kentucky, North Carolina, and Tennessee), where EXPLORE® was administered to all students state-wide who were in grade 8 during the 2012-13 school year.

Concurrent with the grade 8 studies, three states in grade 12 (Massachusetts, Michigan, and Tennessee) also participated. Similar to grade 8, the ACT test was administered state-wide in Michigan and Tennessee and performance on the ACT Reading and Mathematics was linked in these two states to performance on NAEP to establish preparedness benchmarks. In addition, Massachusetts performance on the SAT Critical Reading and Mathematics was linked to performance on NAEP. The SAT is developed and administered by the College Board.

The grade 12 state-level statistical linking studies were designed to pursue the following analysis questions:

1) What are the correlations between grade 12 NAEP and ACT or SAT scores in Reading and Mathematics?
2) What scores on the grade 12 NAEP Reading and Mathematics scales correspond to the ACT college readiness and SAT benchmarks? And what scores on the ACT and SAT scales correspond to grade 12 NAEP Proficient cut scores?

In this session, Andreas Oranje from Educational Testing Service will present research findings from the 2013 grade 12 state statistical linking studies.

---

1 ACT discontinued the use of the EXPLORE® test after fall 2015 for existing users and no new users are now being accepted.
2 Beginning March 2016, College Board discontinued the use of the old SAT test and began to administer the revised SAT.
NAEP Grade 12 Academic Preparedness Research:
Establishing a Statistical Relationship between the NAEP and SAT Assessments in Reading and Mathematics for Grade 12 Massachusetts Students

Nuo Xi
Mei-Jang Lin
Laura Jerry
David Freund
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
Governing Board Project Officer: Sharyn Rosenberg,
Assistant Director for Psychometrics

Prepared by Educational Testing Service with funding by the National Assessment Governing Board, under National Center for Education Statistics contract ED-IES-13-C-0017, Task 9, Option 9(C).
Introduction

Starting in early 2003, the National Assessment Governing Board (Governing Board) embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on academic preparedness. The Governing Board’s working definition of academic preparedness for college is the knowledge and skills in reading and mathematics needed to qualify 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 institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research, scale scores of 302 on the NAEP grade 12 reading assessment (equivalent to the Proficient cut score) and 163 on the NAEP grade 12 mathematics assessment (between the Basic cut score of 141 and the Proficient cut score of 176) were identified to project a reasonable probability of being academically prepared for college. As a result, the percentage of 12th grade students in the U.S. who were academically prepared for college was estimated and reported for the 2013 and 2015 assessments in reading and mathematics.

Extensive details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (https://www.nagb.org/what-we-do/preparedness-research.html).

As part of the first phase of the Governing Board’s preparedness research, Florida participated in the research by providing (via a data sharing agreement) longitudinal data that could be linked to 2009 NAEP grade 12 performance in reading and mathematics. These data were a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

In the current (second) phase of the Governing Board’s academic preparedness research, additional state partners have agreed to provide longitudinal data that can be linked to the 2013 NAEP reading and mathematics assessments at grades 8 and 12. Massachusetts, as one of the state partners, participated in the state-level statistical linking research connecting NAEP and SAT and provided data on students who were part of the NAEP grade 12 sample during the 2012-2013 school year, as well as their SAT data. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.

Discussion Draft
Preparedness Technical Report

Grade 12 Massachusetts
In this report we will describe the NAEP and SAT assessments in (critical) reading and mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

**Linking Assessments**

*The SAT Assessment*

The SAT, owned and published by the College Board, is a college admission test widely used in the United States. Beginning March 2016, College Board started to administer a new SAT that is different from the one students took before (https://collegereadiness.collegeboard.org/sat). The following paragraphs describe the pre-March-2016 SAT, or the “old” SAT, administered in Massachusetts during the 2012-13 school year that was used in this study.

The SAT assessment is offered seven times a year, in October, November, December, January, March, May, and June. College Board states that the SAT tests students’ knowledge and skills in three subjects: critical reading, mathematics, and writing (https://sat.collegeboard.org/why-sat/topic/sat/what-the-sat-tests). The testing time and the number of items vary by subject. The critical reading section of SAT is made up of three multiple-choice sections, two of which are 25-minute sections and the other a 20-minute section. In total, there are 67 critical reading items in SAT. The mathematics section of SAT also contains two 25-minute sections and one 20-minute section. One of the 25-minute math sections contains 8 multiple-choice items and 10 grid-in items. The other two math sections are entirely multiple-choice. In total, there are 54 mathematics items. Each section of the SAT (critical reading, mathematics, and writing) is reported on a 200-to-800 scale, in 10-point increments, for a composite score ranging between 600 and 2400. In this study, only the critical reading and mathematics scores were used to link with the NAEP reading and mathematics assessments.

The SAT assessments were designed to measure a specific student’s skills and knowledge essential for college and career readiness and success (https://collegereadiness.collegeboard.org/about). To help inform the college and career readiness of groups of students, the College Board derived the SAT Benchmark through extensive research (The SAT® College and Career Readiness Benchmark User Guidelines, 2011). The SAT benchmarks were created to “establish a threshold for students that, if met, would ensure a reasonable probability of college success and eventual completion” (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011). Students who meet a benchmark on the SAT test have approximately a 65% chance of earning a first-year grade point average (FYGPA) of 2.67 (B-) or higher (Wyatt et al., 2011). The SAT benchmarks were 1550 for the composite and 500 for each section, i.e., critical reading, mathematics, and writing.
**The National Assessment of Educational Progress (NAEP)**

NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as reading, mathematics, and science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame in order to produce results that are nationally representative and also representative of participating states and urban districts. The NAEP test was administered to a representative sample of 12th graders in Massachusetts public schools during the 2012-2013 school year (with the testing window from the last week of January to the first week of March in 2013). Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 12 reading, each booklet contains two blocks of about 10 items each. For grade 12 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score can be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods (Mislevy, Beaton, Kaplan, & Sheehan, 1992). For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy et al. (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., SAT scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 12 reading and from 0 to 300 in grade 12 mathematics, and the goal is to express the aforementioned SAT benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one subject. Consequently, each student in the matched or linking sample had SAT scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

**Linking**

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some content overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong...
and the instruments have a similarly high reliability, we would be able to claim that the two scales are largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a ‘best’ projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., "at score level X, students have a reasonably high probability to be prepared"). In the case of the preparedness linking studies, and taking past studies into account (e.g., the Phase I preparedness research), a moderate relationship is most probable. We will elaborate further on this in subsequent sections.

Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. It serves as the foundation for most of the preparedness research, especially for the statistical relationship studies. The content alignment studies between NAEP and SAT critical reading and mathematics were conducted by WestEd in 2009, under contract ED-NAG-09-C-0001 with the National Assessment Governing Board. The studies found similar content in NAEP and SAT, and the content overlap was more extensive in mathematics than in reading (https://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

**Methodology**

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

**Data**

This study used data from students who were sampled and assessed in NAEP 12th grade reading or mathematics in 2013 and had also taken the SAT. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered. Thirteen states participated in the pilot state assessment at grade 12, including Massachusetts. About 2,400 public school students in Massachusetts were sampled for each subject. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The SAT is a widely used college admission test but not mandatory in Massachusetts, meaning that a group of self-selected 12th graders participated in SAT and have associated SAT scores. Compared to NAEP assessments, the SAT test is not sample-based and does not apply weights.

The process of matching SAT scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics.
(NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work. In addition, data confidentiality agreements were established between all parties involved and the Massachusetts Department of Education. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of state supplied scores (e.g., SAT scores) was assured through the assignment of a pseudo ID for students taking that assessment and using that pseudo ID as a way to transfer scores to ETS without the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again without any PII. Via the pseudo ID, ETS subsequently matched SAT scores to NAEP files. In the case of Massachusetts, SAT scores were matched at 74% for reading and 76% for mathematics. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) range between 46% and 84%. Notice that the variation in the matching rates across different student subgroups is partly due to the self-selectiveness nature of the SAT assessments. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates.

Table 1. Weighted percentages by gender and race of the Massachusetts linking samples

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian/Alaskan Native</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total^2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>35%</td>
<td>4%</td>
<td>3%</td>
<td>3%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>46%</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>39%</td>
<td>5%</td>
<td>5%</td>
<td>4%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>54%</td>
</tr>
<tr>
<td><strong>Total^2</strong></td>
<td>75%</td>
<td>8%</td>
<td>8%</td>
<td>7%</td>
<td>#</td>
<td>#</td>
<td>2%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Overall Match Rate 74%

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian/Alaskan Native</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total^2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>36%</td>
<td>4%</td>
<td>3%</td>
<td>3%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>47%</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>38%</td>
<td>5%</td>
<td>5%</td>
<td>3%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>53%</td>
</tr>
<tr>
<td><strong>Total^2</strong></td>
<td>74%</td>
<td>9%</td>
<td>9%</td>
<td>6%</td>
<td>#</td>
<td>#</td>
<td>2%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Overall Match Rate 76%

NOTES: 1# Rounds to zero.
2 Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a higher stakes test (i.e., SAT test) and very low on the lower stakes test, the low performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered ‘outliers’ and removed from further analyses. An initial examination of the joint distribution of NAEP and SAT revealed very few potential outlier cases. After this more cursory

Discussion Draft
Preparedness Technical Report

Grade 12 Massachusetts
inspection, standardized residuals from robust regression (Huber, 1973) were used to identify approximately 1.2% of cases in reading and approximately 1.1% of cases in mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). We refer to Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded from the final linking samples and were not used in subsequent analyses.

**Analysis Approach**

After preparatory data identification, matching, merging, and data reconciliation, the linking analyses were conducted. The current study was designed to pursue three specific analysis questions that guide the choices in methodology for the linking and validation:

1. What are the correlations between the grade 12 NAEP and SAT scores in reading and mathematics?
2. What scores on the grade 12 NAEP reading and mathematics scales correspond to the SAT benchmarks?
3. What are the average grade 12 NAEP reading and mathematics scores and IQRs (i.e., the difference between the 75th and 25th percentiles) for students below, at, and at or above the SAT benchmarks?

Questions 2) and 3) have been specified in one particular direction to estimate an academic preparedness cutpoint on the NAEP scale. Conversely and as a complement to these questions, the same analyses can be conducted in the opposite direction to verify: 2*) what scores on the SAT critical reading and mathematics scales correspond to the grade 12 NAEP Proficient cut scores in reading and mathematics and 3*) what the average SAT critical reading and mathematics scores and IQRs are for students below and at or above the NAEP Proficient cut scores.

We will describe pertinent methodological details about the analyses followed by the results of the analyses in the final section. The key steps of the analyses are (a) estimating the correlation between NAEP and SAT, which includes use of the aforementioned latent regression methodology (b) determining the appropriate methodology for linking based on those correlations and (c) applying the selected methodology to effectively estimate cumulative probability functions.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is that NAEP measures constructs that are represented on item response theory based latent scales, which are not measured reliably at the student level. However, pertinent data from students in specified groups of interest can be pooled to estimate reliable scores at the group level. SAT scores, on the other hand, are reliably estimated at the individual level and can be treated as a set of
consecutive (semi-continuous) groups. Correlations between NAEP and SAT can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is 0.74 and for mathematics is 0.89. While these are not low correlations, they do suggest that there is enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score $x$ on NAEP exactly corresponds to a score $y$ on SAT and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of $x$ on NAEP corresponds to the proportion $p$ of students attaining the benchmark score of $y$ or higher on SAT. Subsequently, a choice for $p$ has to be made, where a more conservative claim requires a higher $p$. This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high $p$ has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and SAT reading ($r = 0.74$) is not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is $r = 0.866$ (Dorans, 1999; Dorans & Walker, 2007). The correlation between NAEP and SAT mathematics ($r = 0.89$) met the minimum requirement of 0.866. However, given the very different assessment purposes of NAEP and SAT, as well as the low matching rates for certain reporting subgroups, it was decided to use projection for both reading and math in this study. Typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-SAT distributions\(^1\).

\(^1\) For reading, as part of the loglinear smoothing procedure we preserved the first 3 moments for the NAEP distribution, 4 moments for the SAT distribution, and 4 cross-moments. For math, we preserved the first 3 moments for the NAEP distribution, 4 moments for the SAT distribution, and 4 cross-moments. These loglinear
An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size\(^2\) in order to produce reliable comparisons. For the Massachusetts linking samples, both gender groups met that criterion. For the race/ethnicity groups, only White student subgroups met the criterion. Separate linking functions were established for these subgroups. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. The comparison results showed some variance across the three identified subgroups for reading but not for mathematics. In general, the linking functions for Male and White student subgroups were higher than the overall linking function, and the linking function for Female students was slightly lower than the overall linking function. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projection tables containing conditional cumulative distributions of NAEP proficiencies for SAT scores. The range of possible NAEP scores below, at, and at or above the SAT benchmark (500 on the SAT critical reading scale and 500 on the SAT mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the SAT benchmarks. In addition, the direction of the linking relationship was reversed and the point on the SAT measure that corresponds most closely to the NAEP Proficient cut score was identified using the conditional cumulative distributions of the SAT scores for the NAEP proficiencies. We will discuss the results of the linking study in the following section.

**Results**

**SAT benchmarks projected on the NAEP scale**

The second and third analysis questions ask what scores on the NAEP reading and mathematics scales correspond to the SAT benchmarks. In other words, what would be the scale score on NAEP that corresponds most reasonably to an established benchmark of academic preparedness for college (i.e., SAT).
Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located on the NAEP scales as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and at or above the SAT benchmarks are spread out, though more so for students below the benchmark than above. Note that the mean at the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

Table 2: Descriptive NAEP Statistics for Students Below, At, and At or Above the SAT Benchmarks

<table>
<thead>
<tr>
<th>Subject</th>
<th>SAT Benchmark</th>
<th>Mean</th>
<th>Percentage</th>
<th>SD</th>
<th>25th</th>
<th>75th</th>
<th>IQR1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Below</td>
<td>282</td>
<td>48%</td>
<td>29</td>
<td>263</td>
<td>301</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>At</td>
<td>304</td>
<td>4%</td>
<td>23</td>
<td>289</td>
<td>319</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>323</td>
<td>52%</td>
<td>27</td>
<td>304</td>
<td>340</td>
<td>36</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Below</td>
<td>147</td>
<td>42%</td>
<td>21</td>
<td>134</td>
<td>161</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>At</td>
<td>166</td>
<td>4%</td>
<td>13</td>
<td>157</td>
<td>175</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>187</td>
<td>58%</td>
<td>21</td>
<td>172</td>
<td>200</td>
<td>28</td>
</tr>
</tbody>
</table>

NOTES: 1IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.

To determine the NAEP scale score point that most reasonably corresponds to the SAT benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the SAT benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned previously, a proportion level has to be chosen commensurate with the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows above which point students are more likely to have reached the benchmark than not (i.e., the conditional proportion is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be academically prepared for college. For context, a secondary, light orange line indicates when the conditional proportion $p$ is set at 0.80, indicating a relatively high level of confidence that students have attained the SAT benchmark.

From the graphs it can be deduced that the location on the NAEP reading scale where students have a reasonable probability to be academically prepared for college could be at a NAEP scale score of 302, precisely the Proficient achievement level for NAEP reading at grade 12. The corresponding location on the NAEP math scale could be at 164, about 12 points below the Proficient achievement level for NAEP math at grade 12.
**Figure 1**: Proportion of students meeting the SAT critical reading benchmark of 500 in Massachusetts for NAEP reading scores

**Figure 2**: Proportion of students meeting the SAT mathematics benchmark of 500 in Massachusetts for NAEP mathematics scores
NAEP Proficient cut scores projected on the SAT scale

To conduct the complementing analyses, we find the point on the SAT measure that corresponds most closely to the NAEP Proficient cut score, essentially reversing the direction of the linking relative to the previous analyses. Table 3 provides descriptive statistics of the SAT critical reading and mathematics scores for students below and at or above the grade 12 NAEP Proficient achievement level. The grade 12 NAEP Proficient level cut score was set at 302 for reading and 176 for mathematics.

Table 3: Descriptive SAT Statistics for Students Below, and At or Above the Grade 12 NAEP Proficient Level.

<table>
<thead>
<tr>
<th>Subject</th>
<th>NAEP Proficient</th>
<th>Mean</th>
<th>Percentage</th>
<th>SD</th>
<th>25th</th>
<th>75th</th>
<th>IQR^1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Reading</td>
<td>Below</td>
<td>431</td>
<td>47%</td>
<td>89</td>
<td>370</td>
<td>490</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>565</td>
<td>53%</td>
<td>92</td>
<td>500</td>
<td>620</td>
<td>120</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Below</td>
<td>452</td>
<td>57%</td>
<td>78</td>
<td>400</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>610</td>
<td>43%</td>
<td>78</td>
<td>550</td>
<td>660</td>
<td>110</td>
</tr>
</tbody>
</table>

NOTES: ^1IQR is the Inter Quartile Range or the difference between the 75^th and 25^th percentiles.
^2The "At" category has fewer than 1% students due to the non-continuous nature of the reporting SAT scale score.

Following the same methodology of statistical projection (see Figures 3 and 4) we identified an SAT critical reading score of 490 and a mathematics score of 540 as cut points. The projected point for critical reading is close to the SAT benchmark, and about 40 scale score points higher than the SAT benchmark for mathematics.
Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Massachusetts for SAT critical reading scores

Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Massachusetts for SAT mathematics scores
Summary

The goal of this study was to statistically relate NAEP and SAT and use that relationship to identify a reference point or range on the NAEP 12th grade reading and mathematics scales reasonably associated with SAT benchmarks for critical reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 12th grade who are academically prepared for college for the nation and for states. The state of Massachusetts participated in this study and graciously provided the critical SAT data necessary to conduct the linking study with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for academic preparedness reporting (see Figures 1 and 2 for examples of how the markers were determined).

In addition, we identified the point on the SAT measure that corresponds most closely to the NAEP Proficient achievement level cut score, for grade 12 reading and mathematics scales, in order to explore the relationship between the two measures in the reverse direction (see Figures 3 and 4 for the linking results).

The relationship between NAEP reading and SAT critical reading is moderate (r=0.74), meaning that the kind of relational statements that can be made need to be presented in terms of probability rather than direct one-to-one relationships. The relationship between the two scales for math is quite strong (r=0.89), however, given the very different assessment purposes of NAEP and SAT, as well as the low matching rates for certain reporting subgroups, it was decided to use projection for both reading and math in this study. The results showed that the SAT benchmarks and the NAEP Proficient achievement level cut scores correspond well to each other for reading in both linking directions, but somewhat differ for mathematics. In particular, the NAEP reading Proficient achievement level cut score of 302 could form a reasonable basis for reporting on academic preparedness for college at grade 12 in Massachusetts, while the mathematics counterpart is 164 on the NAEP scale, about 12 points lower than the NAEP Proficient achievement level cut score for grade 12 math. On the other hand, the projection result of the NAEP Proficient reading cut score on the SAT scale is close to the existing SAT Benchmark for critical reading, and about 40 scale score points higher for mathematics.

As part of Phase II of the NAEP 12th grade preparedness research, the current study is closely related to the Phase I statistical linking study that connected NAEP and SAT on the national level (Moran, Oranje, & Freund, 2011). The national NAEP-SAT linking study used data from students who were sampled and assessed in NAEP 12th grade reading or math in 2009 and had also taken the SAT by June 2009. Based on the national linking sample, the correlation between scores on the two reading scales was 0.74, and the correlation was 0.91 between the two math scales. These numbers are very close to the correlations calculated in the current study. The projection results obtained from the national NAEP-SAT linking study (see Table 1 of Moran et al., 2011, p=0.5) also coincide with the
newly identified cutoff points on the NAEP scale for the Massachusetts linking sample, i.e., 302 for reading and 164 for math. The comparison results suggest that the statistical relationship between NAEP and SAT established for the Massachusetts linking sample surveyed in the 2013 NEAP assessment is very similar to that established with the 2009 NAEP-SAT linking samples on the national level.
References


The SAT® College and Career Readiness Benchmark User Guidelines
(http://media.collegeboard.com/digitalServices/pdf/sat/12b_6661_SAT_Benchmarks_PR_1
20914.pdf)

DISCUSSION DRAFT

NAEP Grade 12 Academic Preparedness Research:
Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Michigan Students

Nuo Xi
Mei-Jang Lin
Laura Jerry
David Freund
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
Governing Board Project Officer: Sharyn Rosenberg,
Assistant Director for Psychometrics

Prepared by Educational Testing Service with funding by the National Assessment Governing Board, under National Center for Education Statistics contract ED-IES-13-C-0017, Task 9, Option 9(C).

This research result used data collected and maintained by the Michigan Department of Education (MDE) and/or Michigan’s Center for Educational Performance and Information (CEPI). Results, information and opinions solely represent the analysis, information and opinions of the author(s) and are not endorsed by, or reflect the views or positions of, grantors, MDE and/or CEPI or any employee thereof.
Introduction

Starting in early 2003, the National Assessment Governing Board (Governing Board) embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on academic preparedness. The Governing Board’s working definition of academic preparedness for college is the knowledge and skills in reading and mathematics needed to qualify 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 institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research, scale scores of 302 on the NAEP grade 12 reading assessment (equivalent to the Proficient cut score) and 163 on the NAEP grade 12 mathematics assessment (between the Basic cut score of 141 and the Proficient cut score of 176) were identified to project a reasonable probability of being academically prepared for college. As a result, the percentage of 12th grade students in the U.S. who were academically prepared for college was estimated and reported for the 2013 and 2015 assessments in reading and mathematics. Extensive details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (https://www.nagb.org/what-we-do/preparedness-research.html).

As part of the first phase of the Governing Board’s preparedness research, Florida participated in the research by providing (via a data sharing agreement) longitudinal data that could be linked to 2009 NAEP grade 12 performance in reading and mathematics. These data were a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

In the current (second) phase of the Governing Board’s academic preparedness research, additional state partners have agreed to provide longitudinal data that can be linked to the 2013 NAEP reading and mathematics assessments at grades 8 and 12. Michigan, as one of the state partners, participated in the state-level statistical linking research connecting NAEP and ACT and provided data on students who were part of the NAEP grade 12 sample during the 2012-2013 school year, as well as their ACT data. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.
In this report we will describe the NAEP and ACT assessments in reading and mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

**Linking Assessments**

*The ACT Assessment*

As part of the Michigan Merit Examination (MME), the ACT® plus Writing\(^1\) was administered to almost all 11th graders in the spring of 2012. The ACT test is a curriculum- and standards-based assessment that measure students’ academic readiness for college ([https://www.act.org/aap/index.html](https://www.act.org/aap/index.html)). The assessment includes four multiple-choice tests. Each test measures student’s achievement in one of the following four areas: English, mathematics, reading, and science. The testing time and the number of items in the test vary by subject. For reading, students have 35 minutes to finish 40 multiple-choice items. For mathematics, the test has 60 multiple-choice items and students have 60 minutes to finish. A composite score is provided, which is calculated as the average of the four test scores. The individual test scores, as well as the composite score, range from 1 to 36 and are disseminated to students and schools directly. In this study, only the reading and mathematics scores were used to link with the NAEP reading and mathematics assessments.

The ACT tests were designed to measure students’ knowledge and skills needed for first-year college success. To help students translate test scores into a clear indicator of their current level of college readiness, ACT derived the ACT College Readiness Benchmarks based on a review of normative data, college admissions criteria, and information obtained through ACT's Course Placement Services. Students who meet a benchmark on the ACT test have approximately a 50% chance of obtaining a B or higher and approximately a 75% chance of obtaining a C or higher in the corresponding credit-bearing first-year college courses ([https://www.act.org/content/act/en/education-and-career-planning/college-and-career-readiness-standards/benchmarks.html](https://www.act.org/content/act/en/education-and-career-planning/college-and-career-readiness-standards/benchmarks.html)). The College Readiness Benchmarks for the ACT reading test is 22 and for the ACT mathematics is also 22 (ACT, 2013). These benchmarks were used in this investigation.

*The National Assessment of Educational Progress (NAEP)*

NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as reading, mathematics, and science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame

---

\(^1\) The ACT Writing test is a 40-minute essay test optional to the test takers. It is required as part of the MME in Michigan.
in order to produce results that are nationally representative and also representative of participating states and urban districts. The NAEP test was administered to a representative sample of 12th graders in Michigan public schools during the 2012-2013 school year (with the testing window from the last week of January to the first week of March in 2013). Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 12 reading, each booklet contains two blocks of about 10 items each. For grade 12 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score can be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods (Mislevy, Beaton, Kaplan, & Sheehan, 1992). For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy et al. (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., ACT scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 12 reading and from 0 to 300 in grade 12 mathematics, and the goal is to express the aforementioned ACT benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one subject. Consequently, each student in the matched or linking sample had ACT scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

**Linking**

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some content overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong and the instruments have a similarly high reliability, we would be able to claim that the two scales are largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a ‘best’ projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., "at score level X,
students have a reasonably high probability to be prepared). In the case of the preparedness linking studies, and taking past studies into account, a moderate relationship is most probable. We will elaborate further on this in subsequent sections.

Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. It serves as the foundation for most of the preparedness research, especially for the statistical relationship studies. The content alignment studies between NAEP and ACT reading and mathematics were conducted by ACT in 2009, under subtask 4.3 of contract ED-06-CO-0098 with the National Assessment Governing Board. The studies found similar content in NAEP and ACT, and the content overlap was more extensive in mathematics than in reading (https://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

**Methodology**

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

**Data**

This study used data from students who were sampled and assessed in NAEP 12th grade reading or mathematics in 2013 and had also taken the ACT. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered. Thirteen states participated in the pilot state assessment at grade 12, including Michigan. About 2,900 and 3,100 students at grade 12 were assessed in reading and mathematics, respectively, in Michigan. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The ACT assessment was required in Michigan at the 11th grade level and was offered as part of MME to eligible 12th graders, meaning that almost all students who were sampled for NAEP also participated in ACT and have associated scores. The reverse is not true, given that NAEP is sample-based (i.e., not every student who participated in ACT also participated in NAEP). Notice that the two tests were not administered concurrently. There could be a nine- to eleven-month time span between the state-wide ACT administration (spring of 2012) and the NAEP administration (first quarter of 2013).

The process of matching ACT scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics.
(NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work. In addition, data confidentiality agreements were established between all parties involved and the Michigan Department of Education. A process for matching the student records was developed to protect students’ identity and confidentiality. Confidentiality of state supplied scores (e.g., ACT scores) was assured through the assignment of a pseudo ID for students taking that assessment and using that pseudo ID as a way to transfer scores to ETS without the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again without any PII. Via the pseudo ID, ETS subsequently matched ACT scores to NAEP files. In the case of Michigan, ACT scores were matched at 95% for both reading and mathematics, which is very high. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) were at or above 88%. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates. That matched samples appear to be NAEP representative. In terms of ACT, the weighted average ACT reading and math scores of the matched sample are very close to the average ACT scores of the Michigan graduating class 2013, which are released in the ACT Profile Report (https://forms.act.org/newsroom/data/2013/pdf/profile/Michigan.pdf).

Table 1. Weighted percentages by gender and race of the Michigan linking samples

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38%</td>
<td>6%</td>
<td>2%</td>
<td>2%</td>
<td></td>
<td>#</td>
<td>1%</td>
<td>49%</td>
</tr>
<tr>
<td>Female</td>
<td>39%</td>
<td>7%</td>
<td>2%</td>
<td>1%</td>
<td></td>
<td>#</td>
<td>1%</td>
<td>51%</td>
</tr>
<tr>
<td>Total</td>
<td>77%</td>
<td>13%</td>
<td>5%</td>
<td>3%</td>
<td>1%</td>
<td>#</td>
<td>1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mathematics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>39%</td>
<td>6%</td>
<td>2%</td>
<td>1%</td>
<td></td>
<td>#</td>
<td>1%</td>
<td>50%</td>
</tr>
<tr>
<td>Female</td>
<td>38%</td>
<td>7%</td>
<td>2%</td>
<td>2%</td>
<td></td>
<td>#</td>
<td>1%</td>
<td>50%</td>
</tr>
<tr>
<td>Total</td>
<td>77%</td>
<td>13%</td>
<td>5%</td>
<td>3%</td>
<td>1%</td>
<td>#</td>
<td>1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Overall Match Rate 95%

NOTES: ¹# Rounds to zero.
²Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a higher stakes test (i.e., ACT test) and very low on the lower stakes test, the low performance can be reasonably attributed to motivation rather than performance level. Such cases
would be considered ‘outliers’ and removed from further analyses. An initial examination of the joint
distribution of NAEP and ACT revealed very few potential outlier cases. After this more cursory
inspection, standardized residuals from robust regression (Huber, 1973) were used to identify
approximately 0.8% of cases in reading and approximately 1.4% of cases in mathematics (cases with
absolute standardized residuals greater than 3 were considered outliers and removed). We refer to
Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded
from the final linking samples and were not used in subsequent analyses.

