National Assessment Governing Board

Committee on Standards, Design and Methodology

August 4-5, 2016

AGENDA

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

National Assessment Governing Board

Committee on Standards, Design and Methodology

August 4-5, 2016

AGENDA

Friday, August 5		
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*	

Attachment A -

AnLar Incorporated
1220 N. Fillmore Street, Ste. #330
Arlington, Virginia 22201

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

July 2016

Developed for the National Assessment Governing Board under
contract number ED-NAG-15-C-0001 by AnLar Incorporated.

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 Valididty 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 rigor 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 0.30, 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."[1] 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.[2] 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.

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

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

Identifying Information					Descriptive Characteristics			
Results of Eligibility Screening	**Year Published**	**Source of Study**	**Sample Size**	**Participant Grade(s)**	**Motivation Construct***	**Motivation Construct Categorization****	**Study Type**	**Alignment with Research Question(s)**
Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. Teachers College Record, 113(11), 2309-2344.	2011	Journal Article	2,612	12	Engagement; Effort	Effort; Value	Intervention	Q1, Q2, and Q3
Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.	2003	Journal Article	9,499	12	Motivation	Value; Expectancy	Observational	Q1 and Q2
Craig, M. (2013). Attribution theory in science achievement. (Doctoral dissertation). St. John's University, New York, NY.	2013	Dissertation	11,500	12	Effort; Self-Concept	N/A	Observational	Q1 and Q2
Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Washington, DC: National Center for Education Statistics.	1993	Technical Report (NCES)	26,669	4, 8, 12	Motivation; Effort	Effort; Value; Expectancy	Descriptive	Q2
Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.	2003	Technical Report-NAEP	84	8	Motivation	Value	Descriptive	Q2
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.	1992	Dissertation	3,058	8	Perception	Composite of Expectancy and Value	Observational	Q1 and Q2
Kiplinger, V. L., & 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.	1993	Technical Report (NCES)	80,836	8	Motivation	N/A	Intervention	Q1, Q2, and Q3

Identifying Information					Descriptive Characteristics			
Results of Eligibility Screening	Year Published	Source of Study	Sample Size	Participant Grade(s)	Motivation Construct*	Motivation Construct Categorization**	Study Type	Alignment with Research Question(s)
Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.	2013	Journal Article	160,486	8	Attitude; Self-concept	Expectancy; Value	Observational	Q1 and Q2
O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). Final report of experimental studies on motivation and NAEP test performance (CSE Tech. Rep. No. 427). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.***	1997	Technical Report (NCES)	1,468	8, 12	Motivation	Effort; Expectancy	Intervention	Q1, Q2, and Q3
O'Neil, Jr., H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.	1995	Journal Article	1,468	8, 12	Motivation	Effort; Expectancy	Intervention	Q1, Q2, and Q3
O'Sullivan, C. Y., & Weiss, A. R. (1999). Student work and teacher practices in science: A report on what students know and can do. Washington, DC: National Center for Education Statistics.	1999	Technical Report (NCES)	22,116	4, 8, 12	Motivation	Effort; Value; Expectancy	Descriptive	Q2
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.	2009	Technical Report	11,642	12	Motivation	Effort	Observational	Q1 and Q2
The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. (1991). Washington, DC: National Center for Education Statistics.	1991	Technical Report	8,902	4, 8, 12	Attitude; Perception	Effort; Expectancy; Value	Descriptive	Q2

Identifying Information					Descriptive Characteristics			
Results of Eligibility Screening	Year Published	Source of Study	Sample Size	Participant Grade(s)	Motivation Construct*	Motivation Construct Categorization**	Study Type	Alignment with Research Question(s)
Walberg, H. J., & Ethington, C. A. (1991). Correlates of writing performance and interest: A U.S. national assessment study. The Journal of Educational Research, 84(4), 198-203.	1991	Journal Article	288	12	Motivation	Value; Expectancy	Observational	Q1 and Q2
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.	1996	Technical Report	16	8	Motivation	Value	Observational	Q1 and Q2

* 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.

*** This study is a duplicate of O’Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). Experimental studies on motivation and NAEP test performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

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?

This question was addressed in a variety of ways by the four eligible intervention studies: Braun, Kirsch, and Yamamoto (2011); O'Neil, Sugrue, and Baker (1995), O'Neil, Sugrue, Abedi, Baker, and Golan (1997), and Kiplinger and Linn (1993).

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.

$$z = 0.5 \times 1\text{n}\left(\frac{1+r}{1-r}\right)$$

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 $\overline{X}_1$ and $\overline{X}_2$ are the sample means in the treatment and comparison groups and S_{within} is the within-groups standard deviation, pooled across groups.

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{within}}$$

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^*_i assigned to each correlation or treatment effect was computed as below, where $V^*_{Y_i}$ is the within-study variance for the *i*-th study, plus the between-study variance.

$$W_i^* = \frac{1}{V_{Y_i}^*}$$

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:

$$p^* = 2\left[1 - \Phi\ (|\,Z^*\,|)\right]$$

Note: Φ *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.

$$M = \frac{\sum_{f=1}^{k} p_i n_i}{\sum_{f=1}^{k} n_i}$$

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).

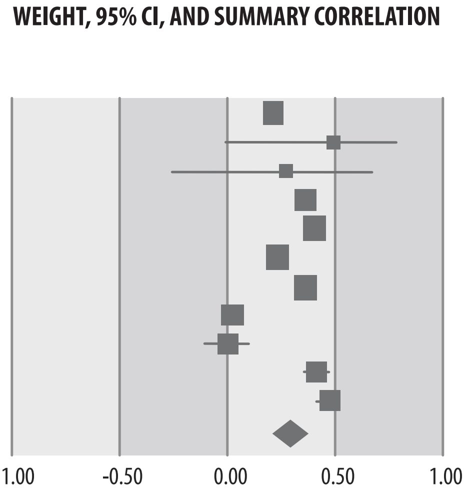
In studies with multiple motivation questions or grade levels, separate correlations were provided for each (e.g., liking math, thinking math is useful). As illustrated in Figure 1, the strongest correlations were found in Yepes-Baraya (8th grade, perceived block difficulty = .50), O'Neil et al. (12th 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

STUDY	**STATISTICS FOR EACH STUDY**				
	Correlation	Lower Limit	Upper Limit	z-Value	p-Value
Kim (1992) – Perception of math	0.22	0.21	0.23	68.64	0.00
Yepes-Baraya (1996) – Perceived ability	0.50	0.01	0.80	1.97	0.05
Yepes-Baraya (1996) – Perceived block difficulty	0.28	-0.25	0.68	1.04	0.30
Lee (2013) – Like to write	0.37	0.37	0.37	155.60	0.00
Lee (2013) – Good at writing	0.41	0.41	0.41	174.51	0.00
Lee (2013) – Writing is favorite	0.24	0.24	0.25	98.06	0.00
Byrnes (2003) – Ability/like math	0.37	0.35	0.39	37.85	0.00
Byrnes (2003) – Math utility/relevance	0.03	0.01	0.05	2.92	0.00
Walberg & ... (1991) – Math utility/imp.	0.00	-0.12	0.12	0.00	1.00
O'Neil et al. (1997) – General effort (8th)	0.42	0.36	0.48	11.25	0.00
O'Neil et al. (1997) – General effort (12th)	0.48	0.42	0.54	13.51	0.00
Summary Correlation	**0.30**	**0.23**	**0.36**	**8.19**	**0.00**

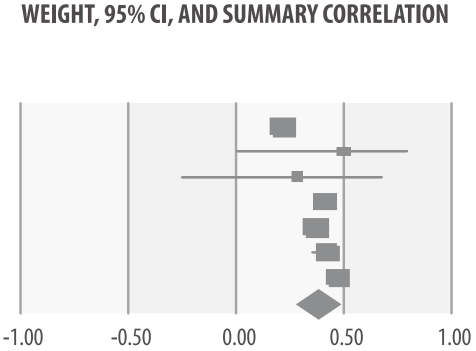


Disaggregating studies by motivation construct (i.e., expectancy vs. value) revealed that both summary correlations are statistically significant, but achievement is more strongly associated with expectancy (.38) than with value (.19).