**Analysis Approach**

After preparatory data identification, matching, merging, and data reconciliation, the linking
analyses were conducted. The current study was designed to pursue three specific analysis
questions that guide the choices in methodology for the linking and validation:

1) What are the correlations between the grade 12 NAEP and ACT scores in reading and
   mathematics?

2) What scores on the grade 12 NAEP reading and mathematics scales correspond to the ACT
   benchmarks?

3) What are the average grade 12 NAEP reading and mathematics scores and IQRs (i.e., the
   difference between the 75th and 25th percentiles) for students below, at, and at or above the
   ACT benchmarks?

Questions 2) and 3) have been specified in one particular direction to estimate an academic
preparedness cutpoint on the NAEP scale. Conversely and as a complement to these questions, the
same analyses can be conducted in the opposite direction to verify: 2*) what scores on the ACT
reading and mathematics scales correspond to the grade 12 NAEP *Proficient* cut scores in reading
and mathematics and 3*) what the average ACT reading and mathematics scores and IQRs are for
students below and at or above the NAEP *Proficient* cut scores.

We will describe pertinent methodological details about the analyses followed by the results of the
analyses in the final section. The key steps of the analyses are (a) estimating the correlation between
NAEP and ACT, which includes use of the aforementioned latent regression methodology (b)
determining the appropriate methodology for linking based on those correlations and (c) applying
procedures to effectively estimate cumulative probability functions.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and
the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is
that NAEP measures constructs that are represented on item response theory based latent scales,
which are not measured reliably at the student level. However, pertinent data from students in
specified groups of interest can be pooled to estimate reliable scores at the group level. ACT scores,
on the other hand, are reliably estimated at the individual level and can be treated as a set of
consecutive (semi-continuous) groups. Correlations between NAEP and ACT can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is 0.73 and for mathematics is 0.83. While these are not low correlations, they do suggest that there is enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score \( x \) on NAEP exactly corresponds to a score \( y \) on ACT and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of \( x \) on NAEP corresponds to the proportion \( p \) of students attaining the benchmark score of \( y \) or higher on ACT. Subsequently, a choice for \( p \) has to be made, where a more conservative claim requires a higher \( p \). This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high \( p \) has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and ACT reading \((r=0.73)\) and mathematics \((r=0.83)\) are not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is \( r = 0.866 \) (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection was used in this study. Typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-ACT distributions.

---

2 Note that if the two assessments were administered closer to each other, the correlation might have been somewhat higher.

3 For reading, as part of the loglinear smoothing procedure we preserved the first 3 moments for the NAEP distribution, 6 moments for the ACT distribution, and 4 cross-moments. For math, we preserved the first 3 moments for the NAEP distribution, 5 moments for the ACT distribution, and 4 cross-moments. These loglinear smoothing models mostly resulted in the smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006), although model complexity and sample size was also taken into consideration.
An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size\(^4\) in order to produce reliable comparisons. For the Michigan linking samples, both gender groups met that criterion. For the race/ethnicity groups, only White student subgroup met the criterion. Separate linking functions were established for these subgroups. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. The comparison results showed some variance across the three identified subgroups for reading but not for mathematics. For reading, the linking functions for Male and White student subgroups were a little higher than the overall linking function, and the linking function for Female students was slightly lower than the overall linking function. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projection tables containing conditional cumulative distributions of NAEP proficiencies for ACT scores. The range of possible NAEP scores below, at, and at or above the ACT benchmark (22 on the ACT reading scale and 22 on the ACT mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the ACT benchmarks. In addition, the direction of the linking relationship was reversed and the point on the ACT measure that corresponds most closely to the NAEP Proficient cut score was identified using the conditional cumulative distributions of the ACT scores for the NAEP proficiencies. We will discuss the results of the linking study in the following section.

**Results**

**ACT benchmarks projected on the NAEP scale**

The second and third analysis questions ask what scores on the NAEP reading and mathematics scales correspond to the ACT benchmarks. In other words, what would be the scale score on NAEP that corresponds most reasonably to an established benchmark of academic preparedness for college (i.e., the ACT).

---

\(^4\) The minimum was set at 500 as a rule of thumb, but based on the idea that there is at least one observation below -3 and above +3 standard deviations (in a standard normal distribution) in expectation.
Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located on the NAEP scales as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and at or above the ACT benchmarks are spread out, though more so for students below the benchmark than above. Note that the mean at the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

### Table 2: Descriptive NAEP Statistics for Students Below, At, and At or Above the ACT Benchmarks

| Subject         | ACT Benchmark | Mean  | Percentage | SD | 25th | 75th | IQR
|-----------------|---------------|-------|------------|----|------|------|----
| Reading         | Below         | 275   | 64%        | 30 | 255  | 295  | 40 |
|                 | At            | 302   | 5%         | 24 | 286  | 318  | 32 |
|                 | At or Above   | 318   | 36%        | 26 | 300  | 335  | 35 |
| Mathematics     | Below         | 140   | 63%        | 23 | 126  | 156  | 30 |
|                 | At            | 168   | 5%         | 15 | 158  | 178  | 20 |
|                 | At or Above   | 184   | 37%        | 19 | 171  | 196  | 25 |

NOTES: 1IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.

To determine the NAEP scale score point that most reasonably corresponds to the ACT benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the ACT benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned previously, a proportion level has to be chosen commensurate with the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows above which point students are more likely to have reached the benchmark than not (i.e., the conditional proportion is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be academically prepared for college. For context, a secondary, light orange line indicates when the conditional proportion $p$ is set at 0.80, indicating a relatively high level of confidence that students have attained the ACT benchmark.

From the graphs it can be deduced that the location on the NAEP reading scale where students have a reasonable probability to be academically prepared for college could be at a NAEP scale score of 308, about 6 points above the *Proficient* achievement level for NAEP reading at grade 12. The corresponding location on the NAEP math scale could be at 169, about 7 points below the *Proficient* achievement level for NAEP mathematics at grade 12.
Figure 1: Proportion of students meeting the ACT reading benchmark of 22 in Michigan for NAEP reading scores

Figure 2: Proportion of students meeting the ACT mathematics benchmark of 22 in Michigan for NAEP mathematics scores
**NAEP Proficient cut scores projected on the ACT scale**

To conduct the complementing analyses, we find the point on the ACT measure that corresponds most closely to the NAEP *Proficient* cut score, essentially reversing the direction of the linking relative to the previous analyses. Table 3 provides descriptive statistics of the ACT reading and mathematics scores for students below and at or above the grade 12 NAEP *Proficient* achievement level. The grade 12 NAEP *Proficient* level cut score was set at 302 for reading and 176 for mathematics.

<table>
<thead>
<tr>
<th>Subject</th>
<th>NAEP Proficient</th>
<th>Mean</th>
<th>Percentage</th>
<th>SD</th>
<th>25th</th>
<th>75th</th>
<th>IQR¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td>Below</td>
<td>17</td>
<td>61%</td>
<td>4</td>
<td>13</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>At or Above²</td>
<td>25</td>
<td>39%</td>
<td>5</td>
<td>20</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td><strong>Mathematics</strong></td>
<td>Below</td>
<td>18</td>
<td>72%</td>
<td>3</td>
<td>15</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>26</td>
<td>28%</td>
<td>4</td>
<td>23</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>

NOTES: ¹IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles. ²The "At" category has fewer than 1% students due to the non-continuous nature of the reporting ACT scale score.

Following the same methodology of statistical projection (see Figures 3 and 4) we identified an ACT reading score of 22, identical with the ACT benchmark, and a mathematics score of 24 as cut points.
Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Michigan for ACT reading scores

Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Michigan for ACT mathematics scores
Summary

The goal of this study was to statistically relate NAEP and ACT and use that relationship to identify a reference point or range on the NAEP 12th grade reading and mathematics scales reasonably associated with ACT benchmarks for reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 12th grade who are academically prepared for college for the nation and for states. The state of Michigan participated in this study and graciously provided the critical ACT data necessary to conduct the linking study with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for academic preparedness reporting (see Figures 1 and 2 for examples of how the markers were determined).

In addition, we identified the point on the ACT measure that corresponds most closely to the NAEP Proficient achievement level cut score, for grade 12 reading and mathematics scales, in order to explore the relationship between the two measures in the reverse direction (see Figures 3 and 4 for the linking results).

A key finding was that the relationship between the two scales is moderate, meaning that the kind of relational statements that can be made need to be presented in terms of probability rather than direct one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct. In addition, in Michigan the grade 12 NAEP assessment was administered almost a year later than the state-wide ACT administration, making interpretation somewhat more challenging. The results showed that, in the state of Michigan, the ACT College Readiness Benchmarks and the NAEP Proficient achievement level cut scores correspond well to each other for reading in both linking directions, but slightly differ for mathematics. In particular, the NAEP reading scale score of 308 could form a reasonable basis for reporting on academic preparedness for college, while the mathematics counterpart is 169 on the NAEP scale. On the other hand, the projection result of the NAEP Proficient reading cut score on the ACT scale coincides with the existing ACT College Readiness Benchmark for reading, and about 2 points higher than the ACT benchmark for mathematics. To what extent these results generalize to other states or the nation is an empirical question.
References


NAEP Grade 12 Academic Preparedness Research:
Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Tennessee Students

Nuo Xi
Mei-Jang Lin
Laura Jerry
David Freund
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
Governing Board Project Officer: Sharyn Rosenberg,
Assistant Director for Psychometrics

Prepared by Educational Testing Service with funding by the National Assessment Governing Board, under National Center for Education Statistics contract ED-IES-13-C-0017, Task 9, Option 9(C).
Introduction

Starting in early 2003, the National Assessment Governing Board (Governing Board) embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on academic preparedness. The Governing Board’s working definition of academic preparedness for college is the knowledge and skills in reading and mathematics needed to qualify 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 institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research, scale scores of 302 on the NAEP grade 12 reading assessment (equivalent to the Proficient cut score) and 163 on the NAEP grade 12 mathematics assessment (between the Basic cut score of 141 and the Proficient cut score of 176) were identified to project a reasonable probability of being academically prepared for college. As a result, the percentage of 12th grade students in the U.S. who were academically prepared for college was estimated and reported for the 2013 and 2015 assessments in reading and mathematics. Extensive details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (https://www.nagb.org/what-we-do/preparedness-research.html).

As part of the first phase of the Governing Board’s preparedness research, Florida participated in the research by providing (via a data sharing agreement) longitudinal data that could be linked to 2009 NAEP grade 12 performance in reading and mathematics. These data were a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

In the current (second) phase of the Governing Board’s academic preparedness research, additional state partners have agreed to provide longitudinal data that can be linked to the 2013 NAEP reading and mathematics assessments at grades 8 and 12. Tennessee, as one of the state partners, participated in the state-level statistical linking research connecting NAEP and ACT and provided data on students who were part of the NAEP grade 12 sample during the 2012-2013 school year, as well as their ACT data. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.
In this report we will describe the NAEP and ACT assessments in reading and mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

**Linking Assessments**

*The ACT Assessment*

The ACT® test was administered to almost all 11th graders in Tennessee in the spring of 2012. It is a curriculum- and standards-based assessment that measure students’ academic readiness for college (https://www.act.org/aap/index.html). The assessment includes four multiple-choice tests. Each test measures student’s achievement in one of the following four areas: English, mathematics, reading, and science. The testing time and the number of items in the test vary by subject. For reading, students have 35 minutes to finish 40 multiple-choice items. For math, the test has 60 multiple-choice items, and students have 60 minutes to finish. A composite score is provided, which is calculated as the average of the four test scores. The individual test scores, as well as the composite score, range from 1 to 36 and are disseminated to students and schools directly. In this study, only the reading and mathematics scores were used to link with the NAEP reading and mathematics assessments.

The ACT tests were designed to measure students’ knowledge and skills needed for first-year college success. To help students translate test scores into a clear indicator of their current level of college readiness, ACT derived the ACT College Readiness Benchmarks based on a review of normative data, college admissions criteria, and information obtained through ACT’s Course Placement Services. Students who meet a benchmark on the ACT test have approximately a 50% chance of obtaining a B or higher and approximately a 75% chance of obtaining a C or higher in the corresponding credit-bearing first-year college courses (https://www.act.org/content/act/en/education-and-career-planning/college-and-career-readiness-standards/benchmarks.html). The College Readiness Benchmarks for the ACT reading test is 22 and for the ACT mathematics is also 22 (ACT, 2013). These benchmarks were used in this investigation.

*The National Assessment of Educational Progress (NAEP)*

NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as reading, mathematics, and science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame in order to produce results that are nationally representative and also representative of participating states and urban districts. The NAEP test was administered to a representative sample of 12th graders in Tennessee public schools during the 2012-2013 school year (with the testing
window from the last week of January to the first week of March in 2013). Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 12 reading, each booklet contains two blocks of about 10 items each. For grade 12 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score can be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods (Mislevy, Beaton, Kaplan, & Sheehan, 1992). For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy et al. (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., ACT scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 12 reading and from 0 to 300 in grade 12 mathematics, and the goal is to express the aforementioned ACT benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one subject. Consequently, each student in the matched or linking sample had ACT scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

**Linking**

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some content overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong and the instruments have a similarly high reliability, we would be able to claim that the two scales are largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a ‘best’ projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., “at score level X, students have a reasonably high probability to be prepared”). In the case of the preparedness linking studies, and taking past studies into account, a moderate relationship is most probable. We will elaborate further on this in subsequent sections.
Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. It serves as the foundation for most of the preparedness research, especially for the statistical relationship studies. The content alignment studies between NAEP and ACT reading and mathematics were conducted by ACT in 2009, under subtask 4.3 of contract ED-06-CO-0098 with the National Assessment Governing Board. The studies found similar content in NAEP and ACT, and the content overlap was more extensive in mathematics than in reading (https://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

Methodology

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

Data

This study used data from students who were sampled and assessed in NAEP 12th grade reading or mathematics in 2013 and had also taken the ACT. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered. Thirteen states participated in the pilot state assessment at grade 12, including Tennessee. About 3,000 and 3,200 students at grade 12 were assessed in reading and mathematics, respectively, in Tennessee. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The ACT assessment was required in Tennessee at the 11th grade level, meaning that almost all students who were sampled for NAEP also participated in ACT and have associated scores. The reverse is not true, given that NAEP is sample-based (i.e., not every student who participated in ACT also participated in NAEP). Notice that the two tests were not administered concurrently. There could be a nine- to eleven-month time span between the statewide ACT administration (spring of 2012) and the NAEP administration (first quarter of 2013).

The process of matching ACT scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics (NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work. In addition, data confidentiality agreements were established between all parties involved and the Tennessee Department of Education. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of state supplied scores (e.g., ACT scores) was assured through the assignment of a pseudo ID for students taking that assessment and

Discussion Draft
Preparedness Technical Report
Grade 12 Tennessee
using that pseudo ID as a way to transfer scores to ETS without the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again without any PII. Via the pseudo ID, ETS subsequently matched ACT scores to NAEP files. In the case of Tennessee, ACT scores were matched at 89% for reading and 90% for mathematics, which is very high. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) were at or above 81%. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates. The matched samples appear to be NAEP representative. In terms of ACT, the weighted average ACT reading and math scores of the matched sample are very close to the average ACT scores of the Tennessee graduating class 2013, which are released in the ACT Profile Report (https://forms.act.org/newsroom/data/2013/pdf/profile/Tennessee.pdf).

Table 1. Weighted percentages by gender and race of the Tennessee linking samples

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian/Alaskan Native</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>36%</td>
<td>10%</td>
<td>2%</td>
<td>1%</td>
<td>#¹</td>
<td>#</td>
<td>#</td>
<td>49%</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>35%</td>
<td>13%</td>
<td>2%</td>
<td>1%</td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>51%</td>
</tr>
<tr>
<td><strong>Total²</strong></td>
<td>70%</td>
<td>23%</td>
<td>4%</td>
<td>2%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Overall Match Rate** 89%

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>American Indian/Alaskan Native</th>
<th>Pacific Islander</th>
<th>2+ races</th>
<th>Total²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>36%</td>
<td>10%</td>
<td>2%</td>
<td>1%</td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>35%</td>
<td>12%</td>
<td>2%</td>
<td>1%</td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Total²</strong></td>
<td>71%</td>
<td>22%</td>
<td>4%</td>
<td>2%</td>
<td>#</td>
<td>#</td>
<td>1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Overall Match Rate** 90%

NOTES: ¹# Rounds to zero.
²Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a higher stakes test (i.e., ACT test) and very low on the lower stakes test, the low performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered ‘outliers’ and removed from further analyses. An initial examination of the joint distribution of NAEP and ACT revealed very few potential outlier cases. After this more cursory inspection, standardized residuals from robust regression (Huber, 1973) were used to identify
approximately 1.3% of cases in reading and approximately 1.4% of cases in mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). We refer to Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded from the final linking samples and were not used in subsequent analyses.

**Analysis Approach**

After preparatory data identification, matching, merging, and data reconciliation, the linking analyses were conducted. The current study was designed to pursue three specific analysis questions that guide the choices in methodology for the linking and validation:

1) **What are the correlations between the grade 12 NAEP and ACT scores in reading and mathematics?**

2) **What scores on the grade 12 NAEP reading and mathematics scales correspond to the ACT benchmarks?**

3) **What are the average grade 12 NAEP reading and mathematics scores and IQRs (i.e., the difference between the 75th and 25th percentiles) for students below, at, and at or above the ACT benchmarks?**

Questions 2) and 3) have been specified in one particular direction to estimate an academic preparedness cutpoint on the NAEP scale. Conversely and as a complement to these questions, the same analyses can be conducted in the opposite direction to verify: 2*) what scores on the ACT reading and mathematics scales correspond to the grade 12 NAEP Proficient cut scores in reading and mathematics and 3*) what the average ACT reading and mathematics scores and IQRs are for students below and at or above the NAEP Proficient cut scores.

We will describe pertinent methodological details about the analyses followed by the results of the analyses in the final section. The key steps of the analyses are (a) estimating the correlation between NAEP and ACT, which includes use of the aforementioned latent regression methodology (b) determining the appropriate methodology for linking based on those correlations and (c) applying procedures to effectively estimate cumulative probability functions.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is that NAEP measures constructs that are represented on item response theory based latent scales, which are not measured reliably at the student level. However, pertinent data from students in specified groups of interest can be pooled to estimate reliable scores at the group level. ACT scores, on the other hand, are reliably estimated at the individual level and can be treated as a set of consecutive (semi-continuous) groups. Correlations between NAEP and ACT can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is 0.73 and for mathematics is 0.83. While these are not low correlations, they do suggest that there is
enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score \( x \) on NAEP exactly corresponds to a score \( y \) on ACT and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of \( x \) on NAEP corresponds to the proportion \( p \) of students attaining the benchmark score of \( y \) or higher on ACT. Subsequently, a choice for \( p \) has to be made, where a more conservative claim requires a higher \( p \). This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high \( p \) has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and ACT reading (\( r = 0.73 \)) and mathematics (\( r = 0.83 \)) are not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is \( r = 0.866 \) (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection was used in this study. Typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-ACT distributions.

An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student

---

1 Note that if the two assessments were administered closer to each other, the correlation might have been somewhat higher.

2 For reading, as part of the loglinear smoothing procedure we preserved the first 3 moments for the NAEP distribution, 4 moments for the ACT distribution, and 4 cross-moments. For math, we preserved the first 2 moments for the NAEP distribution, 5 moments for the ACT distribution, and 4 cross-moments. These loglinear smoothing models mostly resulted in the smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006), although model complexity and sample size was also taken into consideration.
groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size in order to produce reliable comparisons. For the Tennessee linking samples, both gender groups met that criterion. For the race/ethnicity groups, only White student subgroup met the criterion. Separate linking functions were established for these subgroups. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. For NAEP reading, the linking functions for Male and White student subgroups were slightly higher than the overall linking function, and the linking function was slightly lower for Female student subgroup. For NAEP math, no substantial deviation from the overall linking function was detected for White student subgroup. The linking function for Female student subgroup was slightly higher than the overall linking function, and it was slightly lower for Male student subgroup. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projection tables containing conditional cumulative distributions of NAEP proficiencies for ACT scores. The range of possible NAEP scores below, at, and at or above the ACT benchmark (22 on the ACT reading scale and 22 on the ACT mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the ACT benchmarks. In addition, the direction of the linking relationship was reversed and the point on the ACT measure that corresponds most closely to the NAEP Proficient cut score was identified using the conditional cumulative distributions of the ACT scores for the NAEP proficiencies. We will discuss the results of the linking study in the following section.

**Results**

**ACT benchmarks projected on the NAEP scale**

The second and third analysis questions ask what scores on the NAEP reading and mathematics scales correspond to the ACT benchmarks. In other words, what would be the scale score on NAEP that corresponds most reasonably to an established benchmark of academic preparedness for college (i.e., the ACT).

---

3 The minimum was set at 500 as a rule of thumb, but based on the idea that there is at least one observation below -3 and above +3 standard deviations (in a standard normal distribution) in expectation.
Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located on the NAEP scales as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and at or above the ACT benchmarks are spread out, though more so for students below the benchmark than above. Note that the mean at the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

Table 2: Descriptive NAEP Statistics for Students Below, At, and At or Above the ACT Benchmarks

<table>
<thead>
<tr>
<th>Subject</th>
<th>ACT Benchmark</th>
<th>Mean</th>
<th>Percentage</th>
<th>SD</th>
<th>25th</th>
<th>75th</th>
<th>IQR¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Below</td>
<td>269</td>
<td>63%</td>
<td>29</td>
<td>250</td>
<td>289</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>At</td>
<td>295</td>
<td>5%</td>
<td>23</td>
<td>280</td>
<td>310</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>311</td>
<td>37%</td>
<td>25</td>
<td>294</td>
<td>328</td>
<td>34</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Below</td>
<td>135</td>
<td>73%</td>
<td>23</td>
<td>120</td>
<td>151</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>At</td>
<td>164</td>
<td>4%</td>
<td>14</td>
<td>154</td>
<td>173</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>181</td>
<td>27%</td>
<td>18</td>
<td>167</td>
<td>192</td>
<td>25</td>
</tr>
</tbody>
</table>

NOTES: ¹IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.

To determine the NAEP scale score point that most reasonably corresponds to the ACT benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the ACT benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned previously, a proportion level has to be chosen commensurate with the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows above which point students are more likely to have reached the benchmark than not (i.e., the conditional proportion is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be academically prepared for college. For context, a secondary, light orange line indicates when the conditional proportion p is set at 0.80, indicating a relatively high level of confidence that students have attained the ACT benchmark.

From the graphs it can be deduced that the location on the NAEP reading scale students have a reasonable probability to be academically prepared for college could be at a NAEP scale score of 301, slightly lower than the Proficient achievement level. The corresponding location on the NAEP math scale could be at 168, about 8 points below the Proficient achievement level.
Figure 1: Proportion of students meeting the ACT reading benchmark of 22 in Tennessee for NAEP reading scores

Figure 2: Proportion of students meeting the ACT mathematics benchmark of 22 in Tennessee for NAEP mathematics scores
NAEP Proficient cut scores projected on the ACT scale

To conduct the complementing analyses, we find the point on the ACT measure that corresponds most closely to the NAEP Proficient cut score, essentially reversing the direction of the linking relative to the previous analyses. Table 3 provides descriptive statistics of the ACT reading and mathematics scores for students below and at or above the grade 12 NAEP Proficient achievement level. The grade 12 NAEP Proficient level cut score was set at 302 for reading and 176 for mathematics.

Table 3: Descriptive ACT Statistics for Students Below, and At or Above the Grade 12 NAEP Proficient Level.

<table>
<thead>
<tr>
<th>Subject</th>
<th>NAEP Proficient</th>
<th>Mean</th>
<th>Percentage</th>
<th>SD</th>
<th>Percentile 25th</th>
<th>Percentile 75th</th>
<th>IQR¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Below</td>
<td>18</td>
<td>68%</td>
<td>5</td>
<td>14</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>At or Above²</td>
<td>25</td>
<td>32%</td>
<td>5</td>
<td>21</td>
<td>28</td>
<td>7</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Below</td>
<td>18</td>
<td>82%</td>
<td>3</td>
<td>15</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>At or Above</td>
<td>26</td>
<td>18%</td>
<td>4</td>
<td>23</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>

NOTES: ¹IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
²The “At” category has fewer than 1% students due to the non-continuous nature of the reporting ACT scale scores.

Following the same methodology of statistical projection (see Figures 3 and 4) we identified an ACT reading score of 23 and a mathematics score of 25 as cut points. The cut points are about 1 and 3 points higher than the ACT benchmarks for reading and mathematics tests, respectively, for grade 12 students.
Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Tennessee for ACT reading scores

Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Tennessee for ACT mathematics scores
Summary

The goal of this study was to statistically relate NAEP and ACT and use that relationship to identify a reference point or range on the NAEP 12th grade reading and mathematics scales reasonably associated with ACT benchmarks for reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 12th grade who are academically prepared for college for the nation and for states. The state of Tennessee participated in this study and graciously provided the critical ACT data necessary to conduct the linking study with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for academic preparedness reporting (see Figures 1 and 2 for examples of how the markers were determined).

In addition, we identified the point on the ACT measure that corresponds most closely to the NAEP Proficient achievement level cut score, for grade 12 reading and mathematics scales, in order to explore the relationship between the two measures in the reverse direction (see Figures 3 and 4 for the linking results).

A key finding was that the relationship between the two scales is moderate, meaning that the kind of relational statements that can be made need to be presented in terms of probability rather than direct one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct. In addition, in Tennessee the grade 12 NAEP assessment was administered almost a year later than the state-wide ACT administration, making interpretation somewhat more challenging. The results showed that, in the state of Tennessee, the ACT College Readiness Benchmarks and the NAEP Proficient achievement level cut scores correspond well to each other for reading in both linking directions (i.e., the projection results are 1 scale score point different from the ACT benchmark/NAEP Proficient level), but differ more for mathematics. In particular, the NAEP reading scale score of 301 could form a reasonable basis for reporting on academia preparedness for college, while the mathematics counterpart is 168 on the NAEP scale. On the other hand, the projection result of the NAEP Proficient reading cut score on the ACT scale is close to the existing ACT College Readiness Benchmark for reading, and about 3 points higher for mathematics. To what extent these results generalize to other states or the nation is an empirical question.
References


NAEP Academic Preparedness Research: Planned Additional Analyses

In addition to the academic preparedness research studies that have been presented to COSDAM, future analyses using 2013 NAEP Reading and Mathematics assessments will include a national NAEP-ACT linking study and longitudinal studies in grades 8 and 12. Brief overviews are provided for each study:

**National Linking Study with the ACT**

The Governing Board is partnering with ACT, Inc. to conduct a statistical linking study at the national level between NAEP and the ACT in Reading and Mathematics. Through a procedure that protects student confidentiality, the ACT 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 was performed with the SAT in 2009. There will not be a national statistical linking study performed for NAEP and the SAT in 2013.

**Research Questions for National and State Statistical Linking Studies with the ACT:**

1. What are the correlations between the grade 12 NAEP and ACT student score distributions in Reading and Math?
2. What scores on the grade 12 NAEP Reading and Math scales correspond to the ACT college readiness benchmarks? (concordance and/or projection)
3. What scores on the ACT scales correspond to the grade 12 NAEP Reading and Math Proficient cut scores? (concordance and/or projection)
4. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students below, at, and at or above the ACT college readiness benchmarks?
5. What are the average ACT scores and interquartile ranges (IQR) for students below, at, and at or above the grade 12 NAEP Reading and Math Proficient cut scores?
6. Do the results differ by race/ethnicity or gender?