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

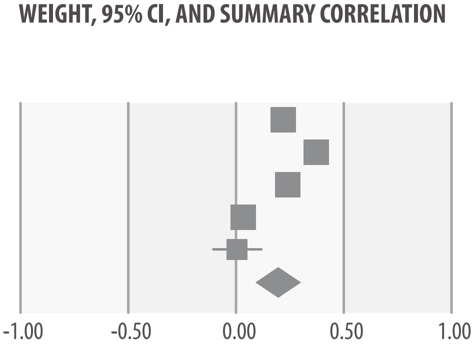
STUDY	STATISTICS FOR EACH STUDY				
	Correlation	Lower Limit	Upper Limit	z-Value	p-Value
Kim (1992) – Perception of math	0.22	0.21	0.23	68.64	0.00
Yepes-Baraya (1996) – Perceived ability	0.50	0.01	0.80	1.97	0.05
Yepes-Baraya (1996) – Perceived block difficulty	0.28	-0.25	0.68	1.04	0.30
Lee (2013) – Good at writing	0.41	0.41	0.41	174.51	0.00
Byrnes (2003) – Ability/liking math	0.37	0.35	0.39	37.85	0.00
O'Neil et al. (1997) – Effort (8th)	0.42	0.36	0.48	11.25	0.00
O'Neil et al. (1997) –Effort (12th)	0.48	0.42	0.54	13.51	0.00
Summary Correlation	**0.38**	**0.28**	**0.48**	**6.54**	**0.00**



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

STUDY	STATISTICS FOR EACH STUDY				
	Correlation	Lower Limit	Upper Limit	z-Value	p-Value
Kim (1992) – Perception of math	0.22	0.21	0.23	68.64	0.00
Lee (2013) – Like to write	0.37	0.37	0.37	155.60	0.00
Lee (2013) – Writing is favorite	0.24	0.24	0.25	98.06	0.00
Byrnes (2003) – Math utility/rel.	0.03	0.01	0.05	2.92	0.00
Walberg & ... (1991) – Math utility/impt.	0.00	-0.12	0.12	0.00	1.00
Summary Correlation	**0.19**	**0.10**	**0.27**	**3.98**	**0.00**



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 ($\beta_{concept} = 5.56$, $p < .001$). Test effort was also found to be related to achievement. Indeed, students who reported not 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")

Study	Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests	Grade	Sample Size
Braun, Kirsch, and Yamamoto (2011)	29%	12	2,612
Data Compendium (1993)	10%	4	8,738
Data Compendium (1993)	20%	8	9,432
Data Compendium (1993)	45%	12	8,499
O'Sullivan and Weiss (1999)	9%	4	22,116
O'Sullivan and Weiss (1999)	16%	8	22,116
O'Sullivan and Weiss (1999)	39%	12	22,116
Stokes and Cao (2009)	49%	12	11,642
Weighted Mean of Percentage	**25%**		

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")

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

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

Study	Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests	Grade	Sample Size
Data Compendium (1993)	10%	4	8,738
O'Sullivan and Weiss (1999)	9%	4	22,116
Weighted Mean of Percentage	**9%**		

Study	Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests	Grade	Sample Size
Data Compendium (1993)	20%	8	9,432
O'Sullivan and Weiss (1999)	16%	8	22,116
Weighted Mean of Percentage	**17%**		

Study	Percentage of Students Indicating That They Did Not Try As Hard on NAEP As on Other Tests	Grade	Sample Size
Braun, Kirsch, and Yamamoto (2011)	29%	12	2,612
Data Compendium (1993)	45%	12	8,499
O'Sullivan and Weiss (1999)	39%	12	22,116
Stokes and Cao (2009)	49%	12	11,642
Weighted Mean of Percentage	42%		