**Longitudinal Statistical Relationships: Grade 8 NAEP**

Using a procedure that protects student confidentiality, secondary and postsecondary data for 2013 NAEP 8th grade test takers in the state samples in North Carolina and Tennessee will be linked to NAEP scores. These studies will examine the relationship between 8th grade NAEP scores and scores on state tests, future ACT scores, placement into remedial versus credit-bearing courses, and first-year college GPA.
Research Questions for Longitudinal Statistical Relationships, Grade 8 NAEP:

1. What is the relationship between NAEP Reading and Math scores at grade 8 and state test scores at grade 4?
2. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students below the ACT benchmarks at grade 11/12? At or above the ACT benchmarks?
3. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students who are placed in remedial and non-remedial courses in college?
4. What are the average NAEP Reading and Math scores (and the IQR) at grade 8 for students who obtain a first-year college GPA of B- or above?

Longitudinal Statistical Relationships: Grade 12 NAEP

In addition to the linking of ACT scores to NAEP 12th grade test scores in partner states, the postsecondary activities of NAEP 12th grade test takers will be followed for up to six years using the state longitudinal databases in Massachusetts, Michigan, and Tennessee. These studies will examine the relationship between 12th grade NAEP scores and scores on placement tests, placement into remedial versus credit-bearing courses, GPA, and persistence.

Research Questions for Longitudinal Statistical Relationships, Grade 12 NAEP:

1. What is the relationship between grade 12 NAEP Reading and Math scores and grade 8 state test scores?
2. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students with placement in remedial and non-remedial courses?
3. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students with a first-year GPA of B- or above?
4. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who remain in college after each year?
5. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who graduate from college within 6 years?
2017 Writing Grade 4 Achievement Levels Setting Contract

The 2017 NAEP writing assessment is the first administration of the grade 4 assessment under the current computer-based Writing Framework (https://www.nagb.org/publications/frameworks/writing/2017-writing-framework.html)\(^1\).

Pursuant to the Governing Board’s legislative mandate, achievement levels must be set for the grade 4 writing assessment. In accordance with the Board policy on setting performance levels for NAEP, the achievement levels setting process includes achievement levels descriptions (ALDs), cut scores, and exemplar items. In 2012, the Board formally approved the updated achievement levels descriptions for writing at all three grade levels. A procurement was issued in March 2016 for a contractor to design and implement studies to recommend cut scores and exemplar items at grade 4.

The 2017 grade 4 writing achievement levels setting will include a field trial (to test logistics associated with any software used to conduct the process), a pilot study, and an operational achievement levels setting study. In addition, the design procedures will require the collection of multiple sources of validity evidence. COSDAM will receive briefings and have the opportunity to provide input on the process throughout the life of the project, with Board action on the grade 4 writing achievement levels planned for the May 2018 Governing Board meeting.

We anticipate awarding the contract shortly before the August 2016 Governing Board meeting. Sharyn Rosenberg of the Governing Board staff will provide an overview of the contract, including key staff, tasks, and milestones.

---

\(^1\) In 2011, NAEP writing assessments were administered at grades 8 and 12 under the current Writing Framework, and achievement levels were set for grades 8 and 12. The grade 4 assessment initially was planned for 2013 administration but was postponed to 2017 due to budgetary constraints.
Lessons Learned from Research on Academic Preparedness for Job Training

For more than a decade, the Governing Board has been working on improving the form, function, and use of NAEP as an indicator of 12th graders’ academic preparedness for postsecondary endeavors. During the May 2016 plenary session, Board members were briefed on the purpose, history, major milestones, and current status of the Board’s preparedness research program.

Between 2005 and 2010, the Governing Board made the following decisions in implementing the preparedness research program:

- The term “academic preparedness” was used rather than “readiness” to indicate that NAEP was not intending to measure other characteristics needed for success in postsecondary endeavors beyond academic knowledge and skills.

- Academic preparedness for college, job training, and the military were not assumed to be the same; separate research strands were pursued for each outcome.

- The working definition of academic preparedness for job training programs refers to the reading and mathematics knowledge and skills needed to qualify for a job training program without remediation in mathematics or reading.

- To operationalize job training programs, five exemplar occupations were selected for use in research studies: Automotive Master Technician; Computer Support Specialist; Heating, Ventilation and Air Conditioning Technician (HVAC); Licensed Practical Nurse (LPNs); and Pharmacy Technician. The exemplar occupations were selected to represent jobs that do not require a 4-year degree and to represent job training programs that require equivalent reading and mathematics knowledge and skills to qualify for entry in both the military and civilian sectors.

Between 2010 and 2015, the Board’s research on using NAEP for academic preparedness for job training programs has included content alignment studies and judgmental standard setting studies. The findings have been inconclusive, largely due to huge variability in the knowledge and skills required by different training programs within a single occupation, let alone across the five exemplar occupations. No work is currently underway for academic preparedness for job training; in 2015, the Governing Board released a summary report of lessons learned (attached).

Michelle Blair of the Governing Board staff will provide an overview of lessons learned from the Board’s extensive research on academic preparedness for job training programs.

Discussion Questions

Should the Governing Board continue to pursue the use of NAEP as an indicator of academic preparedness for job training? If so, what aspects of the original approach should be revisited? What new approaches should be considered?
The National Assessment of Educational Progress (NAEP)

Research on Academic Preparedness for Job Training Programs
National Assessment Governing Board

BOARD MEMBERSHIP (2014-2015)

Terry Mazany  
*President and CEO*  
The Chicago Community Trust  
Chicago, Illinois

Susan Pimentel  
*Vice Chair*  
Educational Consultant  
Hanover, New Hampshire

Terry Mazany  
*President and CEO*  
The Chicago Community Trust  
Chicago, Illinois

Susan Pimentel  
*Vice Chair*  
Educational Consultant  
Hanover, New Hampshire

Honorable Mitchell Chester  
*Commissioner*  
Massachusetts Department of Elementary and Secondary Education  
Malden, Massachusetts

Lucille E. Davy  
*President and CEO*  
Transformative Education Solutions, LLC  
Pennington, New Jersey

Frank Fernandes  
*Principal*  
Kaimuki Middle School  
Honolulu, Hawaii

Honorable Anitere Flores  
*Legislator*  
Florida House of Representatives  
Miami, Florida

Rebecca Gagnon  
*Director*  
Minneapolis Board of Education  
Minneapolis, Minnesota

Shannon Garrison  
*Fourth-Grade Teacher*  
Solano Avenue Elementary School  
Los Angeles, California

Honorable James Geringer  
*Director of Policy and Public Sector Strategies*  
Environmental Systems Research Institute (ESRI)  
Cheyenne, Wyoming

Doris Hicks  
*Principal and CEO*  
Dr. Martin Luther King, Jr. Charter School for Science and Technology  
New Orleans, Louisiana

Honorable Anitere Flores  
*Legislator*  
Florida House of Representatives  
Miami, Florida

Andrew Ho  
*Professor*  
Harvard Graduate School of Education  
Cambridge, Massachusetts

Tonya Mathews  
*President and CEO*  
Michigan Science Center  
Detroit, Michigan

Tonya Miles  
*General Public Representative*  
Mitchellville, Maryland

Honorable Ronnie Musgrove  
*Of Counsel*  
Copeland, Cook, Taylor & Bush, P.A.  
Madison, Mississippi

Dale Nowlin  
*Teacher and Mathematics Department Chair*  
Bartholomew Consolidated School Corporation  
Columbus, Indiana

Father Joseph O’Keefe, S.J.  
*Professor*  
Boston College Lynch School of Education  
Chestnut Hill, Massachusetts

W. James Popham  
*Emeritus Professor*  
University of California, Los Angeles  
Los Angeles, California

B. Fielding Rolston  
*Chairman*  
Tennessee State Board of Education  
Kingsport, Tennessee

Cary Sneider  
*Associate Research Professor*  
Portland State University  
Portland, Oregon

Chasidy White  
*Teacher*  
Brookwood Middle School  
Brookwood, Alabama

Ex-officio Member  
Ruth Neild  
*Director*  
Institute of Education Sciences  
U.S. Department of Education  
Washington, DC

www.nagb.org | www.nagb.gov

Suggested Citation:

145
The National Assessment of Educational Progress (NAEP)

Research on Academic Preparedness For Job Training Programs

Developed for
The National Assessment Governing Board
under contract ED-NAG-11-0005
by Widmeyer Communications/A Finn Partners Company

September 2015
Acknowledgements

The National Assessment Governing Board would like to thank the following individuals and organizations, whose many contributions were instrumental in the 12th grade preparedness research effort and/or the development of this report:

- The National Assessment Governing Board panel of Technical Advisors on 12th Grade Preparedness Research
- The NAEP 12th Grade Preparedness Commission
- ACT, Inc.
- The College Board
- The Florida Department of Education
- Kristopher Kaase
- Widmeyer Communications
I. Introduction

Are the nation’s 12th graders prepared academically for college and job training? The National Assessment Governing Board has been conducting research for more than a decade to determine the potential of the National Assessment of Educational Progress (NAEP) of Reading and Mathematics at Grade 12 to answer this question. The Governing Board’s hope was that NAEP could serve as an indicator of academic preparedness for college and job training. This report provides a summary of the Governing Board’s groundbreaking job training preparedness research.

Measuring achievement at grade 12 is important because it is the end point of mandatory schooling for most students and the start of postsecondary education and training for most adults. However, most standardized tests taken by high school students are taken before 12th grade and are not representative of all students across the nation. NAEP is the only source of nationally representative, 12th grade student achievement results.

The Governing Board commissioned more than 30 research studies to find out if the Grade 12 NAEP could serve as an indicator of students’ academic preparedness for college and job training. The research results support the claim that 12th grade NAEP assessments of reading and mathematics are indicators of academic preparedness for college.

Concurrent with the research on whether NAEP could serve as an indicator of students’ academic preparedness for college, several of the studies commissioned by the Governing Board focused on whether NAEP could serve as an indicator of students’ academic preparedness for job training. This research included:

1. content alignment studies between NAEP and the ACT WorkKeys assessments;
2. comparisons between NAEP and training performance requirements for five exemplar occupations using performance requirements from the U.S. Department of Labor’s occupational information network, or O*NET;
3. a judgmental standard setting study conducted to identify NAEP scale scores at grade 12 representing the knowledge and skills in reading and mathematics needed to qualify for entry into job training programs in five exemplar professions, and
4. a course content analysis study to examine whether NAEP knowledge, skills, and abilities are prerequisite for entering into a job training program in five exemplar professions.

At this time the research results do not support the claim that NAEP Mathematics and Reading at Grade 12 data are indicators of academic preparedness for job training.
Because of the importance of this research, the Governing Board pursued it even though there is no common definition of what is required to prepare high school students for job training, and there is no common process for preparing students for job training. The research highlighted that the knowledge, skills, and abilities required for job training vary widely across occupations. In addition, job training program instructors indicated there is wide variability in job training programs across and within occupations.

The purpose of this report is to summarize the context, methodology, results, and conclusions of the Governing Board’s job training preparedness research studies for NAEP. This report is written for educators, policymakers, researchers, and interested members of the general public who are not assessment experts. Therefore, this report is not intended to provide the full details of each study. For those who would like to review the studies and their results in more detail, links and references to the individual research study reports are provided.
II. The Context for Preparedness Research

The environment for post-secondary education and training is diverse. No single way exists to prepare for college or job training, and post-secondary education and training is provided by a wide array of public, private, and proprietary organizations. When the Governing Board began this initiative in 2004, defining the boundaries for this work was important.

Defining Preparedness

Because NAEP is designed to measure reading and mathematics knowledge and skills, the focus of NAEP is academic preparedness for college or job training, rather than preparedness or readiness in general, which might include important, but non-academic skills such as persistence, time management, teamwork, conflict resolution, and adaptability.

The Governing Board has generally defined preparedness as the academic knowledge and skill levels in reading and mathematics necessary to be qualified for placement into a job training program (for the workplace context) or 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 (for the college context).

For NAEP context, preparedness for job training requires that a student has the reading and mathematics knowledge and skills sufficient to qualify for placement into a job training program. There are a variety of entry points into job training, including apprenticeship programs, community college technical certificates and job training programs, on-the-job training programs, and vocational institute or certification programs.

Additional Research Assumptions

As part of defining the boundaries for this work, the Governing Board made the following assumptions:

Preparedness relates to eligibility rather than success. Preparedness does not mean success in postsecondary job training.

Preparedness relates to qualification to enter rather than being hired for a job. Preparedness for job training refers to the reading and mathematics knowledge and skills needed to qualify for job training; it does not mean that a student is ready to be hired for a job.
Preparedness for civilian job training relates to parallel military jobs. To extend research findings to the military sector, a key assumption is that similar jobs in both the military and civilian sectors require approximately equal reading and mathematics knowledge and skills to qualify for entry.

Multiple research studies and methods should be used. No one study could comprehensively address the feasibility and validity of using NAEP Grade 12 as a measure of academic preparedness for college and job training—including whether the same NAEP content applies to both. Multiple studies and methods should be conducted to see whether there is convergence or divergence of results, and to use these patterns to determine what, if any, valid conclusions can be drawn.
III. Methodology

In determining if NAEP Grade 12 could serve as an indicator of students’ academic preparedness for job training, the Governing Board sought input from a variety of experts, which led to development of a research plan of conducting multiple research studies using multiple methods. The academic preparedness for job training research is organized into three types of studies.

1. Content alignment. These studies are designed to determine the extent to which NAEP and another test measure similar content.

2. Criterion-based judgmental standard setting. These studies are designed to identify NAEP scores at the 12th-grade level representing the knowledge and skills in reading and mathematics needed to qualify for job training programs in five exemplar occupations.

3. Course content analyses. These studies examine whether NAEP knowledge, skills, and abilities are prerequisite for entering into a job training program.

Five Exemplar Occupations

A group of technical experts identified a number of challenges with attempting to use NAEP as a measure of academic preparedness for job training (see Technical Panel on 12th Grade Preparedness Research: Final Report). Among the challenges identified were:

- The wide variety of paths into job training include on-the-job training, in-house training programs, formal apprenticeship programs, training programs in a community college, or training in vocational institutes or programs.

- Although a number of resources exist for identifying knowledge and skills required to qualify for a job, there is very little information on the knowledge and skills to enter training for a job.

- Few occupations have a nationally consistent core knowledge and skills training. Without a nationally consistent expectation for training in an occupation, it is not possible to report on academic preparedness for that occupation in a way that would be meaningful to everyone across the country.

- Some occupations emphasize certain skills (e.g., simple numerical calculations) to the near exclusion of others (e.g., algebra, geometry). Because NAEP assesses comprehensively for a domain (reading or mathematics), using the overall NAEP results for a domain may not provide meaningful information on preparedness for some occupations that only emphasize a subset of the domain assessed by NAEP.
• Equivalence between similar occupations in the military and civilian sectors cannot be assumed. Equivalence of jobs and job training for similar occupations in the military and civilian sectors needs to be confirmed because of the different environments in these job sectors.

To address these challenges, the technical experts recommended selecting exemplar occupations that best represent the entry-level reading and mathematics requirements for multiple sectors of the labor force. The technical experts also recommended a multi-step process for identifying these exemplar occupations. This process excluded occupations that require a bachelor’s degree, although some occupations may require a year or more of training. The Governing Board hired a contractor to conduct the identification process, which resulted in the selection of the following five exemplar occupations (see Identification of Exemplar Occupations – Report, Appendix A, and Appendix B).

### Overview of Types of Research and Studies

To date the following research studies of NAEP as an indicator of academic preparedness for job training have been conducted, which are presented in the table below.

<table>
<thead>
<tr>
<th>Type of Research Study</th>
<th>Status</th>
<th>Reports</th>
</tr>
</thead>
</table>
| Content alignment      | Five studies conducted* | [The Alignment of the NAEP Grade 12 Mathematics Assessment and the WorkKeys Applied Mathematics Assessment](#)  
[The Alignment of the NAEP Grade 12 Reading Assessment and the WorkKeys Reading for Information Assessment](#)  
[The Content Alignment between the NAEP and WorkKeys Assessments](#)  
[Comparisons between NAEP and O*NET on Academic Preparedness for Job Training for Five Target Occupations](#) |
| Criterion-based judgmental standard setting | Two studies conducted | [The Standard for Minimal Academic Preparedness in Mathematics to Enter a Job-Training Program](#)  
[The Standard for Minimal Academic Preparedness in Reading to Enter a Job-Training Program](#) |
| Course content analyses | One study conducted | [Job Training Programs Curriculum Study](#) |

* The report The Content Alignment between the NAEP and WorkKeys Assessments included both reading and mathematics studies.
1. Automotive Master Technician
2. Computer Support Specialist
3. Heating, Ventilation and Air Conditioning (HVAC) Technician
4. Licensed Practical Nurse (LPN)
5. Pharmacy Technician

These five occupations were the focus of studies of content alignment, criterion-based judgmental standard setting, and course content analyses.

In addition to these studies, the Governing Board convened a 10-person technical advisory panel to consider the research conducted to-date, produce ideas for future work, and to provide input on whether the Governing Board should continue to perform research on using NAEP as an indicator of academic preparedness for job training programs (see NAEP Technical Advisory Panel Proceedings of the Symposium on Academic Preparedness Research for more discussion on the challenge of accessing assessments related to job training.)

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple-choice test administered by the United States Military Entrance Processing Command used to determine qualification for enlistment in the United States Armed Forces. It is often offered to U.S. high school students when they are in grade 10, 11, and 12, and it is available to anyone eligible for enlistment. The needed partnerships for NAEP research with ASVAB were not available to the Governing Board when the first phase of the NAEP Preparedness Research Program was being planned and implemented. Hence, statistical linking of NAEP with ASVAB was not possible.

No benchmarking studies, which would involve administering NAEP at grade 12 to a reference group of interest (e.g., military recruits, job trainees), have been conducted. To date, the Governing Board has not successfully established the partnerships that would make a benchmarking study possible.
IV. Results

The Governing Board’s research was designed to explore the question, “Can NAEP Reading and Mathematics at Grade 12 serve as an indicator of academic preparedness for job training?” The results of each of the studies that attempted to answer this question are summarized below. More detailed information about each study and the results can be found by accessing the links provided to the full reports.

Content Alignment

Content alignment between the NAEP and WorkKeys assessments. The WorkKeys assessment is a widely recognized, standardized test related to the workplace created by the ACT. While most content alignment studies examine the alignment of an assessment to a corresponding set of standards, a 2010 study examined the alignment of the NAEP assessment to the WorkKeys assessment.

The findings from the alignment study of the NAEP Grade 12 Mathematics Assessment and the WorkKeys Applied Mathematics Assessment found:

- The WorkKeys Applied Mathematics items that most frequently aligned to the NAEP mathematics standards were related to problem-solving applications of number operations and measurement.
- The WorkKeys Applied Mathematics items do not assess content in the NAEP mathematics standards related to geometry, data analysis, statistics, probability, and algebra.
- The NAEP mathematics items that aligned to the WorkKeys Applied Mathematics standards include geometry content; fractions, ratios, percentages, or mixed numbers; and basic statistical concepts.
- The NAEP mathematics items either infrequently or do not assess at all content in the WorkKeys Applied Mathematics standards related to conversions, determining the best deal, finding errors, and calculating discounts or markups.
- There is content represented by the NAEP mathematics standards that is not covered by the WorkKeys Applied Mathematics assessment, and there is content represented by the WorkKeys Applied Mathematics standards that is not covered by the NAEP mathematics assessment.

The findings from the Alignment Study of the NAEP Grade 12 Reading Assessment and the WorkKeys Reading for Information Assessment found:

- The WorkKeys Reading for Information items that most frequently aligned to the NAEP reading standards were related to information processing and understanding.
- The WorkKeys Reading for Information items do not assess content in the NAEP reading standards related to higher-order thinking skills.
- The NAEP reading items that aligned to the WorkKeys Reading for Information standards include reading comprehension and information processing.
- The NAEP reading items either infrequently or do not assess at all content in the WorkKeys Reading for Information standards related to critical thinking, decision making, and problem solving.
- There is content represented by the NAEP reading standards that is not covered by the WorkKeys Reading for Information assessment, and there is content represented by the WorkKeys Reading for Information standards that is not covered by the NAEP reading assessment.
• The WorkKeys Reading for Information items that aligned to the NAEP reading standards were related to locating and recalling information, causal relations, connecting ideas, drawing conclusions, providing supporting information, and determining word meaning in context.

• The WorkKeys Reading for Information items do not assess content in the NAEP reading standards related to literary reading passages and critiquing or evaluating reading passages.

• The NAEP reading items that aligned to the WorkKeys Reading for Information standards include identifying main ideas, determining word meaning from context, explaining the rationale behind a text, and identifying implied details.

• The NAEP reading items do not assess content in the WorkKeys Reading for Information standards related to understanding, following, and applying instructions; determining and applying general principles contained in workplace documents and applying them to similar and new situations; and to the decoding of workplace jargon.

• Skills measured by both assessments include identifying main ideas, details, and definitions; determining the correct meaning of a word based on context; explaining the rationale of a document; and identifying implied details.

• There is content represented by the NAEP reading standards that is not covered by the WorkKeys Reading for Information assessment, and there is content represented by the WorkKeys Reading for Information standards that is not covered by the NAEP reading assessment.

Content Comparisons Made between NAEP and WorkKeys

Mathematics

• NAEP Grade 12 Mathematics items and WorkKeys Applied Mathematics standards

• NAEP Grade 12 Mathematics standards and WorkKeys Applied Mathematics items

• NAEP Grade 8 and Grade 12 Mathematics Frameworks to WorkKeys cognitive targets for Applied Mathematics and Applied Technology

• NAEP Grade 8 and Grade 12 Mathematics items to WorkKeys cognitive targets for Applied Mathematics and Applied Technology

• NAEP Grade 8 and Grade 12 Mathematics Frameworks to WorkKeys items for Applied Mathematics and Applied Technology

• NAEP Grade 12 Mathematics items and WorkKeys Applied Mathematics standards

• NAEP Grade 12 Mathematics standards and WorkKeys Applied Mathematics items

Reading

• NAEP Grade 12 Reading items and WorkKeys Reading for Information standards

• NAEP Grade 12 Reading standards and WorkKeys Reading for Information items

• NAEP Grade 8 and Grade 12 Reading items to WorkKeys cognitive targets for Reading for Information and Locating Information

• NAEP Grade 8 and Grade 12 Reading Frameworks to WorkKeys items for Reading for Information and Locating Information

• NAEP Grade 8 and Grade 12 Reading Frameworks to WorkKeys cognitive targets for Reading for Information and Locating Information
A 2014 content alignment study examined similarities and overlap in the content and cognitive complexity between NAEP and WorkKeys. This study also included the NAEP grade 8 assessments and frameworks because experts have suggested that NAEP grade 8 may provide a better match to the academic content expectations of job training programs (Kilpatrick, 2012; Loomis, 2012). This study also included WorkKeys assessments for Applied Technology and Locating Information. The major findings from this study were:

- NAEP items do not adequately represent the WorkKeys content domain, as evidenced by the percentages of WorkKeys’ mathematics and reading cognitive targets (52% and 72%, respectively) that were not matched to any NAEP item.

- Sixteen of the 24 (67%) content strands within the NAEP Mathematics Framework and one of the three (33%) cognitive targets within the NAEP Reading Framework were not matched to any WorkKeys item.

- A direct comparison of the content frameworks for the two assessments indicated that the majority of the elements of the NAEP Mathematics Framework, WorkKeys math targets, and WorkKeys applied technology cognitive targets reflected unique content. Unique mathematics elements were calculated for Grade 12 NAEP Math Framework (85%), Grade 8 NAEP Mathematics Framework (75%), WorkKeys math cognitive targets (61%), and WorkKeys applied technology cognitive targets (100%). Unique reading elements included grade 8 and 12 NAEP informational reading framework (50%), WorkKeys reading cognitive targets (46%), and WorkKeys locating information cognitive targets (50%).

Comparisons Between NAEP and O*NET on Academic Preparedness for Job Training for Five Target Occupations. This study identified grade 8 and grade 12 NAEP content that is relevant to training performance requirements for each of the five target occupations (i.e., the exemplar occupations described in the Methodology section), and, conversely, the training performance requirements that are relevant to NAEP content. The job training content was based on performance requirements adapted from O*NET, the U.S. Department of Labor’s occupational information network. The study also compared the levels of knowledge, skills, and abilities (KSAs) needed for proficiency on NAEP reading and mathematics with the levels of KSAs needed for entry into job training. The KSAs included in this study were a subset of KSAs identified as academically relevant by occupational experts from the O*NET covering reading and mathematical related skills (e.g., written comprehension, mathematical reasoning, critical thinking, complex problem solving, deductive reasoning, etc.). The major findings from this study were:

Mathematics
- The NAEP mathematics objectives most relevant to job training content were the objectives associated with the number properties and operations content area and the measurement content area (except for Computer Support Specialists). This was true for both grade 8 and grade 12 NAEP.
• The NAEP mathematics objectives that were least relevant to job training content were the objectives associated with geometry (except for HVAC) and algebra (except for LPNs). This was true for both grade 8 and grade 12 NAEP.

**Reading**
• The NAEP reading objectives most relevant to job training content are the objectives associated with the locate/recall cognitive target for NAEP informational reading.
• The NAEP reading objectives that were least relevant to job training content were the objectives associated with the critique/evaluate cognitive target.

**Mathematics and Reading**
• The range of mathematics and reading skills required by NAEP (both grade 8 and grade 12) is broader than the range of mathematics and reading skills required by job training.
• The percentage of the NAEP mathematics objectives linked to job training requirements for specific occupations decreased considerably from grade 8 to grade 12, indicating that as the complexity of the NAEP objectives increased from grade 8 to grade 12, their relevance to job training decreased. A comparable statement about whether including grade 8 reading resulted in more linked content is not possible because the NAEP reading objectives are the same for grade 8 and for grade 12. (The differentiation at grade 12 relates to the type of texts.)
• Disconnects were found between the levels of KSAs required for proficient performance on NAEP and the levels of KSAs required for entry into job training such that higher levels of the KSAs were required in the NAEP assessments than for job training. The largest disconnects occurred between grade 12 NAEP mathematics and job training. Disconnects also occurred between grade 12 reading and job training. The disconnects in required levels of KSAs tended to be smaller when comparing grade 8 content to job training content, particularly for grade 8 reading, which demonstrated several “matches” with KSA levels for training content (most notably with written comprehension).

The results from the content alignment between the NAEP and WorkKeys assessments and the comparisons between NAEP and O*NET on academic preparedness for job training for five target occupations do not support using NAEP to make judgments about the academic preparedness of 12th grade students to enter job training. These studies indicate that NAEP content covers a much wider domain of reading and mathematics than an assessment of job skills (WorkKeys), and the level of KSAs required for NAEP are higher than the KSAs needed for job training.
Criterion-Based Judgmental Standard Setting

A judgmental standard setting study was conducted to identify grade 12 NAEP scores representing the knowledge and skills in reading and mathematics needed to qualify for job training programs in the five exemplar occupations. Panels of subject matter experts from across the country met to review the NAEP test and determine the minimal level of academic performance on NAEP that demonstrates preparedness for entry into a job training program, as well as for placement in an entry-level credit-bearing college course without need for remediation.

The major findings from the criterion-based standard setting study were:

Mathematics

- Job-training groups struggled to find the mathematics they valued in either the framework or the test items. Because NAEP is more oriented toward pure mathematics than applied mathematics, much of the mathematics at grade 12 is well beyond what job-training groups would expect.

- The areas of number properties and operations and of measurement were the most important content areas for every occupational group, but these areas receive the least emphasis in the NAEP test. Job-training groups all wanted incoming students to know operations with fractions, decimals, and percents and their properties, which are addressed in the NAEP grade 8 objectives.

Reading

- Little agreement was found between job-training and college-entry panelists on the reading knowledge and skills required of students (2 of 25 or 8%). The two reading skills job-training and college-entry panelists agreed on were 1) identify main idea/key concepts/important information and 2) draw conclusions within/across texts. There were two other reading skills with which two of the occupational areas (computer support specialist and LPN) agreed with college-entry panelists: 1) interpret text, and 2) provide evidence to support an interpretation.

- Job-training panelists judged 11 (44%) of the reading skills as required of students for job training, while college-entry panelists did not judge these skills as required. In addition, there were 10 (40%) reading skills which job-training panelists did not rate as required for entry into job training that college-entry panelists rated as required.

The results from this criterion-based judgmental standard setting study do not support using NAEP to make judgments about the academic preparedness of 12th grade students to enter job training. Job-training panelists identified many NAEP 12th grade items they deemed as not required for determining academic preparedness for their job training programs.

In addition, the data collected from the job-training and college-entry panelists do not support the conclusion that minimal academic preparedness for college is the same as minimal academic preparedness for training programs for the five exemplar occupations that were examined. This research indicated the need to determine the prerequisite knowledge, skills, and abilities in reading and mathematics to qualify for placement into entry-level credit-bearing college courses and for job training programs, which led to the course content analyses.
Course Content Analyses

The Job Training Programs Curriculum Study examined course materials from job training programs for the five exemplar occupations. The study objectives were to identify the knowledge, skills, and abilities (KSAs) that are prerequisite and then to compare these prerequisite KSAs with NAEP frameworks and items and with the KSAs identified in the judgmental standard setting study. The major findings from this study were:

Mathematics

- The job training programs studied have few prerequisite expectations represented in the Grade 12 NAEP Mathematics Framework. The largest number of prerequisites across all occupational training programs are found in the number properties and operations domain, specifically: the systems of measurement; variables, expressions, and operations; and equations and inequalities standards.

- The portions of the NAEP mathematics KSA statements that were identified as inapplicable or excluded from the training course content prerequisites, eliminated much of the complex mathematics knowledge and skills that differentiate the grades 8 and 12 frameworks. As a result, some prerequisite KSAs appear to be better described by the grade 8 objectives.

- Many NAEP items at grade 12 were deemed not required for determining academic preparedness for job training programs. Between 64% and 78% of the 130 mathematics objectives were not evident as prerequisite in any course within the five occupations.

Reading

- Across all job training programs, the only grade 12 NAEP reading objectives identified as prerequisites for entry-level courses in all five occupational areas were those related to reading informational texts. Specific reading skills that are prerequisite to all five job training programs include locate or recall causal relations and locate or recall organizing structures of texts, such as comparison/contrast, problem/solution, enumeration, etc.

- The number of reading objectives not evident as prerequisite in any course within the five occupations ranged between 16% and 68% of the 37 objectives.

Mathematics and Reading

- The job-training course prerequisite knowledge, skills, and abilities identified are largely included in the Grade 12 NAEP Frameworks, but the full content of NAEP frameworks is much larger and broader.

The results from the course content analyses do not support using NAEP to make judgments about the academic preparedness of U.S. 12th grade students to enter job training. The NAEP 12th grade frameworks include much more knowledge, skills, and abilities than the job-training course prerequisite knowledge, skills, and abilities.
V. Summary of Findings

After this groundbreaking effort to explore if NAEP could report on preparedness for job training, the Governing Board asked, “What overall conclusions can be made about the NAEP Reading and Mathematics at Grade 12 serving as an indicator of academic preparedness for job training?” Several clear themes emerged from the research studies.

**NAEP’s content coverage is broader than the content covered in job training contexts.** The content alignment study of NAEP and the WorkKeys assessment found that the NAEP items do not adequately represent the WorkKeys content domain. The comparison of NAEP to relevant training performance requirements for each of the five exemplar occupations found the range of reading and mathematics skills required by NAEP (both grade 8 and grade 12) is broader than the range of reading and mathematics skills required by job training. In addition, the levels of knowledge, skills, and abilities (KSAs) required for NAEP were higher than the levels of KSAs required for entry into job training. The job-training panelists in the judgmental standard setting agreed that less than half of the NAEP mathematics and reading content was relevant to preparedness for their programs. Finally, the analysis of job-training course content found that the NAEP frameworks are much larger and deeper than the prerequisite KSAs for job-training.

**Across occupational fields, there is disagreement on which content is important for job training preparedness.** In mathematics, the five exemplar occupations aligned on the importance of number properties and operations followed by measurement. The occupational areas had much less agreement on the other areas of mathematics. In reading, the five exemplar occupations agreed on the importance of understanding vocabulary, identifying important information, summarizing, integrating information within/across texts, drawing conclusions, and applying information to new contexts. Beyond these skills, there was little or no agreement on other skills such as analyzing information, interpreting text, or providing evidence to support an interpretation.

**Within an occupational field, there is disagreement on which content is important for job training preparedness.** Even in occupational fields that have a more common core of training, such as automotive master technicians and LPNs, there is still not agreement on the required content to be prepared for job training. The discrepancies are even greater in fields where there is less of a common core of training (computer support specialists, pharmacy technicians).
A NAEP job training preparedness indicator for the NAEP reading and math assessments is unlikely at this time. Part of the purpose in conducting multiple research studies using multiple methods was to determine if there was mutually confirming evidence. The Governing Board’s interest was whether, when examining these research results in their totality there was: (1) convergence across the two academic preparedness areas (college and job training), or (2) convergence within each academic preparedness area.

First, based on the results and summary above, it is clear that there are wide differences in the required knowledge, skills, and abilities for entry into job training as measured on a standardized measure of job skills, an analysis of relevant job skills, judgment by occupational experts, and analysis of job-training course content as compared to the NAEP frameworks and assessments, which are much wider and deeper. The results indicate no definitive evidence that the academic qualifications needed for job training preparedness and the academic qualifications needed for college preparedness are the same; that is, there is, to date, no convergence across the two academic preparedness areas.

Second, with regard to the convergence of evidence within each academic area, to date, convergence has emerged only for using 12th grade NAEP as an indicator of academic preparedness for college (see Towards The National Assessment of Educational Progress (NAEP) as an Indicator of Academic Preparedness for College and Job Training). Given the evidence compiled to date for academic preparedness for job training, it is unlikely that NAEP will be able to report an indicator for job training academic preparedness for the NAEP mathematics or reading assessments.
VI. Conclusion

The Governing Board began a journey over ten years ago to answer the question of, “Can NAEP Reading and Mathematics at Grade 12 serve as an indicator of academic preparedness for college and job training?” As a part of that question, the Governing Board also sought to find out if NAEP might provide (1) a single indicator of academic preparedness across college and job training, or (2) separate indicators of academic preparedness for college and for job training. Based on more than 30 studies conducted at the direction of the Governing Board answers to this question are emerging.

The evidence to date indicates that 12th grade NAEP can arguably serve as an indicator of academic preparedness for college. The evidence to date does not support using at grade NAEP as an indicator of academic preparedness for job training. An important benefit of this research is the confirming evidence across research studies that there are wide differences in the required knowledge, skills, and abilities for entry into job training as compared to the required knowledge, skills, and abilities for entry into college.

What is next? Although the research findings to date have not supported the establishment of a NAEP academic preparedness for job training indicator, the lessons learned from this research can inform possible future research. Using a subset of the content covered by the grade 12 NAEP as a measure of academic preparedness for job training might be explored. Agreements with partners such as employers, the U.S. Department of Labor, or others may provide the data for statistical linking or benchmarking studies that have not been possible to date.

The Governing Board will consider the lessons learned from this research as they determine the next phases of the academic preparedness research.


WestEd & The Educational Policy Improvement Center. (2013). *National Assessment of Educational Progress grade 12 preparedness research project job training programs curriculum study*. San Francisco, CA, and Eugene, OR: Authors.
# National Assessment Governing Board
## Reporting and Dissemination Committee
### August 4 - 5, 2016

## AGENDA

<table>
<thead>
<tr>
<th>Thursday, August 4</th>
<th>Exploring Best Practices in Innovative Reporting</th>
<th>Attachment A</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:30 pm – 4:00 pm</td>
<td>David Stewart, Founder and CEO, Tembo</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Friday, August 5</th>
<th>Welcome, Introductions, and Agenda Overview</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30 – 10:35 am</td>
<td>Rebecca Gagnon, Chair</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Reporting Updates</th>
<th>Attachment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:35 – 11:05 am</td>
<td>Stephaan Harris, Public Affairs Specialist</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Laura LoGerfo, Assistant Director for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reporting and Analysis</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11:05 am – 12:00 pm</th>
<th>Revising Board Reporting Policy</th>
<th>Attachment C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Laura LoGerfo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dan McGrath, National Center for Education</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistics</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12:00 – 12:15 pm</th>
<th>Information Items:</th>
<th>Attachment D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Media Coverage of 2016 Releases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Progress on Procurements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Schedule of NAEP Releases</td>
<td></td>
</tr>
</tbody>
</table>
Exploring Best Practices in Innovative Reporting

Tembo
Tembo was founded in 2010 in response to a need for better data management, analysis, and visualization services in the education sector. The organization’s mission is to help state departments of education, school districts, and charter management organizations to better define, communicate, and drive school quality and equity. The Philadelphia-based organization undertakes deep analytic projects, helping clients develop and communicate new systems of accountability, building tools to allow parents to better select schools for their children, developing new ways of visualizing information to improve understanding and decision-making, and designing assessment reports that add context and encourage conversations about next steps. Tembo is platform, programming language, and delivery method-agnostic; it works with clients to develop or recommend whatever solution best fulfills their needs.

David Stewart, Founder & CEO
David Stewart has been working to extract meaningful information from student- and school-level data for more than 20 years. Before founding Tembo in 2010, Mr. Stewart led the school accountability portfolio in the New York City Department of Education (NYC DOE) as the Executive Director of Evaluation and Performance Reporting. In that role, he led the development and implementation of the School Progress Reports, the Learning Environment Surveys, and the Quality Reviews.

Prior to joining the NYC DOE, Mr. Stewart served as Vice President of Product Development and Research at The Grow Network/McGraw-Hill. The Grow Network was one of the first organizations in the education sector to recognize that the way information is presented is at least as important as the information itself. David founded Tembo to carry on the strong tradition of reporting excellence that he had learned at The Grow Network and to help charter management organizations, local education agencies, and state education agencies with a wide range of analytic, accountability, and reporting projects.
Events/Activities and Resources

**Events/Activities:** The table below lists the various events schedule and activities planned to promote Technology and Engineering Literacy (TEL) and the assessment results.

<table>
<thead>
<tr>
<th>Date</th>
<th>Event/Activity</th>
<th>Location</th>
<th>Lead</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/30/15</td>
<td>Meeting with Change the Equation</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>1/6/16</td>
<td>Meeting with National PTA</td>
<td>Alexandria, VA</td>
<td>NAGB</td>
</tr>
<tr>
<td>2/23/16</td>
<td>Conference Call with Boston Museum of Science</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>2/24/16</td>
<td>Meeting with National Academy of Engineering</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>3/1/16</td>
<td>Conference Call with Project Lead the Way</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>3/4/16</td>
<td>International Technology and Engineering Educators Association annual meeting panel</td>
<td>Washington, DC</td>
<td>NCES</td>
</tr>
<tr>
<td>3/14/16</td>
<td>Meeting with American Society for Engineering Education</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>3/17/16</td>
<td>Meeting with International Technology and Engineering Educators Association</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>3/17/16</td>
<td>Webinar with Council of Chief State School Officers</td>
<td>Washington, DC</td>
<td>NAGB/NCES</td>
</tr>
<tr>
<td>3/29/16</td>
<td>Meeting with Alliance for Excellent Education</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>3/30/16</td>
<td>Conference call with International Society for Technology in Education</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>4/8/16</td>
<td>National Conference of State Legislatures assessment and technology panel</td>
<td>Denver, CO</td>
<td>Bill Bushaw</td>
</tr>
<tr>
<td>4/10/16</td>
<td>American Educational Research Association TEL panel</td>
<td>Washington, DC</td>
<td>Cary Sneider</td>
</tr>
<tr>
<td>Date</td>
<td>Event/Activity</td>
<td>Location</td>
<td>Lead</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------</td>
<td>---------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>4/12/16</td>
<td>Meeting with STEM Office of U.S. Education Department</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>4/15/16</td>
<td>TEL luncheon co-sponsored with Chevron</td>
<td>Washington, DC</td>
<td>Linda Rosen</td>
</tr>
<tr>
<td>4/21/16</td>
<td>Learning First Alliance webinar</td>
<td>Washington, DC</td>
<td>NAGB/NCES</td>
</tr>
<tr>
<td>4/29/16</td>
<td>Meeting with National Science Teachers' Association</td>
<td>Arlington, VA</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/5/16</td>
<td>Alliance for Excellent Education webinar</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/16/16</td>
<td>Media Embargo Conference Call</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/17/16</td>
<td>The Nation's Report Card: 2014 Technology and Engineering Literacy, in-person and webcast</td>
<td>Michigan Science Center, Detroit, MI</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/17/16</td>
<td>&quot;Technology and Engineering Literacy In the Real World,&quot; in-person and webcast</td>
<td>Michigan Science Center, Detroit, MI</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/18/16</td>
<td>&quot;Building Technology and Engineering Literacy from a Foundation of Contextual Data,&quot; in-person workshop</td>
<td>Wayne State University, Detroit, MI</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/20/16</td>
<td>U.S. News STEM Solutions National Leadership Conference</td>
<td>Baltimore, MD</td>
<td>NAGB</td>
</tr>
<tr>
<td>5/24/16</td>
<td>Change the Equation webinar on contextual variable report</td>
<td>Washington, DC</td>
<td>Linda Rosen</td>
</tr>
<tr>
<td>June 2016</td>
<td>TEL contextual variables roundtable</td>
<td>Washington, DC</td>
<td>Linda Rosen</td>
</tr>
<tr>
<td>6/16/16</td>
<td>Hill briefing with House STEM Caucus (w/Boston Museum of Science)</td>
<td>Washington, DC</td>
<td>NAGB</td>
</tr>
<tr>
<td>6/29/16</td>
<td>American Society of Engineering Educators annual meeting panel</td>
<td>New Orleans, LA</td>
<td>NAGB/NCES</td>
</tr>
</tbody>
</table>

**Resources:** The table below lists the resources and materials developed or in development to promote Technology and Engineering Literacy (TEL) and the assessment results.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Format/Source</th>
<th>Owner</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slide Deck: General TEL presentation</td>
<td>PowerPoint</td>
<td>NAGB</td>
<td>Completed</td>
</tr>
<tr>
<td>Video: &quot;An Introduction to the NAEP Technology and Engineering Literacy Assessment&quot;</td>
<td>YouTube</td>
<td>NCES</td>
<td>Live</td>
</tr>
<tr>
<td>Video: &quot;Exploring a TEL Task&quot;</td>
<td>YouTube</td>
<td>NCES</td>
<td>Live</td>
</tr>
<tr>
<td>Video: NAEP Technology and Engineering Literacy: New Education Assessment (Alan Friedman)</td>
<td>YouTube</td>
<td>NAGB</td>
<td>Live</td>
</tr>
<tr>
<td>FAQ: General TEL</td>
<td>Word, PDF</td>
<td>NAGB</td>
<td>Completed</td>
</tr>
<tr>
<td>FAQ: Scenario-based assessment</td>
<td>Word, PDF</td>
<td>NAGB</td>
<td>Completed</td>
</tr>
<tr>
<td>Infographics</td>
<td>PNG, PDF</td>
<td>NAGB</td>
<td>In Progress</td>
</tr>
<tr>
<td>Videos on task, findings and contextual variables (4)</td>
<td></td>
<td>NCES</td>
<td>Completed</td>
</tr>
<tr>
<td>Infocards on TEL</td>
<td></td>
<td>NCES</td>
<td>completed</td>
</tr>
</tbody>
</table>
**The Nation’s Report Card: Technology and Engineering Literacy**

**Release Events and Webcast**

**May 17–18, 2016, Detroit, Michigan**

**Post-Event Report**

**Executive Summary**

The National Assessment Governing Board hosted an in-person release event and webcast on May 17 at the Michigan Science Center to release results of the first-ever *Nation’s Report Card: Technology and Engineering Literacy*. A discussion-based workshop was held the following day (May 18) at Wayne State University to convene Detroit-based education and industry experts to engage with the data.

Event panelists and speakers included:

- Nina Abubakari, Managing Partner, Fern Health Solutions
- Jeevak Badve, Vice President of Strategic Growth, Sundberg-Ferar
- Mike Balatzis, Infrastructure and Cloud Consultant, C/D/H
- Chris Barman, Vice President, Unit Responsible—Electrical/Electronic Engineering, Fiat Chrysler Automobiles
- Peggy Carr, Acting Commissioner, National Center for Education Statistics
- Virginia Edwards, President and Editor, Editorial Projects in Education
- Tonya Matthews, President and CEO, Michigan Science Center, and Member, National Assessment Governing Board
- Terry Mazany, President and CEO, Chicago Community Trust, and Chairman, National Assessment Governing Board
- Scot Osterweil, Creative Director, Education Arcade at the Massachusetts Institute of Technology
- Honorable Gary Peters, United States Senator representing Michigan (virtual)
- Sonya Pouncy, Co-Founder, 313 Construction
- Cary Sneider, Associate Professor, Portland State University
- Brian J. Whiston, State Superintendent of Michigan

Eighth-grade students from University Prep Science & Math Middle School joined the panel as well to share their experience taking one of the assessment tasks.

A variety of new strategies and tactics, highlighted below, were used to promote the events, and technology and engineering literacy, a relatively new concept to Governing Board stakeholders.
Significant Activities and Results

- The Governing Board hosted the release outside its familiar Washington, D.C., setting, so Team DCG collaborated with Governing Board staff, member Tonya Matthews, and her staff at the Michigan Science Center, as well as staff at the Integrative Biosciences Center (IBio) at Wayne State University, to coordinate two events in Detroit.
- Team DCG developed a new database of approximately 500 Michigan contacts—including community, education, private sector, and public service stakeholders—to invite to the release event. The team also developed a list of nearly 400 priority education stakeholders and a list of 50 new technology subcommittee members and staff contacts (which was added to the existing congressional database) to receive release promotions and webcast invitations.
- Bill Bushaw and Governing Board members supported outreach and promotional efforts by sending personalized emails to invitees and priority organizations, and by sharing content with their social media networks.
- The news release was viewed 303 times on the PR Newswire site and appeared on 172 websites, including Yahoo! Finance, Reuters, and CNBC, reaching a potential audience of nearly 10 million.
- The Governing Board’s live tweets during the webcast were retweeted 53 times, garnering more than 16,500 impressions.
- Forty-seven original pieces were published, 27 of which appeared in national outlets, such as The Washington Post.
- Team DCG facilitated interviews with multiple outlets, such as CNN, Associated Press, and CBS Detroit.

Key Recommendations

- Cohost a release event with Governing Board member(s) to maximize exposure, especially when events are outside of Washington, D.C., and members’ influence within their communities and states can be leveraged.
- Continue to rely on Governing Board staff and members, panelists/speakers, and partners/priority organizations to support promotional and outreach efforts by linking to the event splash and registration pages on websites, sharing information with their networks via social media, and personally sending invitations. This strategy will also help the Governing Board foster long-term relationships with partners/priority organizations.
- Use Governing Board staff and members and panelists/speakers as spokespeople, increasing their availability to reporters by including them on the media pre-conference call and including a list of their names, contact information, and biographies on the embargo site. This will not only give reporters more approved resources, but will also increase the number of Governing Board spokespeople quoted in the news, and reduce the chance of reporters trying to interview anyone without embargoed access.
- Collaborate more closely with the National Center for Education Statistics (NCES) to promote the assessment and its results and relevant resources, regardless of whether they are developed by the Governing Board or NCES. Conducting the kickoff meeting as early as possible, and then scheduling at least one follow-up meeting prior to the release, would increase communications.
- Schedule release events no earlier than 11:00 a.m. to accommodate stakeholders on the West Coast and increase participation.
The Nation’s Report Card: Technology and Engineering Literacy
Congressional Briefing
June 16, 2016
Rayburn House Office Building, Washington, D.C.

Post-Event Report

Executive Summary
The National Assessment Governing Board hosted an in-person congressional briefing and luncheon on June 16 at the Rayburn House Office Building to present the results of the first-ever Nation’s Report Card: Technology and Engineering Literacy and to engage congressional staff, policy-focused stakeholders, and media. The event was also hosted by the Museum of Science, Boston’s National Center for Technological Literacy®, and the luncheon was sponsored by the American Society for Engineering Education (ASEE), American Society of Civil Engineers (ASCE), American Society of Mechanical Engineers (ASME), International Technology and Engineering Educators Association (ITEEA), National Center for Technological Literacy, Oracle, and Society of Women Engineers (SWE).

Event speakers and panelists included:
- Nate Ball, Mechanical Engineer, TV Host, Entrepreneur, and Author
- Bill Bushaw, Executive Director, National Assessment Governing Board
- Peggy Carr, Acting Commissioner, National Center for Education Statistics
- Representative Joe Kennedy (D-Mass.)
- Ioannis Miaoulis, President and Director, Museum of Science, Boston, and Founder, National Center for Technological Literacy
- Tonya Miles, Member, National Assessment Governing Board, and Chief Departmental Administrator, Office of the General Counsel, Maryland-National Capital Park and Planning Commission
- Representative Paul Tonko (D-N.Y.)
- Claus von Zastrow, Chief Operating Officer and Director of Research, Change the Equation

A variety of strategies and tactics, highlighted below, were used to promote the events, and technology and engineering literacy, a relatively new concept to Governing Board stakeholders.
Significant Activities and Results

- The Governing Board hosted one of its most successful congressional briefings to date.
- More than 90 people attended the event.
- Social media-based support from speakers and panelists, Governing Board members and staff, and partners continues to effectively promote events. Prepared, ready-to-post social media content generated a combined 1,040 clicks to the event registration page. The Department of Education; Society of Women Engineers; the Museum of Science, Boston; and Change the Equation generated more than 600 of those clicks.
- A targeted media list of more than 60 contacts—including D.C. national, education, policy, and technology reporters—were contacted, generating interest in the event. Corrine Letsch with EdScoop attended the event in person.
- The Governing Board’s live tweets during the briefing were retweeted 53 times, generating 22,046 impressions.

Key Recommendations

- Continue partnering with relevant organizations with established networks to leverage cross-promotional opportunities and to increase interest in events.
- Live-stream events to expand reach among stakeholders and to increase the likelihood of earning media coverage.
- If no new data is being presented, consider incorporating other opportunities for media that might entice them to attend in person, such as exclusive interview opportunities.
- Continue to provide select Governing Board members and staff, speakers, and panelists with customized, ready-to-use content to support promotional efforts.
Governing Board Guidelines for Releasing, Reporting, and Disseminating NAEP Results

Background
In August 2006, the Governing Board produced a Policy Statement on the Reporting, Release, and Dissemination of NAEP Results, which was accompanied by specific guidance on releasing NAEP results, entitled Guidelines for the Initial Release of The Nation’s Report Card. The Policy Statement delineates the statutory responsibilities for the NAEP program held by the National Center for Education Statistics (NCES) and by the National Assessment Governing Board. This statement also covers principles that drive the preparation, content, release, and dissemination of The Nation’s Report Card. The Guidelines for the Initial Release of The Nation’s Report Card outline the procedures for releasing NAEP data and the elements necessary for inclusion in an initial release of NAEP results.

At the March 2016 meeting of the Reporting and Dissemination (R&D) Committee, members requested that Board staff suggest revisions to the Policy Statement and Guidelines that would lead to an efficient and effective process for the reporting and dissemination of NAEP results and to eliminate outdated language in these documents which no longer bear relevance (e.g., references to print reports).

In May 2016, the Reporting and Dissemination Committee discussed potential innovations to the Board’s reporting of NAEP results, of which the members prioritized initiating several efforts:

- Building a predictable schedule of NAEP releases to facilitate media coverage;
- Presenting valid, well-considered, relevant findings from other data sources alongside NAEP data, such as international data, if feasible and appropriate;
- Developing indicators with NAEP data, analogous to the Consumer Price Index (CPI), that the public can understand and reference easily;
- Staggering the presentation of NAEP data, e.g., releasing some data on one day and releasing other data subsequently, perhaps with deep analyses of those post-release data; and
- Re-organizing reports by type of questions different target stakeholders ask.

This productive discussion set an agenda for pursuing broad, critical changes in reporting and dissemination of NAEP data. And work is underway to realize those goals. But this precluded a conversation about revising specific elements of the policy statement and guidelines which govern the Board’s reporting and disseminating of NAEP findings.

Current Work
At the August 2016 Board meeting, the Committee will discuss the more specific updates, revisions, additions, and deletions to the 2006 Policy Statement and Guidelines. The Governing
Board’s Assistant Director for Reporting and Analysis drafted a new version of the Policy Statement that subsumes updates to the guidelines outlined in the August 2006 Guidelines document. Rather than adopting two separate documents which overlap significantly, combining the guidelines with the policy statement ensures a more streamlined approach. Governing Board staff reviewed the new version, and the revised version was shared with NCES staff.

For the discussion at the August 2016 R&D Committee meeting, the current 2006 Policy Statement and Guidelines is included for reference. The new version will be distributed prior to the meeting. The R&D Committee session will solicit input from R&D members about what to change in the latest version of the Policy Statement to create a draft that could be subject to Committee approval at the next quarterly Board meeting in November 2016.

The session will include a discussion led by Daniel McGrath, Chief of the Reporting & Dissemination Branch in the National Center for Education Statistics’ Assessment Division about the development and review process for NAEP reporting, including how NCES contractors incorporate feedback from focus groups on the release sites. Then the discussion will turn to the Policy Statement itself and reactions from Committee members.
The Nation’s Report Card: Technology and Engineering Literacy
Media Analysis

June 9, 2016

Overview

On May 17, 2016, the National Assessment Governing Board and the National Center for Education Statistics (NCES) released the first-ever Nation’s Report Card for Technology and Engineering Literacy. Team DCG monitored the news coverage from May 17 through June 3 and found 52 original news pieces about the release from national, education trade, technology and engineering trade, and local outlets. In addition, there were 1,361 reprints, rebroadcasts, or broadcast stations that briefly mentioned the results, for a total of 1,413 placements.

For this media analysis, Team DCG read and analyzed all 52 original news pieces. This analysis includes:

- A summary of selected media coverage about the report card, including findings about messaging, common themes, and Governing Board spokespeople
- Multimedia highlights
- Recommendations

Analysis of Media Coverage

Messaging: For this report card release, the Governing Board wanted to emphasize the newness of this assessment, the innovative scenario-based tasks, what this test measures, and why it is important, as well as the contextual variables around in-school and out-of-school experiences.

- Overall, the vast majority of the coverage focused on girls outperforming boys. Of the 52 articles, 43 of the headlines highlighted girls outperforming boys. Only nine headlines did not focus on the gender gap.
- National outlets tended to focus on the girls vs. boys difference in the headline, while local and trade outlets highlighted other statistics, such as racial and socioeconomic differences, in their headlines. Examples of other headlines include Chalkbeat’s “Black and white students score far apart,” and EdTech magazine’s “Digital skills education takes place in schools and at home.”
- Of the messages included in the news release, the most cited results were the overall statistic of 43% of students scoring Proficient, differences between ethnicities, and comparisons between students who were eligible for school lunch.
A total of 33 articles included the number of students who performed at or above Proficient (43%), while 28 articles covered racial differences and 21 mentioned the school lunch eligibility differences. Learning from family members how to fix things was also highly cited, with 18 mentions.

The results around eligibility for school lunch were most often used to say that affluent students performed substantially better than their counterparts (21 mentions).

**Common Themes:** In the coverage of this report card release, journalists related the results to the following newsworthy themes:

- **Lack of women in technology and engineering careers.** Based on the popularity of the girls vs. boys headline mentioned above, the most common theme was the lack of women in technology and engineering careers. Eight articles compared the results to the shortage of women at technology companies.

- **Third annual U.S. News/Raytheon STEM Index.** Politico, Newser, and U.S. News & World Report wrote about the results in the context of the third annual U.S. News/Raytheon STEM index, which was released the same week. The report showed a shortage of STEM workers and mentioned that the current number of STEM graduates is less than the demand for jobs in those fields.