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)

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

** 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)

Study	Expectancy Statements	Average Student Rating of Math Ability Relative to Classmates' Math Ability	Grade	Sample Size
O'Neil et al. (1997)	"Compared to your classmates, your math ability is…" (1 = much lower than most of my classmates; 5 = much higher than most of my classmates)	64% of max score (mean = 3.20/5)	12	670
O'Neil et al. (1997)	"Compared to your classmates, your math ability is…" (1 = much lower than most of my classmates; 5 = much higher than most of my classmates)	68% of max score (mean = 3.40 /5)	8	634
Weighted Mean of Percentage		**66%**		

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

Study	Expectancy Statements*	Percentage of Students Who Agree with Statements	Grade	Sample Size
Data Compendium (1993)	Students who agree with "I am good at mathematics" (1992)	65%	4	8,738
Data Compendium (1993)	Students who agree with "I am good at mathematics" (1990)	64%	4	8,738
National Center for Education Statistics (1991)	Students who agree with "I am good at mathematics" (1990)	62%	4	8,902
O'Sullivan and Weiss (1999)	Students who agree with "I am good at science"	45%	4	22,116
Weighted Mean of Percentage		**55%**		

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

Study	Expectancy Statements*	Percentage of Students Who Agree or Strongly Agree with Statements	Grade	Sample Size
Data Compendium (1993)	Students who agree or strongly agree with "I am good at mathematics" (1992)	60%	8	9,432
Data Compendium (1993)	Students who agree or strongly agree with "I am good at mathematics" (1990)	62%	8	9,432
Kim (1992)	Composite of students who agree with "I like math" and "I am good at math"	67%	8	3,058
Lee (2013)	Students who agree or strongly agree with "I am good at writing"	51%	8	160,486
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990)	63%	8	8,888
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990 TSA)	62%	8	94,979
O'Sullivan and Weiss (1999)	Students who agree with "I am good at science"	47%	8	22,116
Weighted Mean of Percentage		**55%**		

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

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

** 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)

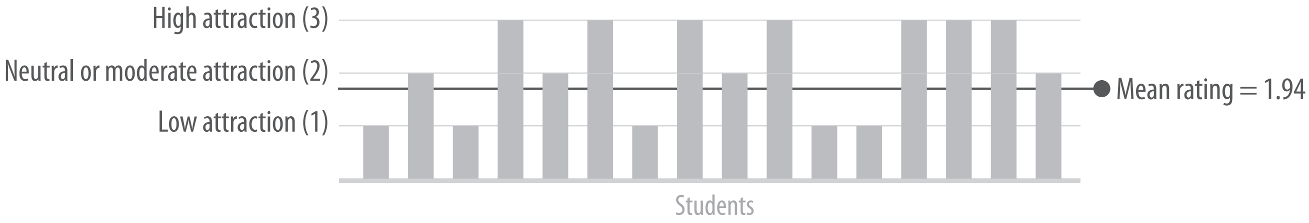
Study	Value Statements*	Percentage of Students Who Agree or Strongly Agree with Statements	Grade	Sample Size
Braun, Kirsch, and Yamamoto (2011)	Students indicating that it was important or very important to do well on the test	44%	12	2,612
Data Compendium (1993)	Students indicating that it was important or very important to do well on the test	89%	4	8,738
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1992)	57%	8	9,432
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1990)	57%	8	9,432
Data Compendium (1993)	Students indicating that it was important or very important to do well on the test	60%	8	9,432
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1992)	51%	12	8,499
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1990)	54%	12	8,499
Data Compendium (1993)	Students indicating that it was important or very important to do well on the test	34%	12	8,499
Jakwerth, Stancavage, and Reed (2003)	Students indicating that it was important or very important to do well on the test	73%	8	84
Lee (2013)	Students who agree or strongly agree with "I like to write"	52%	8	160,486
Lee (2013)	Students who agree or strongly agree with "Writing helps share ideas"	61%	8	160,486
National Center for Education Statistics (1991)	Students who agree with "I like mathematics" (1990)	67%	4	8,902
National Center for Education Statistics (1991)	Students who agree with "I like mathematics" (1990)	67%	4	8,902
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990)	56%	8	8,888
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990 TSA)	57%	8	94,979