- **Gender Gap.** Emily Richmond’s Atlantic article was the only article to question the mass popularity of the National Assessment of Educational Progress (NAEP) girls vs. boys gender difference.
  - She notes, “So why did so much of this week’s media call with reporters focus on the relatively smaller lead girls held over boys on the new assessment? That was because ‘we did not expect this pattern,’ explained Peggy Carr … By comparison, the gaps in socioeconomic status and race have long been evidenced in NAEP scores for other core subjects: ‘It’s sort of the same old story,’ Carr said.”
  - Her article also received the most comments out of all the articles, with over 230.

**Governing Board Spokespeople:**

The most commonly quoted spokespeople (listed in order of most quoted) included:

- Peggy G. Carr, Acting Commissioner, NCES
  - Quoted in 28 articles.
  - Her quote from the media call—“We did not expect this pattern and the pattern does seem to be pretty clear from the data,” referring to the differences in scores between girls and boys—was used in 10 articles (first appeared in The Wall Street Journal).
• Bill Bushaw, Executive Director, National Assessment Governing Board
  ▪ Quoted in nine articles.
  ▪ He had several different quotes taken from the media call and subsequent interviews.
  ▪ His most popular quote (used three times) included how this assessment “measures the ‘T’ and ‘E’ in STEM.”
• Tonya Matthews, President and CEO of the Michigan Science Center; Member, National Assessment Governing Board
  ▪ Quoted in seven articles.
  ▪ All seven of her quotes came from her quote in the press release: “The scores clearly show that when students have opportunities to engage with technology and engineering, they become fluent in the skills that prepare them for living and working in the modern world. But access to these opportunities from place to place is patchy.”
• Terry Mazany, President and CEO, the Chicago Community Trust; Chair, National Assessment Governing Board
  ▪ Quoted in three articles, all taken from the press release.
• Mary Crovo, NAGB Deputy Executive Director, was also quoted once in the U.S. News & World Report article about the U.S. News STEM Solutions Conference.
  ▪ The Wall Street Journal, Politico, and CNN articles all quoted Vince Bertram, CEO of the nonprofit Project Lead the Way. Vince Bertram also wrote an op-ed on the results for U.S. News & World Report a week after the release.
  ▪ Other than those in the list above, no spokespeople were quoted in more than two news items.

Multimedia Highlights

• Due to the unique nature of this release, several different types of outlets outside the usual national and education trades covered the results, including Glamour, The Mary Sue, Machine Design, Chemical Processing, Evaluation Engineering, and ZDNet.
• To help present the data, reporters often used screenshots from the report card site. Ten of the articles used a screenshot of a graph from the site.
• Two articles embedded the NAEP 2014 Technology and Engineering Literacy Results: An Overview for Grade 8 video.
• NBC News included a screenshot of an NAEP tweet in their article: Females outperforming males in tech and engineering literacy.
• Machine Design was the only outlet to create its own graph from the data.
• eSchool News was the only outlet that included the Iggy scenario table from the press release in their article.
• Local outlet CW 39 News Fix created their own video segment of the results, which included interviewing girls from a local all-girl high school. Christian Science Monitor also linked to this video from their article.
Social Media Highlights

Activities
Team DCG did the following:
- Developed promotional social media content, including a “Why TEL Matters” campaign with Governing Board members, and graphics to post on the Governing Board’s social media channels leading up to the event.
- Live-tweeted during the event/webcast and retweeted other tweets about data before, during, and after the event.
- Created social media posts to direct people to the website and to share findings from the assessment after the event.

Observations
Team DCG analyzed the volume of online mentions of “NAEP” and associated keywords from midnight on the day of the release to five days afterward. Findings from before, during, and after the event appear below.

During and after the event:
- From midnight May 17 through May 22, a total of 2,441 posts appeared on social media.
- Most mentions occurred on Twitter (2,183), mainstream news (138), and blogs (101).
- Over five days, our total reach (the number of people who were served any activity from our page) was 2,350.
- Mentions spiked at 6:00 a.m. the morning of the release event, with 28 percent of the coverage occurring then.
- Fifty-three percent of all #NAEP coverage occurred on the day of the event.

NAEP Keyword Mentions Over Time
• The Governing Board’s live tweets during the webcast were retweeted 53 times, garnering 16,549 impressions (the number of times users saw the tweets). Links in tweets were clicked 10 times.
• Mentions focused primarily on the score gap between girls and boys and on the perceived low number of students who performed at Proficient or above on the assessment.
• Social media accounts for media outlets including The Wall Street Journal, Time, Fortune magazine, The Huffington Post, Education Week, and Vox, shared their pieces about the assessment results.
• Interestingly, Mia Farrow (810,000 followers) shared the Time article about the TEL results.

![Image of Time article](https://via.placeholder.com/150)

• Organizations including Change the Equation, The Education Policy Center at AIR, and the FrameWorks Institute, participated in conversations about the data on social media using #NAEP.
• From May 17 through May 22, the Governing Board gained 24 followers on Twitter and 4 Facebook fans.

Before the event:
• Our best-performing post before the event was the Facebook Live stream of Tonya Matthews at the Michigan Science Center. The post reached 3,482 people, got 1,298 views, and got 87 reactions, comments, and shares.
• Governing Board tweets promoting the release earned 5,584 impressions. Links in tweets were clicked 10 times.
• Team DCG created Twitter posts targeted at stakeholder organizations, inviting them to participate, for Alberto Carvalho (16,700 followers on Twitter), Change the Equation (23,400 followers), Michigan Science Center (2,016 followers), Andrew Ho (614 followers), and Bill Bushaw (311 followers).
Recommendations

Prepare specific talking points and messaging framing to lead the media pre-call discussion. Team DCG recommends having an internal discussion to decide how to frame results from the start of the call, knowing that many other outlets will pick up prominent articles and use quotes taken from the call. For example, Leslie Brody from The Wall Street Journal has been the first reporter to ask a question on the last few media calls, which has allowed her to lead off the conversation. For the TEL release, one of the first questions she asked centered on the girls vs. boys gender gap at the beginning of the call, and several other reporters followed her lead and continued to ask questions about the gap. Knowing this, Team DCG recommends creating talking points for prioritizing and addressing results the Governing Board wants to emphasize right at the start of the call to help guide reporter questions.

Offer a graphic that provides a snapshot of the key results. There were 10 articles that took screenshots from the report card site and included them in their story. Instead of screenshots, it might be helpful to offer an easy-to-use graphic that provides a key snapshot of the results broken down by the categories reporters are most interested in, such as gender, ethnicity, geography, and socioeconomic status. This will make it easier for reporters to pull something that gives an overall snapshot of the results at a glance and make it shareable for social media.

Provide reporters with only one news release on the embargo site. On the embargo site, reporters had access to the news release from the Governing Board as well as a media summary from NCES. For future report card releases, the Governing Board and NCES should work together to develop one document for reporters. The two documents presented very similar information, but provided reporters with separate media contacts. Having just one news release will help maintain consistent messaging across all deliverables. This could tie in to the recommendation above, where NCES can format the data they would like to highlight into a shareable graphic instead of a press release format.

When drafting the press release, consider the most newsworthy data. For TEL, what was highlighted in the press release and what was highlighted in most of the news coverage did not match up. The girls vs. boys results were buried in the seventh paragraph of the release, yet were an “unexpected and noteworthy” finding as quoted by Peggy G. Carr. Results that are contradictory in nature, or most easily relatable to a general audience, are going to be the ones national reporters gravitate toward most for their stories. Team DCG recommends considering these results as ones that should be highlighted in the sub-header or toward the beginning of the release to attract the most media attention.

Continue post-release pitching to technology and women’s outlets. Due to the unique results from this release, and to receiving coverage in different outlets such as Glamour and ZDNet, there is an opportunity to continue coverage of the results. Team DCG recommends conducting post-release outreach to technology publications and blogs, as well as women-focused outlets, using the girls vs. boys results and the gender gap in technology and engineering fields as a news hook.
Overview

On April 27, 2016, the National Assessment Governing Board and the National Center for Education Statistics (NCES) released *The Nation’s Report Card: 2015 Mathematics and Reading at Grade 12*. Team DC Group monitored the news coverage from April 27 through May 4 and found 69 original news pieces about the release from national, education trade, and local outlets. In addition, there were 2,048 reprints, rebroadcasts, or broadcast stations that briefly mentioned the results, for a total of 2,117 placements.

For this media analysis, Team DC Group read and analyzed all 69 original news pieces. This analysis includes:

- A summary of selected media coverage about the report card, including findings about messaging, common themes, and Governing Board spokespeople
- Multimedia highlights
- Recommendations

Analysis of Media Coverage

**Messaging:** For this report card release, the Governing Board wanted to emphasize the 12th-grade math and reading scores’ relevance to preparedness, changes in scores since the last assessment, long-term trends, and what can be learned from the contextual variables on the report card site.

- Of the messages included in the news release, 12th-grade students’ lack of academic preparedness for college was the most common to be covered — especially in headlines.
- The results indicating that math scores went down and reading scores were stagnant were mentioned in almost every report, making them another very common message. However, slightly more than half of the pieces focused on preparedness more than the actual report card results. For example, a post on U.S. News & World Report’s High School Notes blog used NAEP results to support three “red flags” for parents to notice about their teens’ preparedness for college.
- Many pieces noted the disparity between the highest- and lowest-performing 12th graders — how scores in the highest-percentile bracket are improving, while scores in the lowest-
percentile bracket are declining. “Struggling seniors fall further behind,” reporter Caitlin Emma wrote in Politico’s Morning Education newsletter.

- Contextual variables were not widely emphasized in the news coverage.
- Several news items included Governing Board member Andrew Ho’s description of NAEP as the “North Star” in the world of assessments.

**Common Themes:** In the coverage of this report card release, journalists related the results to the following newsworthy themes:

- The most common theme — noted in more than half the stories — was the disparity between the report card results that found only 37 percent of seniors academically prepared for college reading and math and the record graduation rate of 82 percent, as reported in December 2015 by NCES. Some journalists wondered whether this makes the legitimacy of a high school diploma questionable. As Anya Kamenetz of NPR noted, “That leaves a potentially large group of kids who got diplomas but who weren’t ready to succeed in college.”
- The impact of the Common Core was another common theme. Some reporters wrote that education experts are speculating whether the adjustment to a new curriculum is taking some time, or if the Common Core is actually less effective for student learning. The San Diego Union-Tribune’s editorial examined this topic, noting that many schools have faced some disruptions when adopting the Common Core.
- Several stories noted that NAEP is not intended to prove or disprove different education policies. The word “misnaepery” was once again brought up in a post on Education Week’s Curriculum Matters blog, where it was defined as “attempting to use NAEP data to explain cause-and-effect relationships.” With a number of reporters discussing this topic, coverage focused more on the results and less on the causes or effects of the results.
- Other common themes included:
  - President Barack Obama’s legacy in the White House, including ESSA’s impact on education
  - The need for remedial classes (and the associated cost) once students get to college
  - Childhood poverty and demographic changes in the United States
  - The decrease in the high school dropout rate

**Governing Board Spokespeople:**

- The most commonly quoted spokespeople (listed in alphabetical order by last name) included:
  - Bill Bushaw, Executive Director, National Assessment Governing Board
  - Peggy G. Carr, Acting Commissioner, NCES
  - Mitchell Chester, Commissioner, Massachusetts Department of Elementary and Secondary Education; Member, National Assessment Governing Board
  - Andrew Ho, Professor, Harvard Graduate School of Education; Member, National Assessment Governing Board
  - John B. King Jr., U.S. Secretary of Education
• Terry Mazany, President and CEO, The Chicago Community Trust; Chair, National Assessment Governing Board

- A handful of pieces interviewed local superintendents, teachers, and students.
- Other than those included in the list above, no other spokespeople were quoted in more than two news items.

**Multimedia Highlights**

- The results were featured in top print publications, including The New York Times and The Wall Street Journal, as well as on NBC’s “Today” show.
- Four local broadcast stations had a segment about the report card results that related the results to their local area.
- To help present the data, reporters used screen shots from the report card site or infographics and graphs produced by their media outlet.

**Recommendations**

**Use the news release and the media pre-call to clearly address any data discrepancies.** After reporters gained access to the news release for *The Nation’s Report Card: 2015 Mathematics and Reading at Grade 12*, Team DC Group received several emails and calls from reporters concerned about the discrepancy in math scores between the news release and the report card website. Team DC Group recommends that for future report card releases, the Governing Board address any discrepancies in the data with reporters from the beginning by including a footnote in the news release and addressing the discrepancies during the media pre-call. This will help to assure reporters that they are reporting accurate information and eliminate last-minute media requests and questions the day before the embargo lifts.

**Now that the Technology and Engineering Literacy report card has been released, use graduation as a news hook to specifically pitch the 12th-grade preparedness results in early June.** While many journalists focused on the preparedness results in their pieces on the grade 12 math and reading report card, some either mentioned them only briefly or quickly tied the data to the record graduation rate. Graduation season will be a great opportunity to prolong the life of the data from *The Nation’s Report Card: 2015 Mathematics and Reading at Grade 12*. Team DC Group can pitch national reporters whose stories focused more on the report card results and can encourage them to specifically cover the preparedness results in a second piece. The pitches can also use the all-time-high graduation rate of 82 percent as a news hook.

**Continue to make panelists and Board members available during the media pre-call and for individual interview requests before the embargo lifts.**

- **Media pre-call.** A number of quotes from the media pre-call, including those from Peggy G. Carr and Bill Bushaw, were used in reports about the results. The Governing Board should continue to host a media pre-call for journalists and should consider allowing more spokespeople to participate.
- **Interview requests.** Andrew Ho and Chris Minnich were quoted in several news reports.
Ho’s quotes appeared in multiple news items and offered excellent insight on the results—showing how Governing Board members can add value through interviews. The
- Governing Board should continue to identify spokespeople for each release and make them available for reporters to interview before the embargo lifts. Reporters appreciate having a variety of spokespeople to choose from, and having a defined list of approved spokespeople helps maintain consistency in the Governing Board’s messaging. Providing this list also may lower the chance that reporters will speak with someone who is unapproved, which could lead to an embargo breach.

Create talking points about “misnaepery” to be distributed to Governing Board members and staff. “Misnaepery” has been defined by some reporters as the misuse of NAEP data—specifically, attempting to use NAEP data to explain cause-and-effect relationships. This term became well-known during the math and reading report card release in October 2015 and resurfaced in the coverage of the grade 12 release. While “misnaepery” is not an official term, it is important that Governing Board members and staff are aware of what it means so they can address it during interviews or in daily conversation. “Misnaepery” could also be a fruitful avenue for opening a conversation about assessment literacy.

Provide reporters with only one news release on the embargo site. On the embargo site, reporters had access to the news release from the Governing Board as well as a media summary from NCES. For future report card releases, the Governing Board and NCES should work together to develop one document for reporters. Both documents presented very similar information, but provided reporters with separate media contacts. Having just one news release will help maintain consistent messaging across all deliverables.
## Upcoming NAEP Reports as of July 2016

<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><strong>2015 Science Report Card</strong></td>
<td>October 2016</td>
</tr>
<tr>
<td><strong>Other NAEP Reports</strong></td>
<td></td>
</tr>
<tr>
<td><em>Focus on NAEP: Sampling</em></td>
<td>July 2016</td>
</tr>
<tr>
<td><em>Focus on NAEP: Simpsons Paradox</em></td>
<td>July 2016</td>
</tr>
<tr>
<td><em>Focus on NAEP 12th Grade Participation &amp; Engagement</em></td>
<td>September 2016</td>
</tr>
<tr>
<td><em>2015 National Indian Education Study</em></td>
<td>October 2016</td>
</tr>
</tbody>
</table>
Releases in 2016

- 2015 Science Report Card
- 2015 National Indian Education Study
- Focus on NAEP: Sampling
- Focus on NAEP: Simpson’s Paradox
- From Algebra to Zoology: How Well Do Students Report Mathematics and Science Coursetaking?
- Focus on NAEP: 12th Grade Participation & Engagement
<table>
<thead>
<tr>
<th>Time</th>
<th>Item</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:30 – 4:35 pm</td>
<td>Welcome and Agenda Overview</td>
<td><em>Terry Mazany, Chair</em></td>
</tr>
</tbody>
</table>
| 4:35 – 4:40 pm | **ACTION:** Nomination of Board Vice Chair for the Term October 1, 2016 – September 30, 2017 | *Terry Mazany, Chair*  
*Anitere Flores, Board Member* |
| 4:40 – 4:45 pm | Strategic Vision                                                      | *Lily Clark, Assistant Director for Policy and Research*                       |
| 4:45 – 5:20 pm | NAEP Research Grants                                                  | *Peggy Carr, Acting Commissioner*                                             |
| 5:20 – 6:00 pm | **CLOSED**                                                            | *NAEP Budget and Assessment Schedule*  
*Bill Bushaw, Executive Director*  
*Peggy Carr, Acting Commissioner* |
|               | **Attachment**                                                       |                                                                              |
Summary of NAEP and NAGB Appropriations – August 2016

Congress provided the National Assessment of Educational Progress (NAEP) with $149 million for fiscal year (FY) 2016. This increased the NAEP budget, which is overseen by the National Center for Education Statistics (NCES), by $20 million. Congress provided these funds in support of NAEP’s transition to digital-based assessments and to expand the Trial Urban District Assessment.

The President’s budget request for FY 2017, if enacted, would continue funding the NAEP program at $149 million and modestly reduce the National Assessment Governing Board’s appropriation by $490,000.

As of August 2016, the Senate’s education Appropriations Subcommittee and the House Appropriations Committee have passed education funding bills. The Senate bill matches the President’s budget request for NAEP and the Governing Board. The House bill would drastically cut NAEP by $20 million, and modestly cut the Governing Board’s budget as proposed in the President’s request.

Table of Recently Approved and Proposed Appropriations for NAEP and the Governing Board:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>National Assessment</td>
<td>$132M</td>
<td>$129M</td>
<td>$149M</td>
<td>$149M</td>
<td>$149M</td>
<td>$129M</td>
</tr>
<tr>
<td>Educational Progress</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Governing Board</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Senate Subcommittee on Appropriations for Labor, Health and Human Services, Education, and Related Agencies passed the FY 2017 appropriations bill on June 9, 2016. It matches the President’s budget request; if enacted it would:

- Continue funding for the NAEP program at $149M.
- Modestly cut NAGB’s budget to $7.745M ($490,000 less than FY16).
- Require NAEP to assess US History, Civics, and Geography at least every four years, in accordance with the current Assessment Schedule (8th grade in 2018, and grades 8 & 12 in 2022). See text below.

The House Appropriations Committee passed the education appropriations bill on July 14, 2016. If enacted it would:

- Cut funding for the NAEP program by $20M.
- Modestly cut NAGB’s budget to $7.745M ($490,000 less than FY16).

**Senate Appropriations Bill Language for FY 2017:**

The Committee recommends $156,745,000 to provide support for the National Assessment of Educational Progress (NAEP), a congressionally mandated assessment created to measure and report the educational achievement of American students in a range of subjects and analyze trends over time. Within the funds appropriated, the Committee recommends $7,745,000 for the National Assessment Governing Board (NAGB), which is responsible for formulating policy for NAEP. The Committee is pleased that the NAGB reinstated assessments for 8th and 12th grade students in United States History, Civics, and Geography. Previous assessments conducted by NAGB indicate that fewer than one in four 4th, 8th, and 12th grade students at all grade levels are proficient in United States History. The Committee directs the NAGB to continue administering the assessments in these three areas at least every four years, in accordance with the current NAEP schedule. According to this schedule, the next administration will be in 2018.
### National Assessment of Educational Progress

#### Schedule of Assessments
Approved November 21, 2015

The National Assessment of Educational Progress (NAEP) Authorization Act established the National Assessment Governing Board to set policy for NAEP, including determining the schedule of assessments. (P.L. 107-279)

<table>
<thead>
<tr>
<th>Year</th>
<th>Subject</th>
<th>National Grades Assessed</th>
<th>State Grades Assessed</th>
<th>TUDA Grades Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>U.S. History*</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Civics*</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Geography*</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>TECHNOLOGY AND ENGINEERING LITERACY</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2015</td>
<td>Reading*</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics*</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Science**</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2016</td>
<td>Arts*</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2017</td>
<td>Reading</td>
<td>4, 8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>4, 8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2018</td>
<td>U.S. History</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Civics</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Geography</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Technology and Engineering Literacy</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2019</td>
<td>Reading</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>Reading</td>
<td>4, 8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2022</td>
<td>U.S. HISTORY</td>
<td>8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>CIVICS</td>
<td>8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>GEOGRAPHY</td>
<td>8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Economics</td>
<td>8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Technology and Engineering Literacy</td>
<td>8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2023</td>
<td>Reading</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2024</td>
<td>ARTS</td>
<td>8</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>FOREIGN LANGUAGE</td>
<td>12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Long-term Trend</td>
<td>~</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
</tbody>
</table>

**NOTES:**

*Assessments not administered by computer. Beginning in 2017 all operational assessments will be digitally based.

**Science in 2015 consisted of paper-and-pencil and digital-based components.

~Long-term Trend (LTT) assessments sample students at ages 9, 13, and 17 and are conducted in reading and mathematics.

Subjects in **BOLD ALL CAPS** indicate the year in which a new framework is implemented or assessment year for which the Governing Board will decide whether a new or updated framework is needed.
Researchers’ Use of NAEP Data

Friday, August 5th, 12:30 p.m. – 2:30 p.m.

The National Assessment of Educational Progress not only represents the gold standard in large-scale assessment but also pioneers new approaches to measuring, analyzing, and reporting student achievement for the nation. The National Assessment Governing Board sets the policies and assessment schedule for the National Assessment of Educational Progress (NAEP), and the National Center for Education Statistics administers NAEP.

Researchers can drive the impact of NAEP on national, state, and local policy, because they can serve as conduits for key stakeholders. Generally, policymakers, educators, business leaders, administrators, parents, and members of the public drawn to NAEP will not invest the time, effort, and resources necessary to delve into the trends, patterns, and relationships in the data. Researchers do make this investment. This session features prominent researchers who know and use NAEP data:

- **Thomas Cook**, Northwestern University and Mathematica Policy Research
- **Susan Dynarski**, University of Michigan
- **Charles Payne**, University of Chicago
- **Sean Reardon**, Stanford University

**Andrew Ho**, Governing Board member and the Chair of the Board’s Committee on Standards, Design and Methodology, will invite the panelists to share their work with the Board. He will then moderate a discussion about the potential and the challenges of analyzing NAEP data. This discussion will include an opportunity for Board members to ask questions of the panelists.

The session should encompass myriad topics and promote lively discussion among the panelists and Board members. For example:

- What are the most effective ways of using NAEP to address national, state, and district policy questions?
- How does NAEP’s design both foster and limit the capacity to answer causal questions that most people want to ask of NAEP data?
- How can the Board encourage more analyses and modeling of NAEP data?
- What alternative approaches can NAEP take to reporting results?
Thomas Cook

Cook is a Senior Fellow at Mathematica Policy Research and the Joan and Sarepta Harrison Chair of Ethics and Justice where he is a professor of sociology, psychology, education and social policy, and is a fellow of the Institute for Policy Research.

Thomas Cook’s work focuses on social science research methodology, program evaluation, school reform, and contextual factors that influence adolescent development, particularly for urban minorities.

Cook has written and/or edited 10 books, including several books on social science methods for inferring causation and conducting evaluations, primarily in the areas of education and community health. His books are considered foundational in the fields of evaluation and social science, especially *Quasi-Experimentation Design and Analysis Issues for Field Settings*, *Qualitative and Quantitative Methods in Evaluation Research*, and *The Foundations of Evaluation Theory*. His work appears regularly in peer-reviewed journals and book chapters.

Cook is a Fellow of the American Academy of Arts and Sciences, and has been a Trustee and Board Chairman of the Russell Sage Foundation in New York and of the Textile Museum in Washington, DC. Cook’s professional awards include the Rossi Award from the Association for Public Policy Analysis and Management, the Sells Award for Lifetime Achievement, Society of Multivariate Experimental Psychology, and the Distinguished Scientist Award of Division 5 of the American Psychological Association. He is a Margaret Mead Fellow of the American Academy of Political and Social Science and a fellow at the American Academy of Arts and Sciences. Cook served on the congressionally appointed committee evaluating Title I (No Child Left Behind) from 2006 to 2008.

Cook holds a Ph.D. from Stanford University.
Susan Dynarski

Susan Dynarski is a professor of public policy, education and economics at the University of Michigan, where she holds appointments at the Gerald R. Ford School of Public Policy, School of Education, Department of Economics and Institute for Social Research and serves as co-director of the Education Policy Initiative. She is a faculty research associate at the National Bureau of Economic Research and the Center for Analysis of Postsecondary Education and Employment. She is a nonresident senior fellow in the Economic Studies Program at the Brookings Institution.

 Dynarski’s research focuses on the effectiveness of charter schools, the optimal design of financial aid, the price elasticity of private school attendance, the relationship between postsecondary schooling and labor market outcomes, and the effect of high school reforms on academic achievement and educational attainment.

 Dynarski has been a visiting fellow at the Federal Reserve Bank of Boston and Princeton University as well as an associate professor at Harvard University. She is an editor of Educational Evaluation and Policy Analysis, former editor of The Journal of Labor Economics and Education Finance and Policy, and is currently on the board of Educational Evaluation and Policy Analysis. She has been elected to the boards of the Association for Public Policy and Management and the Association for Education Finance and Policy, for which she currently serves as president-elect. The National Association of Student Financial Aid Administrators awarded her the Robert P. Huff Golden Quill Award for excellence in research on student aid. Amidst this work, she posts regularly to the New York Times “The Upshot” blog.

 Dynarski earned an A.B. in Social Studies from Harvard, a Master of Public Policy from Harvard and a Ph.D. in Economics from MIT.
Charles M. Payne

Charles M. Payne is the Frank P. Hixon Distinguished Service Professor in the School of Social Service Administration at the University of Chicago, where he is also an affiliate of the Urban Education Institute. His interests include urban education and school reform, social inequality, social change and modern African American history. He is the author of *Getting What We Ask For: The Ambiguity of Success and Failure In Urban Education* (1984) and *I've Got the Light of Freedom: The Organizing Tradition in the Mississippi Civil Rights Movement* (1995). The latter has won awards from the Southern Regional Council, Choice Magazine, the Simon Wisenthal Center and the Gustavus Myers Center for the Study of Human Rights in North America. He recently published *So Much Reform, So Little Change* (Harvard Education Publishing Group), which discusses lessons learned about the persistence of failure in urban districts.

Payne is the recipient of a Senior Scholar grant from the Spencer Foundation and was a Resident Fellow at the foundation for 2006-2007. He has served on the Board of the Chicago Algebra Project, the Steering Committee for the Consortium on Chicago School Research, the Research Advisory Committee for the Chicago Annenberg Project, as well as the editorial boards of *Catalyst*, *The Sociology of Education* and *Educational Researcher*. He currently serves on the Board of MDRC, the editorial board of *High School Journal*, and the advisory board for Teacher College Press' series on social justice. He is the co-founder of the Duke Curriculum Project, which involves university faculty in the professional development of public school teachers, and also co-founder of the John Hope Franklin Scholars, which tries to better prepare high school youngsters for college. He is among the founders of the Education for Liberation Network, which encourages the development of educational initiatives that encourage young people to think critically about social issues and understand their own capacity for addressing them; i.e., freedom schools, social justice schools, rites of passage programs and so on.

Payne holds a bachelor's degree in Afro-American studies from Syracuse University and a doctorate in sociology from Northwestern.
Sean Reardon is the endowed Professor of Poverty and Inequality in Education and is Professor (by courtesy) of Sociology at Stanford University. He also serves as the Director of the Stanford Interdisciplinary Doctoral Training Program in Quantitative Education Policy Analysis. This program is designed to provide doctoral students in social science disciplines and in the Graduate School of Education with advanced training in state-of-the-art quantitative methods of discipline-based education policy analysis.

His research focuses on the causes, patterns, trends, and consequences of social and educational inequality, the effects of educational policy on educational and social inequality, and in applied statistical methods for educational research. In addition, he develops methods of measuring social and educational inequality (including the measurement of segregation and achievement gaps) and methods of causal inference in educational and social science research. In particular, his work focuses on issues of residential and school segregation and of racial/ethnic and socioeconomic disparities in academic achievement and educational success.