Study	Value Statements*	Percentage of Students Who Agree or Strongly Agree with Statements	Grade	Sample Size
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990)	54%	12	8,862
O'Sullivan and Weiss (1999)	Students who agree with "I like science"	67%	4	22,116
O'Sullivan and Weiss (1999)	Students indicating that is was important or very important to do well on the test	85%	4	22,116
O'Sullivan and Weiss (1999)	Students who agree with "I like mathematics" (1992)	71%	4	22,116
O'Sullivan and Weiss (1999)	Students who agree with "I like mathematics" (1990)	70%	4	22,116
O'Sullivan and Weiss (1999)	Students who agree with "Science is useful for solving everyday problems"	35%	4	22,116
O'Sullivan and Weiss (1999)	Students who agree with "Science is useful for solving everyday problems"	40%	8	22,116
O'Sullivan and Weiss (1999)	Students who agree or strongly agree with "I like science"	50%	8	22,116
O'Sullivan and Weiss (1999)	Students indicating that it was important or very important to do well on the test	58%	8	22,116
O'Sullivan and Weiss (1999)	Students who agree with "Science is useful for solving everyday problems"	50%	12	22,116
O'Sullivan and Weiss (1999)	Students who agree with "I like science"	52%	12	22,116
O'Sullivan and Weiss (1999)	Students indicating that is was important or very important to do well on the test	34%	12	22,116
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

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

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

Study	Value Statements*	Percentage of Students Who Agree or Strongly Agree with Statements	Grade	Sample Size
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1992)	57%	8	9,432
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1990)	57%	8	9,432
Data Compendium (1993)	Students indicating that it was important or very important to do well on the test	60%	8	9,432
Jakwerth, Stancavage, and Reed (2003)	Students indicating that it was important or very important to do well on the test	73%	8	84
Lee (2013)	Students who agree or strongly agree with "I like to write"	52%	8	160,486
Lee (2013)	Students who agree or strongly agree with "Writing helps share ideas"	61%	8	160,486
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990)	56%	8	8,888
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990 TSA)	57%	8	94,979
O’Sullivan and Weiss (1999)	Students who agree with "Science is useful for solving everyday problems"	40%	8	22,116
O’Sullivan and Weiss (1999)	Students who agree with "I like science"	50%	8	22,116
O’Sullivan and Weiss (1999)	Students indicating that it was important or very important to do well on the test	58%	8	22,116
Weighted Mean of Percentage		54%		

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

Study	Value Statements*	Percentage of Students Who Agree or Strongly Agree with Statements	Grade	Sample Size
Braun, Kirsch, and Yamamoto (2011)	Students indicating that it was important or very important to do well on the test	44%	12	2,612
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1992)	51%	12	8,499
Data Compendium (1993)	Students who agree or strongly agree with "I like mathematics" (1990)	54%	12	8,499
Data Compendium (1993)	Students indicating that it was important or very important to do well on the test	34%	12	8,499
National Center for Education Statistics (1991)	Students who agree or strongly agree with "I like mathematics" (1990)	54%	12	8,862
O'Sullivan and Weiss (1999)	Students who agree with "Science is useful for solving everyday problems"	50%	12	22,116
O'Sullivan and Weiss (1999)	Students who agree with "I like science.	52%	12	22,116
O'Sullivan and Weiss (1999)	Students indicating that is was important or very important to do well on the test	34%	12	22,116
Weighted Mean of Percentage		**46%**		

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

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

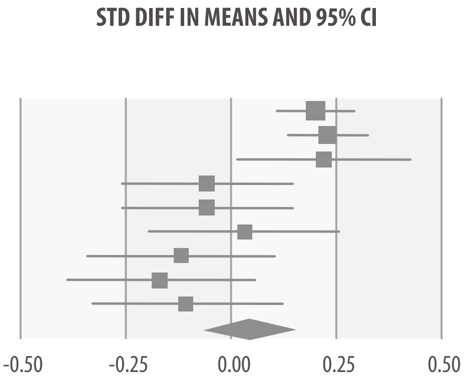
This question was answered through two random effects meta-analyses of treatment effects in intervention studies. One meta-analysis compared the self-reported effort of treatment groups that received different incentives to the self-reported effort of a control group. The other compared the NAEP achievement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

Figure 13: Intervention Effects on Students' Self-Reported Effort

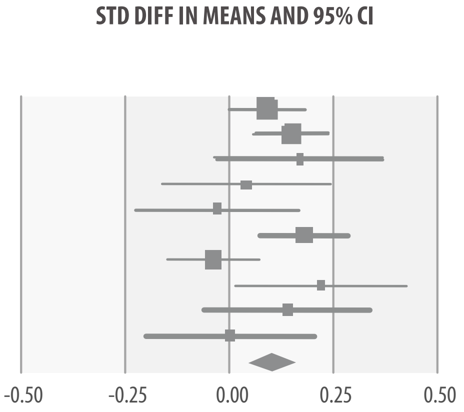
STUDY	STATISTICS FOR EACH STUDY						
	Std diff in means	Standard Error	Variance	Lower Limit	Upper Limit	z-Value	p-Value
Braun et al. (2011) – $20	0.20	0.05	0.00	0.10	0.30	4.09	0.00
Braun et al. (2011) – $5 + ≤ $30	0.22	0.05	0.00	0.13	0.33	4.72	0.00
O'Neil et al. (1995) – $1 (8th)	0.22	0.11	0.01	0.01	0.43	2.06	0.04
O'Neil et al. (1995) – Ego (8th)	-0.06	0.11	0.01	-0.27	0.15	-0.57	0.57
O'Neil et al. (1995) – Task (8th)	-0.06	0.11	0.01	-0.27	0.15	-0.57	0.57
O'Neil et al. (1995) – $1 (12th)	0.03	0.12	0.01	-0.20	0.26	0.26	0.80
O'Neil et al. (1995) – Ego (12th)	-0.12	0.12	0.01	-0.35	0.11	-1.03	0.30
O'Neil et al. (1995) – Task (12th)	-0.17	0.12	0.01	-0.40	0.06	-1.47	0.14
O'Neil et al. (1995) – Cert. (12th)	-0.11	0.12	0.01	-0.34	0.12	-0.94	0.35
Summary Effect	**0.04**	**0.06**	**0.00**	**-0.07**	**0.15**	**0.67**	**0.50**



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.

Figure 14: Intervention Effects on NAEP Achievement

STUDY	STATISTICS FOR EACH STUDY						
	Std diff in means	Standard Error	Variance	Lower Limit	Upper Limit	z-Value	p-Value
Braun et al. (2011) – $20	0.09	0.05	0.00	-0.01	0.19	1.55	0.06
Braun et al. (2011) – $5 + ≤ $30	0.15	0.05	0.00	0.06	0.25	2.11	0.00
O'Neil et al. (1995) – $1	0.17	0.11	0.01	-0.03	0.38	1.60	0.11
O'Neil et al. (1995) – Ego	0.04	0.11	0.01	-0.17	0.25	0.38	0.70
O'Neil et al. (1995) – Task	-0.03	0.10	0.01	-0.23	0.17	-0.26	0.77
Kiplinger & Linn (1993) – Blk. 1	0.18	0.06	0.00	0.07	0.30	3.08	0.00
Kiplinger & Linn (1993) – Blk. 2	-0.04	0.06	0.00	-0.16	0.08	-0.88	0.49
O'Neil et al. (1997) - $1	0.22	0.11	0.01	0.01	0.43	2.08	0.04
O'Neil et al. (1997) – Ego	0.14	0.11	0.01	-0.07	0.35	1.34	0.18
O'Neil et al. (1997) – Task	0.00	0.10	0.01	-0.20	0.20	0.00	1.00
Summary Effect	**0.10**	**0.02**	**0.00**	**0.05**	**0.14**	**4.30**	**0.00**