He teaches graduate courses in applied statistical methods, with a particular emphasis on the application of experimental and quasi-experimental methods to the investigation of issues of educational policy and practice. He is a member of the National Academy of Education, and has been a recipient of a William T. Grant Foundation Scholar Award, a Carnegie Scholar Award, and a National Academy of Education Postdoctoral Fellowship.

Sean received his doctorate in education in 1997 from Harvard University.
The National Assessment Governing Board began its Strategic Planning Initiative (Initiative) by developing a framework to set priorities and goals. The Strategic Planning Framework was unanimously approved by the Board on August 8, 2015, concluding Phase I of the Initiative. At the August 2015 meeting, the Board met in small cross-committee groups to develop an initial list of activities the Board could accomplish within five years to achieve its priorities specified in the Strategic Planning Framework. These conversations formed the basis for the first draft of the Strategic Plan.1

Upon approval of the Framework, Chair Mazany asked Vice Chair Lucille Davy to lead Phase II of the Initiative to guide the Board’s development of its Strategic Plan. At the Board’s November 2015 meeting, Vice Chair Davy led the Governing Board in its first plenary discussion of the Draft Strategic Plan Activities document.

The Board’s plan for Phase II includes soliciting feedback from external education stakeholders to inform the Strategic Plan. The Board hired a consultant, Jim Kohlmoos, to conduct conversations with 22 individuals who are respected education leaders, familiar with NAEP, and represent a diverse range of perspectives to generate ideas for the Strategic Plan. In addition, the staff discussed the priorities and Draft Strategic Plan Activities document with the Board’s joint Policy Task Force with the Council of Chief State School Officers, representing state assessment experts. The external feedback was prepared for the Board’s consideration at its March 2016 meeting in a plenary session where Jim Kohlmoos participated.2

At the May 2016 meeting, the Governing Board engaged in a detailed discussion of the revised draft Strategic Plan. The Board met first in small, cross-committee groups and then as the full Board in a plenary session. What emerged from the May 2016 meeting was a call for a significantly revised strategic document that would focus purely on the Board’s work and be an inspiring, succinct, and effective public communications tool.

Chair Mazany tasked the Executive Committee to revise the draft for discussion at the August 2016 Board meeting. The Executive Committee members were highly engaged in its June and July teleconference meetings. As a result, the May 2016 draft has transformed into the now-called “Strategic Vision” draft. In addition to the Executive Committee’s work, the draft Strategic Vision also went through numerous rounds of Board staff review and reflects feedback from NCES.

At the August 2016 meeting, Board members are asked to review the Strategic Vision draft with a focus on ensuring the draft comprehensively includes all of the initiatives the Board wants to pursue.

The August 2016 meeting agenda is organized to provide Board members the opportunity to discuss the draft in small, cross-committee groups prior to the plenary session. It is the Board’s expectation that the Strategic Vision will be ready for action at the November 2016 meeting. Upon approval of the Strategic Vision, the Board will conclude Phase II of the Initiative. Phase III is the final stage to implement the vision.

---


2 A summary of the feedback provided for the March 2016 meeting is available here: [https://www.nagb.org/content/members/assets/documents/meetings/board-meetings/2016/2016-03/10-strategic-planning-initiative.pdf](https://www.nagb.org/content/members/assets/documents/meetings/board-meetings/2016/2016-03/10-strategic-planning-initiative.pdf)
Overview

The Nation's Report Card is the largest nationally representative, continuing evaluation of the condition of education in the United States and has served as a national yardstick of student achievement since 1969. Through the National Assessment of Educational Progress (NAEP), The Nation's Report Card informs the public about what America's students know and can do in various subject areas, and compares achievement data over time and among student demographic groups. The assessment provides objective information on student performance to policymakers, educators, and the public at the national, state, and urban district levels. It has served an essential role in evaluating the condition and progress of American education for more than four decades.

The independent, bipartisan National Assessment Governing Board was created by Congress in 1988 to set policy guidelines for The Nation’s Report Card. In overseeing The Nation's Report Card, the Governing Board identifies subjects to be tested, determines the content and achievement levels for each assessment, approves all test questions, and takes steps to improve the form, reporting, and use of results. The Governing Board also informs the public about The Nation's Report Card by communicating its results to a wide range of stakeholders including policymakers, educators, researchers, the media, and the general public. The Governing Board partners with the National Center for Education Statistics (NCES), which administers the NAEP program, to accomplish this work. By law, The Nation’s Report Card is expressly prohibited from influencing state and local curriculum or instruction and from reporting individual student or school results. Therefore, the program maintains its reputation as an independent, low stakes assessment providing valuable and trustworthy information to the country.

The Nation’s Report Card provides the country with the tools to measure student progress in achieving the knowledge and skills necessary for success as citizens in a democratic society and a rapidly changing world. This allows the nation to learn where more work must be done to improve learning among all students. NAEP was established to answer the important question: “How are our nation’s students doing?” to which we add a second question: “How can NAEP provide information about how our students are doing in the most innovative, informative, impactful ways?” Through NAEP, the Governing Board seeks to answer these questions by creating the following Strategic Vision to guide its work through the year 2020.

*The Governing Board will increase the value of The Nation’s Report Card as a resource to impact student achievement by continuing to innovate the form and content of NAEP and expand its use and dissemination.*
INFORM

The National Assessment Governing Board will increase stakeholders’ awareness of The Nation’s Report Card’s wealth of information (such as results, contextual variables, items, measurement innovations, frameworks, studies, etc.) and will facilitate stakeholders’ uses of NAEP in appropriate and meaningful ways.

1. Strengthen and expand partnerships to promote awareness and use of NAEP results and other resources.
2. Increase opportunities to link NAEP to other assessments and data sources, including state, national and international student assessments.
3. Expand the availability, utility, and use of NAEP resources, in part by creating new resources targeted to stakeholders’ needs and promoting research to inform education policy initiatives.

INNOVATE

The National Assessment Governing Board will keep The Nation’s Report Card at the forefront of measuring student achievement—both in its form and content—through implementation of the following strategies.

1. Use flexible approaches to update NAEP subject area frameworks to support the Board’s duties to both measure the changing expectations and requirements for students and maintain trend to report on educational progress.
2. Continue improving the content, analysis, and reporting of NAEP contextual variables by considering the relevance and sensitivity of the questions; highlighting meaningful context in the data; and enhancing the reporting of these data.
3. Research policy and technical implications related to the future of NAEP Long-Term Trend assessments in reading and mathematics.
4. Develop policy approaches to ensure the NAEP Assessment Schedule matches the Board’s policy priorities based on the nation’s evolving needs and NAEP’s funding.
5. Explore the feasibility of developing new approaches to measuring the complex skills required for transition to postsecondary education and career.
IMPACT

The National Assessment Governing Board will pursue all of the strategies identified under the Inform and Innovate priorities to increase the impact of The Nation’s Report Card. Together, these strategies form the focus of the Governing Board’s work to inform and influence educational progress. This Strategic Vision will guide Board decisions about which initiatives must be continued, ceased, or commenced in response to new opportunities and the ever-changing external environment.

Conclusion

The Governing Board seeks for The Nation’s Report Card to be used as a resource to prepare students for their future and will continually improve it to deliver on this commitment. An essential role of the Governing Board is to safeguard public trust in NAEP’s measurement of our nation’s elementary and secondary students’ academic achievement. The Nation’s Report Card provides our country with information to understand the opportunities, challenges, and trends in our decentralized system of education. When there are discussions about student achievement, NAEP is often relied upon as a trusted source of information.

The educational landscape of the 21st century demands increased academic rigor, greater technological sophistication, improved civic participation, and expanded global perspectives for all students. In this time of rapid and accelerating change, it is essential for The Nation’s Report Card to propel innovation to ensure NAEP’s leadership role and to address the challenges of improving student achievement, while maintaining its timeless promise to serve as the constant and unassailable measure of student achievement for our nation.
National Assessment Governing Board
Nominations Committee

August 6, 2016
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>Update on Nominations for Board Terms Beginning on October 1, 2016</td>
</tr>
<tr>
<td></td>
<td><em>Mary Crovo, Governing Board Staff</em></td>
</tr>
<tr>
<td></td>
<td>Discussion of Nominations for Board Terms Beginning on October 1, 2017</td>
</tr>
<tr>
<td></td>
<td><em>Committee Members</em></td>
</tr>
</tbody>
</table>
Inside NAEP: Training on Contextual Variables

Brief overview of contextual variables history (including Board policies), key milestones, current status, and plans for 2017 and beyond

Contextual questionnaires have been included in NAEP assessments for years although their prominence in reporting as well as in media coverage has increased recently. The primary objective of these questionnaires is to provide context to NAEP student achievement data and assess changes over time.

In October 2011, the National Assessment Governing Board convened an expert panel to recommend how to make better use of existing contextual variables and to propose an agenda for additional topics and questions to include in NAEP that would be useful in developing education policy and valuable to the public. The six-member panel was chaired by Marshall Smith, former U.S. Under Secretary of Education, and the panel delivered a report to the Board in March 2012. One of the results of this report was the Board’s “Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting.” This policy statement was unanimously adopted by the Board on August 4, 2012 (and accompanies this background document below).

During this session, NCES will provide an overview of the current research-based development process for the NAEP contextual questionnaires as well as progress to date with implementing the policy principles and implementation guidelines listed within the Board’s policy statement. This will include spotlighting some of the recent changes and innovations made to contextual questionnaire development, and how these innovations help the NAEP program transition from paper and pencil assessments to a digital-based platform in 2017 and beyond. These include:

- Conducting cognitive interviews for all new and revised questions.
- Creating questionnaire modules around emerging topics of policy interest that are rooted in decades of published research, directly related to understanding academic achievement (e.g., non-cognitive student factors, school climate), and useful in the transition of NAEP to digital-based assessment (e.g., technology use).
- Developing questionnaire indices and a psychometric process for more robust reporting on constructs.
- Using pilot studies to test multiple versions of a question or construct to help inform the best version for the operational assessments.
- Spiraling questions in pilot studies to test a larger number of questions without increasing respondent burden.
• Balancing the qualitative art of writing contextual questions with the quantitative science of evaluating contextual questions (e.g., timing data, factor analysis, correlation with achievement, response rate distribution, high percentage missing).

• Establishing better response options for survey questions that replace the widely criticized agree-disagree format through systematic study of different response scales in cognitive labs and pilot administrations.

• Implementing improvements to the teacher and school online contextual questionnaires to reduce respondent burden.

Issues related to sensitivity and privacy concerns among states

Over the past several NAEP administrations, NCES has received an increase in notifications from states that are electing not to participate in at least some portion of the contextual questionnaires. The vast majority of these concerns relate to the “Core” (i.e., non-subject specific) section of the student questionnaires. At the state level, there have been virtually no concerns communicated to NAEP regarding the subject-specific questionnaires.

During this presentation, NCES will provide an update on state participation with contextual questionnaires, including in recent NAEP administrations as well as in the plans for 2017. NCES will differentiate the nuances between sensitivity concerns and privacy concerns, detailing how these two related issues might be best messaged to various NAEP constituencies.

The presentation will also feature recent activities NCES has done to help assuage these concerns along with considerations for the future.
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. 7 and 9]

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. 7 and 9]

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. 13]

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. 13 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. 13]

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. 12 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. 11 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. 12 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. 13 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. 13 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]
# Governing Board and NAEP Resources

## Table of Resources and Links

Attached documents are listed with page numbers. Click underlined links to access unattached documents online.

<table>
<thead>
<tr>
<th>Page No.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>National Assessment Governing Board: Authority and Organization</td>
</tr>
<tr>
<td></td>
<td>- NAEP Law</td>
</tr>
<tr>
<td></td>
<td>- Board By-laws</td>
</tr>
<tr>
<td></td>
<td>- Board Composition and Responsibilities</td>
</tr>
<tr>
<td></td>
<td>- Board Members and Categories</td>
</tr>
<tr>
<td></td>
<td>- Ethics Primer for the National Assessment Governing Board</td>
</tr>
<tr>
<td></td>
<td>- Board Staff Organization</td>
</tr>
<tr>
<td></td>
<td>- NAEP Organizational Model (relationship to other organizations)</td>
</tr>
<tr>
<td></td>
<td>- Board Current Contracts</td>
</tr>
<tr>
<td></td>
<td>- Board 2014 Strategic Communications Plan</td>
</tr>
<tr>
<td></td>
<td>NAEP Schedule of Assessments</td>
</tr>
<tr>
<td></td>
<td>- NAEP Schedule of Assessments (latest version)</td>
</tr>
<tr>
<td></td>
<td>- History of Changes to the NAEP Schedule of Assessments</td>
</tr>
<tr>
<td></td>
<td>- Recent NAEP releases</td>
</tr>
<tr>
<td></td>
<td>General Web-based Resources</td>
</tr>
<tr>
<td></td>
<td>- Home page of Governing Board web site</td>
</tr>
<tr>
<td></td>
<td>- Home page of the Nation’s Report Card web site</td>
</tr>
<tr>
<td></td>
<td>- Materials for previous Board meetings</td>
</tr>
<tr>
<td></td>
<td>Board Policies for NAEP</td>
</tr>
<tr>
<td></td>
<td>- General Policy: Conducting and Reporting NAEP</td>
</tr>
<tr>
<td></td>
<td>- Framework Development</td>
</tr>
<tr>
<td></td>
<td>- Item Development and Review</td>
</tr>
<tr>
<td></td>
<td>- Developing Student Performance Levels for NAEP</td>
</tr>
<tr>
<td></td>
<td>- Reporting, Release, and Dissemination of NAEP Results</td>
</tr>
<tr>
<td></td>
<td>- Guidelines for the Initial Release of The Nation's Report Card</td>
</tr>
<tr>
<td></td>
<td>- Resolution on Reporting 12th Grade Academic Preparedness for College</td>
</tr>
<tr>
<td></td>
<td>- Resolution on Reporting on Preparedness of 12th Grade Students</td>
</tr>
<tr>
<td></td>
<td>- Collection and Reporting of Background Data by NAEP</td>
</tr>
<tr>
<td></td>
<td>- NAEP Testing and Reporting on Students with Disabilities and English Language Learners</td>
</tr>
</tbody>
</table>
• Trial Urban District Assessment: Eligibility Criteria and Selection Procedures
  o List of Eligible TUDA Districts
• Resolution on Linking NAEP and International Assessments

NAEP Assessment Design

• Overview of NAEP Assessment Design .................................................................32
• NAEP Alliance Contractors ..................................................................................35

Selected Board-commissioned research reports and papers (from most to least recent)

• Technical Report: NAEP 12th Grade Preparedness Research
• Technical Panel on 12th Grade Preparedness Research – Final Report
• The Future of 12th Grade NAEP: Report of the Ad Hoc Committee on Planning for NAEP 12th Grade Assessments in 2009
• Redesigning the National Assessment of Educational Progress

Previous “Inside NAEP” presentations

• Developing NAEP Frameworks: A Look Inside the Process
• Developing NAEP Test Questions
• Introduction to Validity
• NAEP Achievement Levels
• Sampling Concepts

Glossary of Acronyms and Other Terms .................................................................36
National Assessment Governing Board

Composition
The Board is non-partisan, with 26 members representing gender, geographic, and racial-ethnic diversity. Specific categories of members specified in the NAEP law:

- **Policymakers**: governors or former governors (2), state legislators (2), chief state school officers (2), local school district superintendent (1), state (1) and local (1) school board members, nonpublic school administrator or policymaker (1)
- **Educators**: classroom teachers (3), principals (2), curriculum specialists (2)
- **Public**: general public representatives (2), parents (2), business representative (1)
- **Technical experts**: testing and measurement experts (3)

*The director of the Institute of Education Sciences serves as an ex-officio 26th member.*

Responsibilities
The responsibilities of the Board are mandated by Congress, and include:

- **Test Development**
  - Select subject areas to assess
  - Develop assessment objectives and test specifications
  - Ensure all items are free from bias
  - Have final authority on appropriateness of all items

- **Technical Methodology**
  - Develop appropriate student achievement levels
  - Design the methodology of the assessment to ensure that assessment items are valid and reliable

- **Reporting and Dissemination**
  - Develop guidelines for reporting and disseminating results
  - Plan and execute the initial public release of NAEP reports
  - Take appropriate actions needed to improve the form, content, use, and reporting of results
## National Assessment Governing Board

### Members and Categories by Term Expiration Date

<table>
<thead>
<tr>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anitere Flores</strong>&lt;br&gt;State Legislator (Republican)</td>
<td><strong>Lucille Davy</strong>&lt;br&gt;General Public Representative</td>
<td><strong>Mitchell Chester</strong>&lt;br&gt;Chief State School Officer</td>
<td><strong>Alberto Carvalho</strong>&lt;br&gt;Local School Superintendent</td>
</tr>
<tr>
<td><strong>Rebecca Gagnon</strong>&lt;br&gt;Local School Board Member</td>
<td><strong>James Geringer</strong>&lt;br&gt;Governor (Republican)</td>
<td><strong>Shannon Garrison</strong>&lt;br&gt;Fourth Grade Teacher</td>
<td><strong>Carol Jago</strong>&lt;br&gt;Curriculum Specialist</td>
</tr>
<tr>
<td><strong>Andrew Ho</strong>&lt;br&gt;Testing &amp; Measurement Expert</td>
<td><strong>Doris Hicks</strong>&lt;br&gt;Elementary School Principal</td>
<td><strong>Frank Fernandes</strong>&lt;br&gt;Secondary School Principal</td>
<td><strong>Dale Nowlin</strong>&lt;br&gt;Twelfth Grade Teacher</td>
</tr>
<tr>
<td><strong>Terry Mazany</strong>&lt;br&gt;General Public Representative</td>
<td><strong>Tonya Miles</strong>&lt;br&gt;General Public Representative</td>
<td><strong>Tonya Matthews</strong>&lt;br&gt;General Public Representative</td>
<td><strong>Fielding Rolston</strong>&lt;br&gt;State School Board Member</td>
</tr>
<tr>
<td><strong>Joseph O’Keefe</strong>&lt;br&gt;Non-public School Administrator or Policymaker</td>
<td><strong>Ronnie Musgrove</strong>&lt;br&gt;Governor (Democrat)</td>
<td><strong>Chasidy White</strong>&lt;br&gt;Eighth Grade Teacher</td>
<td><strong>Linda Rosen</strong>&lt;br&gt;Business Representative</td>
</tr>
<tr>
<td>(Vacancy)&lt;br&gt;State Legislator (Democrat)</td>
<td><strong>W. James Popham</strong>&lt;br&gt;Testing &amp; Measurement Expert</td>
<td><strong>Chasidy White</strong>&lt;br&gt;Eighth Grade Teacher</td>
<td><strong>Cary Sneider</strong>&lt;br&gt;Curriculum Specialist</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Ken Wagner</strong>&lt;br&gt;Chief State School Officer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Joe Willhoft</strong>&lt;br&gt;Testing &amp; Measurement Expert</td>
</tr>
</tbody>
</table>

* Member currently serving 2nd term; not eligible for reappointment

Updated 2/12/16
EXECUTIVE SUMMARY

Now that you are a member of the National Assessment Governing Board (“NAGB”) you need to know what ethics laws and rules apply to you. The following is a very brief summary of these rules. For a more detailed discussion of how these rules apply to you, please refer to the attached summary entitled “Ethics Laws and Rules Applicable to SGEs.”

Your Status as a Special Government Employee

You are considered an SGE and not a regular federal employee because NAGB anticipates that you will be serving the federal government through your position for only 130 days or less during any period of 365 consecutive days. Whether or not you are paid by the Board for your service is irrelevant. This summary discusses how the ethics rules apply to SGEs.

Criminal Statutes Apply to Your Activities

Some of the ethics laws that apply to you carry criminal penalties. Below is a brief summary of the most important of these laws.

- The chief conflict of interest law bars you from participating personally and substantially in your capacity as a member of NAGB in any particular matter before the federal government that has a direct and predictable effect on your own financial interests or the financial interests of others with whom you have certain relationships. See 18 U.S.C. Section 208.

- If you find yourself with a financial conflict of interest, you have four options: (1) disqualify yourself (you don’t participate in any way in the matter); (2) resign from the outside entity that is the basis for the conflict; (3) sell or divest the stock or other financial interest that is the basis for the conflict; or (4) request and obtain a statutory waiver.¹

- Two other laws prohibit you from representing a third party, with or without compensation, before any court or agency in connection with any particular matter involving specific parties in which the United States is a party or has a direct and substantial interest and in which you have participated personally and substantially as an SGE. In addition, if you serve the federal government for more than 60 days during the immediately preceding period of 365 consecutive days, these restrictions apply to any matter that is pending with NAGB. But remember that these restrictions do not apply to particular matters of general applicability, such as broadly applicable policies, rulemaking proceedings or legislation, that do not involve specific parties. See 18 U.S.C. Sections 203 and 205.

¹ In rare circumstances, with the concurrence of the U.S. Office of Government Ethics, you may obtain a waiver of the conflict of interest.
• Another criminal law limits some of your activities after your service on NAGB ends. This law prohibits you from representing others in connection with the same particular matter involving specific parties in which you participated personally and substantially during your service to NAGB. This prohibition lasts for your lifetime. See 18 U.S.C. Section 207.

Standards of Ethical Conduct for Employees of the Executive Branch

The Standards of Ethical Conduct for Employees of the Executive Branch (Standards), 5 C.F.R. Part 2635, are regulations that apply both to regular federal government employees and to SGEs. However, a few exceptions exist in the Standards in recognition of the fact that SGEs are working for the government only in a very limited way. A brief synopsis of some these rules and their exceptions follow.

• **Fundraising:** You may not use your official title, position and authority to engage in fundraising.

• **Gifts:** You may not accept gifts from a “prohibited source” or offered to you because of your official position on NAGB. A prohibited source includes any person: seeking official action from NAGB; doing or seeking to do business with NAGB; conducting activities regulated by NAGB; or having interests that may be substantially affected by your official duties. There are many exceptions to this rule that are discussed in more detail in the accompanying memorandum.

• **Lobbying:** In your role as a member of NAGB, you may not urge others to contact Congress or a state legislature to urge the passage or defeat of legislation. Additional restrictions exist regarding lobbying. You should contact Department of Education’s Ethics Division before engaging in any type of lobbying.

• **Misuse of Position:** You may not use your position on NAGB or nonpublic information gained through your service on NAGB to seek advantage for yourself or others. In addition, you may not use your NAGB title in a manner that makes it appear that NAGB is sanctioning your views, products, services or personal enterprises.

• **Political Activities:** You may not engage in political activity when you are on duty or in a federal government building or car, and you may never use your official title as a member of NAGB in connection with political activities.

• **Teaching, Speaking and Writing:** You may not receive compensation for teaching, speaking or writing if: (1) the invitation was offered to you because of your position on NAGB; (2) the information conveyed by you draws substantially on nonpublic information that you obtained by working on NAGB; (3) the invitation was extended to you by an organization or person who has interests that may be substantially affected by your performance on NAGB; or (4) the subject of your work deals in a significant way
with a matter involving specific parties that you worked on while on NAGB. Again, there are some exceptions to this rule that are outlined in more detail in the accompanying memorandum.

**Required Filing of a Financial Disclosure Report By SGEs**

As a member of the NAGB, you are required to file a confidential financial disclosure report (also referred to as a “450” Report) when you are first appointed, and annually thereafter if you are reappointed. The purpose of the financial disclosure form is to protect you from inadvertently violating any of the criminal conflict of interest statutes and so that NAGB can know that your advice is free from any real or perceived conflicts of interest.

Please do not rely solely on this “Executive Summary” before undertaking your duties. There are many subtle nuances that are not discussed in this summary that may apply to your specific situation. The attached expanded summary provides additional detail that will help you better understand the ethics rules. Please feel free to call or e-mail Marcella Goodridge in the Ethics Division of the Office of the General Counsel at the U.S. Department of Education at (202) 401-8309, or Marcella.Keiller@ed.gov, for answers to any specific ethics questions that may arise in the course of your service on NAGB.
ETHICS LAWS AND RULES APPLICABLE TO SGES

I. INTRODUCTION

Although the ethics rules are numerous and detailed, a single, simple principle underlies these rules: You should never use your public office for private gain, either for yourself, or for any third party. In addition, you must refrain not only from engaging in any activity that violates the ethics rules, but you must also refrain from any activity that creates the appearance of a violation of any of these rules. The summary below is designed to help you avoid violating any ethics rules covering your activities as a member of NAGB.

II. YOUR STATUS AS A SPECIAL GOVERNMENT EMPLOYEE

A. What is a “special Government employee”?

Because you have been appointed to be a member of the NAGB and you are expected to perform your duties for not more than 130 days during the 365 days subsequent to the date of your appointment, you are, by law, a “special Government employee” (SGE). As an SGE, you are a federal government employee. This means that upon appointment, you assume the responsibilities, obligations, and restrictions that are part of public service. Because SGEs are not full-time employees, several of these restrictions apply only in limited circumstances.

B. Do the ethics restrictions apply when I am not working for NAGB?

Yes, any restrictions concerning your private activities (representational services, expert witness activities, etc.) apply equally on days when you serve the federal government through your position on NAGB and on days when you do not, except with respect to political activity. If you have not provided any services for the federal government for some time, but have not received a termination date for your appointment, you must seek a formal resolution of the matter before engaging in conduct prohibited by the ethics rules.

III. CONFLICTS OF INTEREST

A. What criminal conflict of interest statutes apply to SGES?

While you are employed as an SGE, you need to pay particular attention to four criminal conflict of interest laws found in Chapter 11, Title 18 of the United States Code: 18 U.S.C. Sections 203, 205, 207 and 208. These criminal laws include some special provisions for the treatment of SGEs. A discussion of these laws and certain related requirements found in other laws and regulations follows.
B. What financial conflicts of interest may arise for SGEs under section 208?

Section 208 prohibits you from participating personally and substantially in any particular matter that has a direct and predictable effect on your financial interests, including certain interests of others that are imputed to you under the statute. This means that you may not work on NAGB matters if you have certain connections – through the ownership of stock, through employment, or by virtue of other circumstances – with an organization that has a financial interest in the matter. For example, you may not work at all on a contract competition if you own stock valued at a certain amount in a company competing for the contract. You may not participate in a discussion of whether to modify an existing contract with a company if you work for that company. And, you may not assist in the development of a scope of work for a contract competition if you know that an organization on which you serve on the Board of Directors plans to compete for that contract.

In addition to your own personal financial interests, the financial interests of the following persons or organizations are imputed to you and also disqualify you from participating in a particular matter:

(1) your spouse;
(2) your minor child;
(3) your general partner;
(4) an organization for which you serve as an officer, director, trustee, general partner or employee; and
(5) any prospective employer.

Example 1 You are on the governing board of ABC, a nonprofit organization. ABC’s financial interests are imputed to you under the statute. This means that for the purpose of determining whether you have a conflict of interest, ABC’s financial interests are treated as if they were your own. Accordingly, you may not participate in any NAGB matter in which ABC has a financial interest. Similarly, if you were in the process of discussing employment with ABC, you would be barred from participating in any NAGB matter affecting the financial interests of ABC.

Example 2 You are on the governing board of ABC (or employed by ABC, own stock in ABC, seeking employment with ABC, etc). You are asked to participate in the process of reviewing and scoring contract proposals for a contract competition for a NAGB project. Fifteen organizations have submitted a bid. When you open the proposal from one organization, you note that ABC’s name is one of the organizations that has submitted a bid. Or, perhaps ABC is listed as a subcontractor in one of the proposals. This contract competition is a “particular
NOTE: Apart from the criminal conflicts of interest statutes discussed above, a regulation also exists that prohibits you from participating in a matter involving specific parties if a reasonable person would question your impartiality.

matter” that will have a “direct and predictable effect” upon the financial interests of ABC. In other words, as a result of the contract competition, ABC will either gain business or not, and this decision will affect ABC financially – either negatively or positively. The amount of financial interest is not relevant – as long as ABC’s finances will be affected, unless a regulatory exemption or waiver permits you to do so, you may not work on this competition. And, because each proposal is competing against all of the others, your evaluation of competing proposals will affect the chances ABC has of winning the contract. Accordingly, you may not review any of the proposals.

You must recuse yourself from a matter as soon as you realize that you have a conflict. If, for example, you notice that you have a conflict when you are in the middle of reviewing contract proposals, you put the proposal back in its envelope and call up an NAGB staff member and let that person know that you think that you are disqualified from working on the competition. If there is any question, you should contact the U.S. Department of Education Office of the General Counsel’s Ethics Division for guidance. Once you have determined that you may not work on this matter, send the proposal back to NAGB staff.

You are permitted to participate in a particular matter affecting one campus of a multi-campus institution of higher education, where the disqualifying interest arises from your employment with a separate campus of the same institution, provided that you have no multi-campus responsibilities at the institution. If you are employed with a large university with multiple campuses and you do not have any multi-campus responsibilities, you may participate in official matters--such as grants, contracts, applications, and other particular matters--that affect the financial interests of another campus in the same university system where you are employed. Below are some examples of how section 208 may apply to your activities.

Example 3 You are employed as a professor at the University of California-Berkeley. NAGB is planning to evaluate the impact of computer-based testing on students with disabilities and English language learners. UC-Berkeley’s science and technology department has submitted a bid. NAGB’s actions will have a direct and predictable effect on the university’s financial interest. Therefore, you may not participate in any way on this matter.

Example 4 You are employed as a researcher at the University of California-Berkeley. NAGB is planning to evaluate the impact of computer-based testing on students with disabilities and English language learners. The University of California-Los Angeles (UCLA) has submitted a bid to be the contractor for NAGB’s evaluation. You may participate in this matter because it will not have a direct and predictable effect on either your financial interests or UC-Berkeley’s.
C. How do I resolve a conflict of interest?

1. Disqualification

A common method of resolving a conflict of interest is to disqualify yourself from participating in the matter.

*Example 5* You are serving on NAGB’s Ad Hoc Committee that will examine issues related to computer-based testing for students with disabilities and English language learners, including developing a study of computer-based testing methodologies. The Request for Proposals has been disseminated. One of the bids submitted is from ABC Corporation (ABC). You own $20,000 worth of stock in ABC. You must advise the U.S. Department of Education Office of the General Counsel’s Ethics Division that you own stock in ABC and you will not be able to participate in any way in the entire contract competition. If ABC is awarded the contract, you will also need to disqualify yourself from the entire matter.

2. Divestiture

Divestiture of a disqualifying interest (usually through the sale of stock) is another remedy available to avoid a potential violation of section 208. SGEs are not eligible for a Certificate of Divestiture (CD). A CD is a tax benefit that allows the deferral or nonrecognition of capital gain where an employee divests a financial interest in order to comply with conflict of interest requirements. Unfortunately, Congress specifically excluded SGEs from eligibility to receive CDs. 26 U.S.C. § 1043(b)(1)(A).

3. Resignation

On some very rare occasions when none of the aforementioned options are available or feasible, an SGE may need to resign from participating in an outside activity with an entity if his or her official activities as an SGE have a direct and predictable effect on the financial interest of that entity creating an irreconcilable conflict.

4. Waiver or Authorization

Another remedy to avoid a conflicting financial interest is to request and obtain a statutory waiver by contacting the Department of Education’s Ethics Division (an authorization is similar to a waiver, but only applies to non-statutory conflicts of interest - what are often referred to as “appearances of a conflict”). You may be granted a waiver only if your financial interest is not so substantial as to be deemed to be likely to affect the integrity of your services.

*Example 6* In the scenario described in Examples 1 and 2 above, you are granted a waiver permitting you to participate in a general policy matter that affects ABC’s financial interests as...
long as the matter affects all similarly situated entities in the same manner. But you would remain disqualified from participating in a matter that specifically involves ABC, which in this case means the entire contract competition.

D. What restrictions apply to my representation of third parties under sections 203 and 205?

With regard to particular matters in which you have participated personally and substantially while serving NAGB, you are prohibited from representing a third party on those particular matters, with or without compensation, before any court or agency, when the United States is a party or has a direct and substantial interest in the matter. See 18 U.S.C. Sections 203 and 205.

In addition, if you serve the federal government for more than 60 days during the immediately preceding period of 365 consecutive days, you are prohibited from representing a third party on any matter involving specific parties pending before NAGB, even if your work at NAGB did not involve these matters. These restrictions do not apply to particular matters of general applicability, such as broadly applicable policies, rulemaking procedures or legislation that does not involve specific parties.

IV. POST-EMPLOYMENT

After your appointment terminates at NAGB, you need to pay particular attention to one more criminal statute that subjects you to restrictions regarding certain matters that you may have worked on as a member of NAGB. Pursuant to 18 U.S.C. Section 207, you may never represent any third party, other than in the performance of your official government duties, in connection with the same particular matter involving specific parties in which you participated personally and substantially as a member of NAGB. This is a lifetime prohibition. For example, if you participated in a NAGB discussion concerning a contract to State University, you may never represent State University with respect to that same contract before any official of the Executive Branch of the federal government and you may never represent State University with respect to that contract in any federal court.

Further, if you serve on NAGB more than sixty days and are compensated above a certain level, you may be subject to a one-year “cooling-off” period during which you would be barred from representing before NAGB certain third parties in connection with any matter. There are some exceptions to this law as well, and you should contact the Department of Education’s Ethics Division for guidance.

V. STANDARDS OF ETHICAL CONDUCT AND OTHER ETHICS RULES

The Standards of Ethical Conduct for Employees of the Executive Branch (Standards), 5 C.F.R. Part 2635, are regulations that apply both to regular federal government employees and to SGEs. Although you are treated generally the same as regular employees under the Standards, a few
exceptions do exist for SGEs in recognition of the fact that SGEs are working for the government only in a very limited way. In addition, there are other rules that govern your conduct as an SGE, including the Hatch Act, anti-lobbying rules, the Federal Acquisition Regulation, and rules about accepting gifts and compensation from foreign governments. A brief synopsis of some of these rules follows.

A. What restrictions apply if I want to engage in fundraising?

You may not use your NAGB title, position or authority to solicit funds for any organization. In addition, you may not personally solicit funds or other support from persons whose interests may be affected substantially by the performance or nonperformance of your official duties.

B. What restrictions are there on my acceptance of gifts?

You are prohibited from accepting gifts (almost anything of monetary value) from a “prohibited source” or gifts given because of your official position as a member of NAGB, unless a specific exception applies. The definition of “prohibited source” includes any person:

- seeking official action from NAGB;
- doing or seeking to do business with NAGB; or
- having interests that may be substantially affected by your official duties at NAGB.

The definition also includes organizations the majority of whose members fall within any of these categories. You may accept various benefits resulting from your outside business or employment activities, if a reasonable person would conclude that such benefits are not offered or enhanced because of your official position. The most commonly applicable exceptions to the gift rule allow you to accept:

- Modest items of food other than a meal, such as coffee, soft drinks, or donuts;
- Most plaques, certificates and trophies;
- Discounts available to all Government employees;
- Anything for which you pay market value;
- Gifts valued at $20 or less per occasion, totaling no more than $50 in a calendar year from any one source;
- Gifts clearly motivated by friendship or family relationship;
- Gifts resulting from your outside business activities, including those of your spouse; and
- Free attendance or meal which is provided by:

1. the sponsor of the event for the day on which you are speaking at the event, or for a widely-attended gathering of mutual interest to a number of parties when the necessary determination of agency interest has been made; or

2. someone other than the sponsor of a widely-attended gathering of mutual interest to a number of parties when more than 100 people are expected to attend, the
aggregate value of the gift is under $335, and the necessary determination of
agency interest has been made.

C. What restrictions apply if I want to “lobby” Congress?

NAGB and its members are permitted to communicate directly with Congress in their official
capacity on matters that are related to legislation or appropriations deemed necessary to conduct
NAGB’s “public business” (i.e., the NAGB’s statutory functions and responsibilities). However,
the Anti-Lobbying Act, 18 U.S.C. Section 1913, prohibits you, in your official capacity at
NAGB, from engaging in “grass-roots lobbying” (i.e., directly or indirectly suggesting or
requesting that others contact Congress or a state legislature to urge the passage or defeat of
proposed or pending legislation), even if it is related to the NAGB’s public business. The Anti-
Lobbying Act also requires that any permissible direct communications with Congress in your
official capacity at NAGB be made only through official channels.

None of these restrictions prohibit you from lobbying members of Congress or state legislatures,
or urging others to do so, on your own time in your personal capacity. If you lobby Congress or
state legislatures in your personal capacity, and the issue is related to NAGB’s business, you
should make it clear that you are not representing NAGB and not acting in your official capacity
as a member. Also, please note that when you are lobbying as a private citizen, you are not
permitted to use government resources or equipment (including, but not limited to, computers,
telephones, fax machines, copy machines, stationery), or seek assistance from NAGB staff.

D. What does “misuse of position” mean?

You may not use your position on NAGB to seek advantage for yourself or others. You also
may not use nonpublic information gained through your service at NAGB to seek advantage for
yourself or others. Finally, you may not use your NAGB title in a manner that makes it appear
that the NAGB is sanctioning your views, products, services or personal enterprises. Of course,
you may list your membership on NAGB on your curriculum vitae, but you may never use your
status as an NAGB member to advertise or promote your personal activities. Please seek advice
from the Department of Education Office of the General Counsel’s Ethics Division if you have
any questions in this area.

E. May I keep my day job and still serve on NAGB?

Yes, you may continue to collect your regular salary from an outside employer for days on which
you are providing services to the federal government (whether your federal government service
is paid or unpaid). However, if you have another consultant or advisory position with NAGB or
any other federal department or agency, you may not receive per diem or salary from NAGB for
the same day for services performed for the two positions.

F. Are there any restrictions on my political activities?

You may not engage in any political activities while you are on duty (i.e., performing
government services) or when you are in a government building or vehicle. Although you are not subject to any restrictions on your political activities when you are not performing government services, you may never use your official title as a member of NAGB in connection with any political activities.

**G. What restrictions do I face if I want to teach, speak, or write on matters that are related to the duties I perform for NAGB?**

You may not receive compensation for teaching, speaking, or writing if:

- the activity is performed as part of your official duties (e.g., a speech on behalf of NAGB);

- the invitation to engage in the activity was extended primarily because of your official position at NAGB, rather than expertise in the subject matter;

- the invitation or offer of compensation was extended to you by someone with interests that may be affected substantially by your duties;

- the information conveyed through the activity draws substantially on nonpublic information obtained through your service at NAGB; or

- the activity deals, in significant part, with a matter involving specific parties to which you are currently assigned or had been assigned during your current NAGB appointment.

Notwithstanding the restrictions in bold type you may accept compensation for teaching a course requiring multiple presentations offered as part of: (a) the regularly established curriculum of various specified types of educational institutions; or (b) educational or training programs sponsored and funded by federal, State, or local government. However, if you teach at an educational institution, you must not participate in any NAGB matters that involve that institution.

**H. What restrictions apply if my government duties involve the awarding of contracts?**

If you are involved in the awarding of any contracts, please seek advice from the Ethics Division. There are special provisions that cover your involvement in the awarding of contracts. For example, you may not accept compensation as an employee, officer, director, or consultant of a contractor within the one-year period after leaving Government service where you participated in certain procurement matters pertaining to that contractor. In addition, if you disclose certain information pertaining to Federal procurements that you obtained during your service on a committee, you may face sanctions, including criminal penalties.
I. What restrictions apply to my interaction with foreign entities?

The emoluments clause of the U.S. Constitution prohibits you from receiving any emolument, office or title of any kind from a foreign government, including political subdivisions of a foreign government. An emolument is compensation received by virtue of holding an office or having employment with a foreign government and includes, for example, salary, honoraria, transportation, per diem allowances, household goods, shipment costs, and housing allowances. This clause has been interpreted to be broader than the traditional notion of employment and includes, for example, income received through a partnership when an identifiable portion of the partnership draw can be attributed to the partnership’s fees from such foreign government. This provision has particular relevance to positions with foreign universities that are government-operated, as opposed to private institutions. United States Constitution, art. I § 9, cl. 8. There are also statutory provisions restricting acceptance of gifts from foreign governments. 5 U.S.C. § 7342. You should seek advice from the Ethics Division regarding the details about these restrictions. Additionally, a criminal statute bars employment or consultation with a foreign entity for the purpose of providing foreign agent representation or lobbying. 18 U.S.C. § 219.

The ban on participating in foreign agent activities covered by the Foreign Agents Registration Act (FARA) prohibits representation of foreign governments or foreign political parties before the United States Government, as well as a number of other activities conducted within the United States on behalf of such entities. There are certain FARA exceptions related to trade or commerce, legal representation, humanitarian fundraising, and religious, scholastic, or scientific pursuits. The Lobbying Disclosure Act of 1995 requires certain covered Federal officials who serve as agents of foreign principals (other than foreign governments or foreign political parties) to register if they work on behalf of foreign corporations, associations, or other organizations.

Finally, certain restrictions apply after your position with NAGB terminates. Specifically, 18 U.S.C. § 207 includes restrictions on former employees who participated in trade or treaty negotiations on behalf of the United States (18 U.S.C. § 207(b)) and on former senior employees who wish to represent, or aid or advise in the representation of, a foreign entity with the intent to influence a decision of a Federal employee or agency (18 U.S.C. § 207(f)).

J. What do I do if I am called to be an expert witness?

Government employees generally may not participate as an expert witness, with or without compensation, other than on behalf of the United States, in any proceeding before a federal court or agency in which the United States is a party or has a direct and substantial interest. This restriction applies to most SGEs only if the SGE actually participated officially in the same proceeding or in the particular matter that is the subject of the proceeding. If you are appointed by the President, serve on a commission established by statute, or serve (or are expected to serve) for more than 60 days in a period of 365 days, the restriction on expert service also applies to any proceeding in which NAGB is a party or has a direct and substantial interest.
K. May I keep and use frequent flyer miles that I earn when I am on official NAGB travel?

Yes, you may use frequent flyer miles or other airline awards or promotions accumulated on official NAGB travel for your own personal use.

VI. CONCLUSION

We understand that these laws are complex and may not be intuitive. Again, we caution you that this summary is merely an introduction to the ethics laws and rules that apply to you. You should always feel free to contact the Department of Education Office of the General Counsel’s Ethics Division with any questions or concerns.

Marcella Goodridge Keiller, Attorney
U.S. Department of Education
Office of the General Counsel
400 Maryland Avenue, S.W., Room 6E237
Washington, D.C. 20202-2110
(202) 401-8309
(202) 260-5104 (fax)

Marcella.Keiller@ed.gov
# NATIONAL ASSESSMENT GOVERNING BOARD
## CURRENT CONTRACTS

<table>
<thead>
<tr>
<th>Board Committee &amp; Activity</th>
<th>Area of Work</th>
<th>Contractor</th>
<th>Staff Member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Committee on Standards, Design and Methodology (COSDAM)</td>
<td>Statistical Linking Studies and Related Data Sharing Agreements</td>
<td>Via states and NAEP Alliance contractors ETS and Westat</td>
<td>Sharyn Rosenberg</td>
</tr>
<tr>
<td></td>
<td>Participant Engagement in NAEP: Critical Review and Synthesis of Research</td>
<td>AnLar Incorporated</td>
<td>Michelle Blair</td>
</tr>
<tr>
<td></td>
<td>12th Grade Preparedness Reports</td>
<td>Widmeyer Communications (Year 5 of 5 Years)</td>
<td>Mary Crovo Michelle Blair Sharyn Rosenberg</td>
</tr>
<tr>
<td>Reporting and Dissemination Committee</td>
<td>Outreach and Dissemination</td>
<td>The District Communications Group (Year 2 of 3 Years)</td>
<td>Stephaan Harris</td>
</tr>
<tr>
<td></td>
<td>World Wide Web Services</td>
<td>Quotient, Inc. (Year 2 of 3 Years)</td>
<td>Stephaan Harris Mary Crovo</td>
</tr>
<tr>
<td>Executive Committee</td>
<td>NAGB/Council of Chief State School Officers: State Outreach</td>
<td>Council of Chief State School Officers</td>
<td>Michelle Blair Mary Crovo</td>
</tr>
</tbody>
</table>

July 2016
In 2014 and beyond, the National Assessment Governing Board seeks to focus its communication efforts strategically and cost effectively to "Make Data Matter" for various target audiences. The Board is well-positioned to increase the impact of its outreach, but it must prioritize its audiences and identify its objectives for each, while integrating innovative strategies to elevate the Board’s work—and NAEP—as a thought leader in education.

Reingold proposes three goals the Board can pursue to amplify its outreach efforts.

I. Make a Connection With Target Audiences
II. Engage Audiences Between Report Card Releases
III. Maximize Impact Through Innovation

Reingold’s assumption in developing strategic priorities for the Board is that reporting and dissemination activities must support a vision to make an impact in education through engagement with NAEP that will enable the use, discussion, and sharing of NAEP data and information. A time-phased action plan, including specific outreach tactics and metrics, will be developed with Governing Board staff on the Board’s approval of this strategic communications plan.

The members of the Reporting and Dissemination Committee have identified three key audiences it believes the Board should focus on—parents; teachers and administrators; and policymakers—as each of these audiences is in a position to make an impact through NAEP data. Working with staff, we will identify the Board’s goals and expectations of each audience and the key messages needed to engage each one effectively.

Potential outcomes of the audience-focused outreach are listed below:

**Parents**
- Understand the value of NAEP and its implication for parents.
- Ask informed questions about their child’s education and the school system.
- Use NAEP to consider out-of-school factors that might affect their child’s education.
- Share NAEP information and messages with their parent peers.

**Teachers and Administrators**
- Understand the value of NAEP and its implication for teachers and administrators.
- Use NAEP to influence change within their classroom or school system.
- Educate parents about NAEP data and resources.
- Share and distribute NAEP information to their peers.
**Policymakers**

- Understand the value of NAEP and its implication for education policy.
- Use and cite NAEP data in policy decisions, public statements, and white papers.
- Distribute NAEP information and messages to constituents and peers to help advocate for change.

It is important to remember that messages and calls to action are intended to move the Board’s priority audiences along an engagement continuum, from awareness and education to trial, buy-in, and, ultimately, action. But creating the right messages is only the beginning. It is critical to know which information to deliver first, which should follow, and who are the most credible messengers. We will lay out a cohesive, practical, comprehensive roadmap for reaching the Board’s target audiences that identifies how to take advantage of existing opportunities, what new strategies to develop, and optimal methods of dissemination. The action plan will include a variety of opportunities to connect with each audience to maximize the reach and frequency of each message. The proposed strategies involve cultivating and leveraging partnerships that will include stakeholders or champions. There will also be collaboration with the National Center for Education Statistics (NCES) to ensure efforts are not duplicated, with Board and NCES staff coordinating on roles, responsibilities, and resources on various strategies as needed.

To illustrate the strategies identified above, below we discuss what the execution of each one could involve for the Board’s three priority audiences.

**I. Make a Connection With Target Audiences**

The goal is personal and powerful: “Communicate the Value of NAEP.” This means going beyond the distribution of NAEP data to highlighting, developing, and sharing relevant messages, content, stories, and calls to action for key audiences. Communicating the “So what?” and “Why should we care?” can help the Board move beyond the scores and headlines to clarify the value of NAEP and its important role as an indicator of student achievement.

- **Develop key messages and calls to action for priority audiences.** The Governing Board’s audience is widely diverse—in their knowledge of and experience with NAEP, in their intended uses and consumption of data and information, and in their communications networks, favored channels, and approaches. With these differences in mind, it is imperative that the Governing Board tailor messages for each of its audiences to inspire deeper engagement with NAEP data. Instead of a one-size-fits-all approach, we will define and continually test and adjust the messages that are the most relevant to each audience.

  **Example of the strategy in action for parents:** Include the tailored messages and calls to action on the website’s “Information For” parent pages. The parent landing page could have calls to action including “Learn about NAEP,” “Download NAEP resources,” or “Test yourself on NAEP questions.” The page could also have a section devoted to the Board’s assessment literacy efforts (including resources, information and questions to ask) once outreach strategies from the work group are finalized.
Example of the strategy in action for teachers and administrators: The American Federation of Teachers and National Education Association could include a NAEP toolkit with messages for teachers on its website in a resources section.

Example of the strategy in action for policymakers: Minneapolis Board of Education and Governing Board member Rebecca Gagnon could use and reference data from *Science in Action: Hands-On and Interactive Computer Tasks From the 2009 Science Assessment* in a discussion with the Minnesota Department of Education and the Minnesota Education Technology Task Force about the importance of science computer labs.

**Impact metrics:** The number of downloads of materials such as a PowerPoint or frequently asked questions PDF; number of clicks on links for calls to action (e.g., “Test yourself on NAEP questions”); number of champions—that is, advocates—who commit to using or distributing the NAEP messaging and toolkit.

- **Expand communications beyond reporting on the scores.** We need to get beyond the typical report presentations of the data and find meaningful ways to elevate the data (and their implications) through materials, messaging, and outreach activities. We will identify and highlight hidden gems of NAEP data, connecting the dots between data and practice and leveraging resources to reach specific audiences to deliver important messages in a meaningful and memorable way. The Governing Board must be a storyteller that educates its audiences about the relevancy of NAEP data and resources in a way that resonates with its audiences’ interests and needs in an actionable manner.

Example of the strategy in action for parents: Develop a parent leader discussion guide to assist parent leaders in using NAEP and other assessment data in their conversations with school administrators about improving student achievement for all children.

Example of the strategy in action for teachers and administrators: Develop an interactive Prezi presentation (a visually animated storytelling tool for presenting ideas and messages) on NAEP achievement gap data from the recent 2013 Mathematics and Reading, Grade 12 report card for New Leaders, a national nonprofit organization that develops transformational school leaders and designs effective leadership policies and practices for school systems across the country.

Example of the strategy in action for policymakers: Governing Board member Anitere Flores could host a Florida Senate session on parent involvement in education to highlight NAEP contextual variables data in reading from the 2013 Mathematics and Reading, Grade 12 report card. For example, when asked whether students discussed what they read, students who reported discussing their reading every day or almost every day had higher reading scores.
**Impact metrics:** The number of guides distributed at stakeholder conferences or downloaded from the website; number of groups posting the guide on their websites; number of Prezi and data downloads; parent-submitted testimonials and feedback on using the guide to speak with school and district leaders.

- **Tell the NAEP story through user testimonials.** NAEP data become more impactful when stakeholders learn how others use the data to fulfill their missions and advance their educational goals. Working through key groups, we will collect and disseminate real-life testimonials from the priority audiences to become an authentic author of the NAEP story.

  **Example of the strategy in action for parents:** Collaborate with National PTA to solicit testimonials from parents about how they use NAEP and other assessment data, and then promote the testimonials through the Board’s and PTA’s online networks. These testimonials and other NAEP information could also be featured on the websites of other national education groups, encouraging parents to learn about different assessments their children might take and how the data can be used.

  **Example of the strategy in action for teachers and administrators:** Coordinate with elementary school principal and Board member Doris Hicks and future Board member chosen for the secondary school principal slot to collaborate with the National Association of Elementary School Principals and the National Association of Secondary School Principals to solicit testimonials from principals and teachers within their districts about how they use NAEP and the importance of at-home and out-of-school activities that enhance learning, then promote testimonials through the school communication channels.

  **Example of the strategy in action for policymakers:** Collaborate with the National Association of State Boards of Education to collect testimonials from state board members on how data, including NAEP data, are used to inform policy-level decisions and improvements.

  **Impact metrics:** The number of NAEP user testimonials received; number of testimonial views online; number of social media shares and engagement; quality of the engagements and comments about parents using data.

  - **Potential action taken by key audiences under this goal:** Using NAEP materials and resources on organization websites to inform questions of school and education leaders about school curriculum and district progress; downloading NAEP sample questions to test student knowledge or supplement classroom lessons;

II. Engage Audiences Between Report Card Releases

The goal is ongoing and impactful: “Continual Engagement.” This means building tangible connections—outside of report card release events—between NAEP and its stakeholders, and equipping them with the insight, information, and tools to make a difference in educational quality and student achievement. This important strategy cannot be executed by staff alone, and will require the contributions of Board members and the partnership of stakeholder groups and other NAEP champions, including former Board members.
Expand the report card release life cycle. There is great opportunity for the Governing Board to enliven data and engage target audiences by taking a comprehensive, reimagined view of releasing and reporting on NAEP results that goes beyond the one-day release event. The entire life cycle of an assessment—from developing the framework to fielding assessments to disseminating results—offers content and commentary that, if shared more strategically, will powerfully support the NAEP brand and use of NAEP by target audiences. The Board can both enhance the report card releases and extend the life cycle to make meaningful connections with target audiences by developing pre- and post-release content, and recording and sharing video or audio which tease out and illuminate NAEP data.

**Example of the strategy in action for parents:** For each report card release develop a highlight reel with panelist quotes, select data points, and facts on reading, mathematics, and science contextual variables to send to parent stakeholder groups to distribute to their networks and on the Web.

**Example of the strategy in action for teachers and administrators:** Governor Board member Terry Mazany could host a meeting with the executive director of the Chicago Principals & Administrators Association to discuss the value of NAEP state and TUDA achievement data.

**Example of the strategy in action for policymakers:** Host a briefing with the California State Board of Education on the performance of fourth-grade students in the *NAEP 2012 Writing Grade 4 Pilot* with a diverse panel to include California fourth-grade teacher and Governing Board member Shannon Garrison, the executive director of the National Writing Project, and authors Carol Bedard and Charles Fuhrken.

**Impact metrics:** The numbers of video views and shares; number of groups posting the video; quality of comments and conversations under the video; feedback from stakeholder groups about the impact of the video and parent engagement with the content; number of participants at the meeting or briefing.

**Leverage partnerships with stakeholder organizations and champions.** As a trusted messenger of information to key audiences, the Governing Board needs to mobilize its existing networks, engaging stakeholder groups and champions to share and shape future outreach. Stakeholders and champions are diverse and can be from education associations or news outlets like NBC News. They could also be politicians, celebrities, athletes, or prominent individuals like First Lady Michelle Obama. We will help the Board identify key partnership opportunities for its priority audiences and develop specific recommendations for engagement, to put their distinct capabilities to work in promoting NAEP and extending the Governing Board’s reach. For example, we could keep working with the Alliance for Excellent Education to produce and promote post-release webinars, provide data infographics to the National Council of Teachers of Mathematics, and collaborate with the National Council of La Raza in sponsoring Facebook chats in addition to consistently pursuing new opportunities with key stakeholder organizations.
**Example of the strategy in action for parents:** Collaborate with NBC News’ Education Nation and Pearson on their Parent Toolkit (www.parenttoolkit.com) including NAEP materials, graphics, and downloadable resources on the website that position the Governing Board as an authoritative source of information on student assessment data.

**Example of the strategy in action for teachers and administrators:** Collaborate with Danica McKellar, actress, author, and STEM education advocate, to submit an article to the National Science Teachers Association’s NSTA Express newsletter on the importance of STEM education and girls’ involvement in STEM, and include data from NAEP’s *Technology and Engineering Literacy* assessment.

**Example of the strategy in action for policymakers:** Arrange for James Geringer and/or Ronnie Musgrove, Board members and former governors, to present at the annual National Governors Association conference on an important policy issue affecting states in which NAEP data and contextual variables are relevant. Additionally, the Board and he governors can collaborate with the Center on Education Policy to include NAEP reading data and contextual variables (such as frequency of discussing what they read or finding reading enjoyable) in their research papers, publications and annual progress report.

**Impact metrics:** The number of clicks on the NAEP content; number of downloads of NAEP materials; use of presented NAEP data by governors and state policy leaders in media citations, state websites and other materials; volume of referral traffic from the Parent Toolkit site back to the Governing Board’s website; Education Nation engagement that identifies stories of the Toolkit in action; number of newsletter opens and clicks; number of research report downloads.

- **Equip, empower, and display thought leadership.** The Governing Board and NCES are well-positioned as thought leaders among researchers and many national policymakers but could expand their influence with other audiences, such as parents, local policymakers, and education practitioners. Governing Board members and staff should be seen by media representatives and stakeholders as valued spokespersons on educational assessment and achievement, including specific topics such as computerized assessments, achievement gap trends, 12th-grade academic preparedness, and the importance of technology, engineering, and literacy. The Board can also continually secure speaking engagements at a variety of events such as the International Reading Association’s annual conference or local PTA chapter meetings, or pitch quotes for inclusion in news articles and op-eds on relevant topics.