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

Blagg, K., & Chingos, M. (May 2016). Varsity Blues: Are High School Students Being Left Behind? The Urban Institute.

Brophy, J. & Ames, C. (September 2005). NAEP Testing for Twelfth Graders: Motivational Issues. Prepared for the National Assessment Governing Board. Michigan State University.

Byrnes, J.P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.

Craig, M. (2013). Attribution theory in science achievement. (Unpublished doctoral dissertation). St. John's University, New York, NY.

Data compendium for the NAEP 1992 mathematics assessment of the nation and the states. (1993). Educational Testing Service, Washington, DC.; National Assessment of Educational Progress, Princeton, NJ.

Dweck, C. (2006). *Mindset: The new psychology of success*. Random House.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of educational research, 75(3), 417-453.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.

Institute of Education Sciences. (2014). Procedures and Standards Handbook Version 3.0. What Works Clearinghouse. http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.

Kim, L. Y. (1992). Factors affecting student learning outcomes: A school-level analysis of the 1990 NAEP mathematics trial state assessment. (Unpublished doctoral dissertation). University of Southern California, Los Angeles, CA.

Kiplinger, V. L., & 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.

Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.

National Center for Education Statistics (1991). The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

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.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). Experimental studies on motivation and NAEP test performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). Final report of experimental studies on motivation and NAEP test performance (CSE Tech. Rep. No. 427). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

O'Neil, Jr, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.

Osborne, J. (2010). Correlation and other measures of association. In Hancock, G. R. & Mueller, R. O. (eds.) *Reviewer's guide to quantitative methods*. Routledge: New York. (pp-55-69).

O'Sullivan, C. Y., & Weiss, A. R. (1999). Student work and teacher practices in science: A report on what students know and can do. Washington, DC: National Center for Education Statistics.

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.

Walberg, H. J., & Ethington, C. A. (1991). Correlates of writing performance and interest: A U.S national assessment study. *The Journal of Educational Research*, 84(4), 198-203.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. C*ontemporary Educational Psychology*, 25(1), 68-81.

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.

IX. ENDNOTES

[1] Student Engagement in the National Assessment of Educational Progress (NAEP): Critical Review and Synthesis of Research. (June 18, 2015). National Assessment Governing Board Solicitation ED-NAG-15-R-0004.

[2] Institute of Education Sciences. (2014). WWC Procedures and Standards Handbook v.3.0, Table 111.1, page 12.

Appendix A. Design Document

LITERATURE REVIEW DESIGN DOCUMENT

Contract # ED-NAG-15-C-0001

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

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 reported[4])
- Study design (i.e., observational design, intervention study)

[3] Osborne, J. (2010). Correlation and other measures of association. In Hancock, G. R. & Mueller, R. O. (eds.) Reviewer's guide to quantitative methods. Routledge: New York. (pp-55-69).

[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

  - Data findings summary
  - Data findings critique
  - Inferences/conclusions summary
  - Additional information, questions, or concerns
  - 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.[5] 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[6] 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.

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

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:

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

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

Description: In text box, type in the APA-compliant citation. APA citation examples for various sources can be found at http://www.umuc.edu/library/libhow/apa_examples.cfm. One example is: O'Neil, H., Abedi, J, Lee, C., Miyoshi, J., & Mastergeorge, A. (April 2004). Monetary Incentives for Low-Stakes Tests (CSE Report 625). Center for the Study of Evaluation National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

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
    - 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: Observation al 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 = (3117 – 1719) / 3117 = 45%

TA = (1565-884)/1565 = 43.5%

CA = (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.