**Example of the strategy in action for parents:** Work with Board member and parent Tonya Miles and develop and pitch op-eds that connect NAEP data with important year-round education events, emphasizing the role parents can play in raising student achievement. During Black History Month, pitch a piece to HuffPost Parents that spotlights achievement gap success stories, or pitch a piece about technology and engineering skill-building beyond the classroom to *Sacramento Parent* magazine.
Example of the strategy in action for teachers and administrators: Co-host a webinar discussion on NAEP state achievement trends with the American Federation of School Administrators, with members weighing in on state-level changes and education initiatives that are aimed at increasing achievement.

Example of the strategy in action for policymakers: Submit a proposal to the National School Board Association’s annual conference for a Board member and NCES to co-host a breakout session to share and discuss the recent 2013 Mathematics and Reading, Grade 12 report card, academic preparedness data, and recent graduation rate research.

Impact metrics: The numbers of op-ed placements, shares, and comments; quality of user engagements and comments; number of follow-up questions from readers; number of new emails collected (from a “Subscribe to the Governing Board” call to action); number of webinar and conference participants and follow-up requests.

Potential action taken by key audiences under this goal: Inspired by op-ed on racial achievement gaps, exploring gaps in their own districts and talking with school leaders about parity of resources; noting performance trends in subjects by state and/or urban district and then using that knowledge to inform state, local, or school district-level decisions regarding academic programs.

III. Maximize Impact Through Innovation

The goal is proactive and cutting-edge: “Lead the Way.” This means reaching and making meaningful connections with priority audiences, customizing events, fostering and driving online conversations, and creating tech-savvy materials with compelling content.

- Customize release event formats. Report cards are not one-size-fits-all; innovative release event strategies are needed to achieve the specific goals of each release. Each release event strategy should have distinct goals, audiences messages, materials, strategies, and tactics to Make Data Matter. The Governing Board has expanded the report card release event structure from physical events for every release to include webinars and live-streaming during events, a post-release social media Facebook chat, and an online townhall event. We will continue to refine this approach to customizing every release to maximize the immediate release impact and create a sustained conversation that continues to reach and engage key audiences.

Example of the strategy in action for parents: Host a Google Hangout for parents after a NAEP release that can feature panelists from the National Council of La Raza talking about the importance of parent involvement in education, and encourage parent participants to share how they use data to help their students achieve.
**Example of the strategy in action for teachers and administrators:** Develop a Twitter town hall guide (NAEP data points, question-and-answer content, best-practice tips, and facilitation instructions) for teachers and school administrators to host their own facilitated chats with parents and the school district on state-level NAEP data and areas for application.

**Example of the strategy in action for policymakers:** Host an in-person round-table discussion with members of the Massachusetts Mayors’ Association on the latest state-level NAEP reading and mathematics results and their state-based implications.

**Impact metrics:** The number of promotions of the online events and shares of the URL; numbers of event participants and total users viewing them or reached; numbers of comments or participants sharing their testimonials; number of follow-up testimonials received for inclusion in materials or on the website.

- **Engage in the online conversation.** It is important to be aware of the conversations on important education issues, but to influence and help shape public understanding and perceptions the Governing Board needs to participate in the conversation with key messages. We will help the Governing Board foster conversations through real-time engagement on social media platforms, develop content such as an article written by a Governing Board member to post on NAEP’s upcoming blog coordinated by NCES, and create a strategy to join or host online chat events, sponsor Q&A sessions, or solicit feedback. Champions are key to the success of this effort, providing greater reach and often a more powerful story than the Governing Board can tell alone.

**Example of the strategy in action for parents:** Hold a webinar with the Governing Board’s Education Summit for Parent Leaders attendees and parent leader champions to review the NAEP website workshop tutorial and obtain feedback through a moderated chat on how they have used NAEP data since the event. Compile feedback to create a one-pager and share it with participants.

**Example of the strategy in action for teachers and administrators:** Collaborate with the National Council of Teachers of Mathematics (NCTM) on an online Q&A chat session based on the NAEP Mathematics Curriculum Study data, educating NCTM about the wide variance of content in mathematics courses and books with the same name. Board member and math teacher Dale Nowlin could be a participating panelist.

**Example of the strategy in action for policymakers:** Reach out to the National Governors Association (NGA) on Twitter and provide NGA with content and data about the 2013 Mathematics and Reading, Grade 12 report card.

**Impact metrics:** Numbers of campaign participants and user submissions; numbers of engagements (“likes,” comments, shares, retweets, views) for the multimedia submissions; quality of comments on the multimedia submissions; growth in the Governing Board social media audience and number of engaged users discussing assessment data.
Create multimedia, digital content and materials. The Governing Board must present messages, graphics, and images that resonate with target audiences. A wealth of materials has been developed by the Governing Board and NCES, and the first step will be to audit and catalog resources that may be repurposed through outreach and promotional activities. For the materials gaps that are identified, it is imperative to develop interactive, multimedia content and materials that deliver key messages to target priority audiences and include a call to action. Examples include infographics that embellish key report card findings to facilitate understanding and encourage engagement with NAEP data among nonexperts; videos, Prezi, and other presentation tools allowing exploration of the relationships between ideas and numbers and visual presentations of NAEP; and an email newsletter with new content and specific calls to action.

Example of the strategy in action for parents: Create a “NAEP for Parents” email newsletter with information on the latest report card data and trends, multimedia content such as video clips or NAEP data user testimonials, and links to other resource or news content and the interactive data maps on the Board’s parent Web pages, to be distributed bimonthly or consistently throughout the year.

Example of the strategy in action for teachers and administrators: Create an infographic with “hidden data” gems from the NAEP Grade 8 Black Male Students report and accompanying language to share with the National Alliance of Black School Educators to post on social media.

Example of the strategy in action for policymakers: Work with Board member Terry Holliday to create an interactive presentation at CCSSO’s annual large-scale assessment conference on NAEP computer-based assessments, or work with Board member Tom Luna to distribute the dynamic 12th-grade preparedness video highlighting the new college preparedness data to Chiefs for Change members.

Impact metrics: Email open rate; numbers of email shares, clicks from email to website, and new email subscribers; number of release participants who list the email as their referral source; numbers of email replies or responses with inquiries about NAEP or acquiring NAEP materials and resources; number of video and infographic views and shares.

Potential action taken by key audiences under this goal: Using contextual data to influence out-of-school factors that have been shown to correlate with achievement; using curriculum study findings to investigate course rigor and influence change for exposure to challenging subject matter.

By pursuing these three fundamental communication goals and identifying priority strategies and tactics, the Governing Board can more effectively reach its target audiences to Make Data Matter and, ultimately, make an impact.
### National Assessment of Educational Progress
### Schedule of Assessments
### Approved November 21, 2015

The National Assessment of Educational Progress (NAEP) Authorization Act established the National Assessment Governing Board to set policy for NAEP, including determining the schedule of assessments. (P.L. 107-279)

<table>
<thead>
<tr>
<th>Year</th>
<th>Subject</th>
<th>National Grades Assessed</th>
<th>State Grades Assessed</th>
<th>TUDA Grades Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>U.S. History*</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Civics*</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geography*</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TECHNOLOGY AND ENGINEERING LITERACY</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Reading*</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Mathematics*</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>Science**</td>
<td>4, 8, 12</td>
<td>4, 8</td>
<td>4, 8</td>
</tr>
<tr>
<td>2016</td>
<td>Arts*</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>Reading</td>
<td>4, 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>4, 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>U.S. History</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Civics</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geography</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technology and Engineering Literacy</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>Reading</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>Reading</td>
<td>4, 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>U.S. HISTORY</td>
<td>8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CIVICS</td>
<td>8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GEOGRAPHY</td>
<td>8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Economics</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technology and Engineering Literacy</td>
<td>8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2023</td>
<td>Reading</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>4, 8, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2024</td>
<td>ARTS</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FOREIGN LANGUAGE</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-term Trend</td>
<td>~</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTES:**
*Assessments not administered by computer. Beginning in 2017 all operational assessments will be digitally based.
**Science in 2015 consisted of paper-and-pencil and digital-based components.
~Long-term Trend (LTT) assessments sample students at ages 9, 13, and 17 and are conducted in reading and mathematics.
Subjects in **BOLD ALL CAPS** indicate the year in which a new framework is implemented or assessment year for which the Governing Board will decide whether a new or updated framework is needed.
History of Changes to the NAEP Schedule of Assessments

Historical Schedule Changes
The major schedule changes adopted by the Board since 2000 are listed below:

1. Added grade 4 and 8 state-level Reading and Mathematics every two years. (2002) [Prior to the 2002 ESEA reauthorization (NCLB), state assessments at grades 4 and 8 were given every two years with reading and writing in one biennium and mathematics and science in the next, i.e., these subjects and grade 12 subjects were tested once every four years.]
2. Added the High School Transcript Study (HSTS) as a regularly scheduled study. (2005)
4. Added Technology and Engineering Literacy (TEL) to the NAEP subjects assessed. (2005)
5. Added grade 12 state-level Reading and Mathematics for volunteer states with a periodicity of every four years. (2008)
6. Adjusted the periodicity of science to correspond to the periodicity of TIMSS to conduct international benchmarking studies in mathematics and science. (2010)
7. Scheduled Writing as a technology based assessment, beginning with national data collections only and delaying fourth grade in order to complete a special study. (2010)

Other schedule changes and program adjustments from 2000 through 2015 have been due primarily to budget constraints and/or technical challenges, considering options such as:
- Assessing fewer grade levels in non-required subject areas (e.g., U.S. History, Civics, and Geography; Writing; TEL).
- Postponing a state-level assessment
- Postponing a full assessment/study (e.g., World History, Foreign Language, HSTS).
- Changing the sample size and reporting depth for jurisdictions (e.g., alternating subjects with a smaller sample size in a model called focal and non-focal).

Guiding Principles for Schedule Changes
Guiding principles and priorities that have been used to guide planned updates to the NAEP schedule of assessments include:

1. Follow the guidance in the NAEP Act (303(b)(2)),
2. Administer all assessments using technology beginning in 2017,
3. Continue to assess broad-based curricular areas with a priority for science, technology, engineering, and mathematics (STEM),
4. Providing state-level data in curricular areas beyond reading and mathematics,
5. Include more districts in the TUDA program.

Guidance for the schedule is found in Title 303 Sec. 303(b)(2) which addresses the use of random sampling (A), testing in reading and mathematics at grades 4 and 8 once every two years (B), and testing in reading and mathematics at grade 12 at regularly scheduled intervals (at least as often as prior to NCLB (C).

After this initial guidance, Sec. 303(b)(2)(D) provides guidance for including other subjects in grades 4, 8, and 12 to the extent time and resources allow. It says, including assessments “... in regularly scheduled intervals in additional subject matter, including writing, science, history, geography, civics, economics, foreign languages, and arts, and the trend assessment described in subparagraph (F).”
Overview of NAEP Assessment Design

The content and format for each NAEP subject-area assessment is determined by a NAEP assessment framework, developed under the Governing Board’s direction. General details about the structure of NAEP assessments include:

Long Test, Short Student Test Booklet
- Each student gets a small part of the test
- No individual student scores

Common Block Structures Across Subjects
- Items are within blocks, blocks are within booklets
  Example:
  At grade 4: Reading has 10 blocks and Math has 10 blocks

Test Question Types
- Multiple-choice
- Open-ended
- Computer-based tasks (Writing, Science, TEL)

Contextual Questions
- Student, teacher, administrator questionnaires

Student Booklet Block Design

While some NAEP assessments are conducted on a technology-based platform (TEL, Writing), for paper-based assessments NAEP uses a focused balanced incomplete block (BIB) or partially balanced incomplete block (pBIB) design to assign blocks or groups of cognitive items to student booklets. Because of the BIB and pBIB booklet designs and the way NAEP assigns booklets to students, NAEP can sample enough students to obtain precise results for each test question while generally consuming an average of about an hour and a half of each student's time.

The "focused" aspect of NAEP's booklet design requires that each student answer questions from only one subject area. The "BIB" or "pBIB" design ensures that students receive different interlocking sections of the assessment forms, enabling NAEP to check for any unusual interactions that may occur between different samples of students and different sets of assessment questions.

In a BIB design, the cognitive blocks are balanced; each cognitive block appears an equal number of times in every possible position. Each cognitive block is also paired with every other cognitive block in a test booklet exactly the same number of times. In a pBIB design, cognitive blocks may not appear an equal number of times in each position, or may not be paired with every other cognitive block an equal number of times. NAEP booklet design varies according to subject area (e.g., geography, mathematics, reading, science, U.S. history, writing).
Once the instrument developer has laid out the configuration of all blocks for each booklet in a booklet map shown here with the following column headings,

<table>
<thead>
<tr>
<th>Booklet number</th>
<th>Cognitive block 1</th>
<th>Cognitive block 2</th>
<th>Contextual question directions</th>
<th>General student contextual questions</th>
<th>Subject-specific contextual questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the number of rows (booklet numbers) provides the booklet spiral design information needed for the bundling of the student booklets.


### NAEP Assessment Sample Design

Each assessment cycle, a sample of students in designated grades within both public and private schools throughout the United States (and sometimes specified territories and possessions) is selected for assessment. In addition, in state assessment years, of which 2007 is an example, the samples of public schools and their students in each state are large enough to support state-level estimates. In all cases, the selection process utilizes a probability sample design in which every school and student has a chance to be selected, and standard errors can be calculated for the derived estimates.

### Public School Selection in State Assessment Years

The selection of a sample of public school students for state assessment involves a complex multistage sampling design with the following stages:

- Select public schools within the designated areas,
- Select students in the relevant grades within the designated schools, and
- Allocate selected students to assessment subjects.

The Common Core of Data (CCD) file, a comprehensive list of operating public schools in each jurisdiction that is compiled each school year by the National Center for Education Statistics (NCES), is used as the sampling frame for the selection of sample schools. The CCD also contains information about grades served, enrollment, and location of each school. In addition to the CCD list, a set of specially sampled jurisdictions is contacted to determine if there are any newly formed public schools that were not included in the lists used as sampling frames. Considerable effort is expended to increase the survey coverage by locating public schools not included in the most recent CCD file.

As part of the selection process, public schools are combined into groups known as strata on the basis of various school characteristics related to achievement. These characteristics include the physical location of the school, extent of minority enrollment, state-based achievement scores, and median income of the area in which the school is located. Stratification of public schools
occurs within each state. Grouping schools within strata by such selected characteristics provides a more ordered selection process with improved reliability of the assessment results.

On average, a sample of approximately 100 grade-eligible public schools is selected within each jurisdiction; within each school, about 60 students are selected for assessment. Both of these numbers may vary somewhat, depending on the number and enrollment size of the schools in a jurisdiction, and the scope of the assessment in the particular year. Students are sampled from a roster of individual names, not by whole classrooms. The total number of schools selected is a function of the number of grades to be assessed, the number of subjects to be assessed, and the number of states participating.

**Private School Selection in State Assessment Years**

In years in which state-level samples are drawn for public schools, private schools are classified by type (e.g., Roman Catholic, Lutheran, etc.), and are grouped for sampling by geography (Census region), degree of urbanization of location, and minority enrollment. About 700 private schools, on average, are included, with up to 60 students per school selected for assessment. These samples are not large enough to support state-level estimates for private schools. Thus, inferences for private schools are limited to the national level, even in years when public school assessments are state-specific.

A national sample of private schools in all grades is then drawn from a list compiled through the Private School Universe Survey (PSS), which is a mail survey of all U.S. private schools carried out biennially by the U.S. Census Bureau under contract to NCES. The PSS list is updated for new schools only for a sample of Roman Catholic dioceses.

**National-Only Assessment Years**

In years when the NAEP samples are intended only to provide representation at the national level and not for each individual state, the public and private school selection process is somewhat different. Rather than selecting schools directly from lists of schools, the first stage of sampling involves selecting a sample of some 50 to 100 geographic primary sampling units (PSUs). Each PSU is composed of one or more counties. They vary in size considerably, and generally about 1,000 PSUs are created in total, from which a sample is selected. Within the set of selected PSUs, public and private school samples are selected using similar procedures to those described above for the direct sampling of schools from lists. The samples are clustered geographically, which results in a more efficient data collection process. The selection of PSUs is not necessary when the sample sizes are large in each state, as in state assessment years.


**NAEP Alliance Contractors**

NAEP is conducted by the Assessment Division of NCES, which also works with a series of contractors. The following chart presents the structure of the collaboration between these contractors.
To learn more about NAEP contractors in addition to the NAEP Alliance contractors, visit:
http://nces.ed.gov/nationsreportcard/contracts/history.aspx
## Glossary of Acronyms and Other Terms

The following acronyms and terms are commonly used in the work of the National Assessment Governing Board.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASA</td>
<td>American Association of School Administrators</td>
</tr>
<tr>
<td>ACT</td>
<td>Formerly American College Testing</td>
</tr>
<tr>
<td>ADC</td>
<td>Assessment Development Committee (Board Committee responsible for test development on all NAEP subjects)</td>
</tr>
<tr>
<td>AERA</td>
<td>American Educational Research Association</td>
</tr>
<tr>
<td>AFT</td>
<td>American Federation of Teachers</td>
</tr>
<tr>
<td>AIR</td>
<td>American Institutes for Research</td>
</tr>
<tr>
<td>ALDs</td>
<td>Achievement Level Descriptions</td>
</tr>
<tr>
<td>ALS</td>
<td>Achievement Levels Setting</td>
</tr>
<tr>
<td>ARRA</td>
<td>American Recovery and Reinvestment Act of 2009</td>
</tr>
<tr>
<td>AYP</td>
<td>Adequate Yearly Progress (From the No Child Left Behind Act)</td>
</tr>
<tr>
<td>BOTA</td>
<td>Board on Testing and Assessment, National Academy of Sciences</td>
</tr>
<tr>
<td>CCSS</td>
<td>Common Core State Standards</td>
</tr>
<tr>
<td>CCSSO</td>
<td>Council of Chief State School Officers</td>
</tr>
<tr>
<td>CGCS</td>
<td>Council of the Great City Schools</td>
</tr>
<tr>
<td>COSDAM</td>
<td>Committee on Standards, Design and Methodology (Board committee responsible for technical issues)</td>
</tr>
<tr>
<td>CRESST</td>
<td>Center for Research on Evaluation, Standards, and Student Testing (Research Center at UCLA)</td>
</tr>
<tr>
<td>DAC</td>
<td>Design and Analysis Committee (Advisory panel to ETS on technical issues in NAEP operations)</td>
</tr>
<tr>
<td>ECS</td>
<td>Education Commission of the States (First NAEP contractor and organization supporting state policy leaders)</td>
</tr>
<tr>
<td>EIMAC</td>
<td>Education Information Management Advisory Consortium (Advisory committee to CCSSO, mostly state testing directors)</td>
</tr>
<tr>
<td>ELs or ELLs</td>
<td>English Learners or English Language Learner (Pronounced &quot;Ls&quot;; formerly called Limited English Proficient or LEP)</td>
</tr>
<tr>
<td>ELPA</td>
<td>English Language Proficiency Assessment (Also ELPA21)</td>
</tr>
<tr>
<td>EPIC</td>
<td>Education Policy Improvement Center</td>
</tr>
<tr>
<td>ESEA</td>
<td>Elementary and Secondary Education Act</td>
</tr>
<tr>
<td>ETS</td>
<td>Educational Testing Service</td>
</tr>
<tr>
<td>FAR</td>
<td>Federal Acquisition Regulations</td>
</tr>
<tr>
<td>GAO</td>
<td>Government Accountability Office</td>
</tr>
<tr>
<td>GPO</td>
<td>Government Printing Office</td>
</tr>
<tr>
<td>GSA</td>
<td>General Services Administration</td>
</tr>
<tr>
<td>HSTS</td>
<td>High School Transcript Study (A special NAEP data collection)</td>
</tr>
<tr>
<td>IEP</td>
<td>Individualized Education Plan (A required document under the Individuals with Disabilities Education Act, which specifies learning objectives for an individual student found with a disability)</td>
</tr>
<tr>
<td>IES</td>
<td>Institute of Education Sciences (The Department of Education office in which NCES is located. The Director of IES is an ex-officio member of the Governing Board.)</td>
</tr>
<tr>
<td>IRA</td>
<td>International Reading Association</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory&lt;br&gt;<em>(A theory for design, analysis, and scoring of tests)</em></td>
</tr>
<tr>
<td>KaSA</td>
<td>Knowledge and Skills Appropriate&lt;br&gt;<em>(A series of NAEP research studies to improve measurement precision)</em></td>
</tr>
<tr>
<td>KSA</td>
<td>Knowledge, Skill, and/or Ability&lt;br&gt;<em>(A statement describing a subset of academic content)</em></td>
</tr>
<tr>
<td>LEP</td>
<td>Limited English Proficient&lt;br&gt;<em>(Term formerly used for an English Language Learner)</em></td>
</tr>
<tr>
<td>LTT</td>
<td>Long Term Trend Assessment&lt;br&gt;<em>(Series of NAEP tests that began in the early 1970’s)</em></td>
</tr>
<tr>
<td>MST</td>
<td>Multi-stage Testing&lt;br&gt;<em>(A testing format where subsets of test items are presented to students based on item difficulty and student performance)</em></td>
</tr>
<tr>
<td>NAE</td>
<td>National Academy of Education</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment of Educational Progress&lt;br&gt;<em>(Pronounced &quot;nape&quot;)</em></td>
</tr>
<tr>
<td>NAESP</td>
<td>National Association of Elementary School Principals</td>
</tr>
<tr>
<td>NAGB</td>
<td>National Assessment Governing Board&lt;br&gt;<em>(Pronounced &quot;nag bee&quot;)</em></td>
</tr>
<tr>
<td>NAS</td>
<td>National Academy of Sciences</td>
</tr>
<tr>
<td>NASBE</td>
<td>National Association of State Boards of Education</td>
</tr>
<tr>
<td>NASSP</td>
<td>National Association of Secondary School Principals</td>
</tr>
<tr>
<td>The Nation’s Report Card</td>
<td>Alternate reference for NAEP assessments</td>
</tr>
<tr>
<td>NCES</td>
<td>National Center for Education Statistics&lt;br&gt;(Project office for NAEP in the U.S. Department of Education and IES)</td>
</tr>
<tr>
<td>NCLB</td>
<td>No Child Left Behind Act of 2001</td>
</tr>
<tr>
<td>NCME</td>
<td>National Council on Measurement in Education</td>
</tr>
<tr>
<td>NCTE</td>
<td>National Council of Teachers of English</td>
</tr>
<tr>
<td>NCTM</td>
<td>National Council of Teachers of Mathematics</td>
</tr>
<tr>
<td>NEA</td>
<td>National Education Association</td>
</tr>
<tr>
<td>NEA</td>
<td>National Endowment for the Arts</td>
</tr>
<tr>
<td>NEH</td>
<td>National Endowment for the Humanities</td>
</tr>
<tr>
<td>NGSS</td>
<td>Next Generation Science Standards</td>
</tr>
<tr>
<td>NRC</td>
<td>National Research Council</td>
</tr>
<tr>
<td>NSBA</td>
<td>National School Boards Association</td>
</tr>
<tr>
<td>NSLP</td>
<td>National School Lunch Program</td>
</tr>
<tr>
<td>NVS</td>
<td>NAEP Validity Studies Panel</td>
</tr>
<tr>
<td>OGC</td>
<td>Office of the General Counsel&lt;br&gt;(in the U.S. Department of Education)</td>
</tr>
<tr>
<td>OMB</td>
<td>Office of Management and Budget</td>
</tr>
<tr>
<td>PARCC</td>
<td>Partnership for Assessment of Readiness for College and Careers</td>
</tr>
<tr>
<td>PIRLS</td>
<td>Progress in International Reading Literacy Study</td>
</tr>
<tr>
<td>PISA</td>
<td>Program for International Student Assessment</td>
</tr>
<tr>
<td>POC</td>
<td>Principal Operating Components&lt;br&gt;(Divisions of the U.S. Department of Education)</td>
</tr>
<tr>
<td>PTA</td>
<td>Parent Teacher Association</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>Reporting and Dissemination Committee (Board Committee responsible for NAEP reporting issues)</td>
</tr>
<tr>
<td>RFP</td>
<td>Request for Proposals</td>
</tr>
<tr>
<td>RP</td>
<td>Response probability (probability of correct response on a test question)</td>
</tr>
<tr>
<td>RTT</td>
<td>Race to the Top (also referred to as RTTT)</td>
</tr>
<tr>
<td>SBAC</td>
<td>SMARTER Balanced Assessment Consortium</td>
</tr>
<tr>
<td>SD</td>
<td>Students with Disabilities</td>
</tr>
<tr>
<td>SES</td>
<td>Socio-economic Status</td>
</tr>
<tr>
<td>TBA</td>
<td>Technology-based Assessment</td>
</tr>
<tr>
<td>TEL</td>
<td>Technology and Engineering Literacy (A content area assessed by NAEP)</td>
</tr>
<tr>
<td>The Department</td>
<td>United States Department of Education</td>
</tr>
<tr>
<td>The Secretary</td>
<td>Secretary of Education (Honorable Arne Duncan during the Obama administration)</td>
</tr>
<tr>
<td>TIMSS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
<tr>
<td>TUDA</td>
<td>Trial Urban District Assessment (NAEP component that measures students in large urban districts)</td>
</tr>
<tr>
<td>DATE AND TIME</td>
<td>EVENT</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Wednesday, August 3</td>
<td>Chicago Cubs vs Miami Marlins</td>
</tr>
<tr>
<td>1:20 pm OPTIONAL</td>
<td></td>
</tr>
<tr>
<td>Wednesday, August 3</td>
<td>Group Pizza Dinner</td>
</tr>
<tr>
<td>6:30 pm OPTIONAL</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Outreach Visit with leaders of organizations providing innovative education opportunities for Chicago area youth.</td>
</tr>
<tr>
<td>9:00 – 11:00 am</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Assessment Development Committee</td>
</tr>
<tr>
<td>12:00 – 4:00 pm</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Committee on Standards, Design and Methodology</td>
</tr>
<tr>
<td>1:00 – 4:00 pm</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Reporting and Dissemination Committee</td>
</tr>
<tr>
<td>2:30 - 4:00 pm</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Executive Committee</td>
</tr>
<tr>
<td>4:30 – 6:00 pm</td>
<td></td>
</tr>
<tr>
<td>Thursday, August 4</td>
<td>Outreach Dinner</td>
</tr>
<tr>
<td>6:30 – 9:00 pm</td>
<td></td>
</tr>
<tr>
<td>Friday, August 5</td>
<td>Run/Walk</td>
</tr>
<tr>
<td>5:30 am/6:00 am</td>
<td></td>
</tr>
<tr>
<td>Friday, August 5</td>
<td>Full Board Meeting General Session</td>
</tr>
<tr>
<td>8:30 – 10:15 am</td>
<td>(Committee meetings: 10:30 am – 12:15 pm)</td>
</tr>
<tr>
<td>Closed Working Lunch:</td>
<td>12:30 – 2:30 pm</td>
</tr>
<tr>
<td>Open Session</td>
<td>2:30 – 2:45 pm</td>
</tr>
<tr>
<td>Breakout Sessions</td>
<td>3:00 – 4:30 pm</td>
</tr>
<tr>
<td>Friday, August 5</td>
<td>Full Board Working Dinner</td>
</tr>
<tr>
<td>6:30 – 9:30 pm</td>
<td></td>
</tr>
<tr>
<td>Saturday, August 6</td>
<td>Nominations Committee</td>
</tr>
<tr>
<td>7:30 – 8:15 am</td>
<td></td>
</tr>
<tr>
<td>Saturday, August 6</td>
<td>Full Board meeting</td>
</tr>
<tr>
<td>8:30 am – 12:00 pm</td>
<td></td>
</tr>
</tbody>
</table>

The group shuttle to O’Hare Airport will depart the hotel 15 minutes following the conclusion of the Board meeting.