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}$$

s, the pooled standard deviation, as (for two independent samples):

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

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

Study Citation	Status after Phase 2
Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. Teachers College Record, 113(11), 2309-2344.	Proceed to Phase 3
Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.	Proceed to Phase 3
Cohen, A. S., Cho, S., Li, F., Shutz, P., Hong, J. Y. (2005). A mixture IRT model analysis of Grade 12 examinee motivation on the 2002 NAEP Reading Test. Athens, GA: Georgia Center for Assessment, University of Georgia.	Eliminate- psychometric study that does not address the research questions
Cohen, A., Li, F., & Cho, S. (2005). A mixture model analysis of examinee motivation on a standardized achievement test. Athens, GA: Georgia Center for Assessment, University of Georgia.	Eliminate- psychometric study that does not address the research questions
Craig, M. (2013). Attribution theory in science achievement. (Doctoral dissertation). St. John's University, New York, NY.	Proceed to Phase 3
Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Washington, DC: National Center for Education Statistics.	Proceed to Phase 3
Freund, D. S., & Rock, D. A. (1992). A preliminary investigation of pattern-marking in 1990 NAEP data. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).	Eliminate- psychometric study that does not address the research questions
Guthrie, J. T., Schafer, W. D., & Huang, C. W. (2001). Benefits of opportunity to read and balanced instruction on the NAEP. The Journal of Educational Research, 94(3), 145-162.	Eliminate- does not address any of the research questions; focus on "engaged reading"
Hoffman, R. G., & Trippe, D. M. (2005). The impact of grade 12 students' non-response to NAEP open-response items. Washington, DC: National Center for Education Statistics.	Eliminate- did not answer research questions; analyzes non-response on tests
Hong, J. Y., Li, F., Cho, S., Schutz, P. A., & Cohen, A. S. (2006). Why students do not respond to NAEP reading questions: The relationship between students' response patterns and reading motivation. Sun-Chung-Uh-Mun, 34, 179-199.	Eliminate- psychometric study that does not address the research questions
Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.	Proceed to Phase 3
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.	Proceed to Phase 3
Kiplinger, V. L., & 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.	Proceed to Phase 3
Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.	Proceed to Phase 3
Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. Large-scale Assessments in Education, 2(1), 1-24.	Eliminate- psychometric study that does not address the research questions

Study Citation	Status after Phase 2
Mullis, I. V. S., & Stancavage, F. (n.d.). Analyzing the NAEP data on testing conditions in schools.	Eliminate- not original, empirical research
Ogut, B., Walton, E., & Dogan, E. (2010). Examining 12th-graders' engagement/motivation in NAEP mathematics assessment using data from high-stakes assessments. Washington, DC: NAEP Education Statistics Service Institute.	Eliminate- study measures are not a reliable proxy for motivation
O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). Final report of experimental studies on motivation and NAEP test performance (CSE Tech. Rep. No. 427). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.	Proceed to Phase 3
O'Neil, Jr., H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.	Proceed to Phase 3
O'Sullivan, C. Y., & Weiss, A. R. (1999). Student work and teacher practices in science: A report on what students know and can do. Washington, DC: National Center for Education Statistics.	Proceed to Phase 3
Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1998). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Landover, MD: National Academy Press.	Eliminate- does not contain original, empirical research
Qian, J. (2014). An investigation of position effects in large-scale writing assessments. Applied Psychological Measurement, 38(7), 518-534.	Eliminate- does not address the research questions; focus on effects of question position
Schultz, S. R., Deatz, R. C., & Gladden, F. L. (2008). NAEP-QA grade 12 motivation study: Summary of assessment site visits (Report No. FR-08-10). Alexandria, VA: Human Resources Research Organization.	Eliminate- intervention study that did not pass methodological rigor because of inability to calculate effect size
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.	Proceed to Phase 3
The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. (1991). Washington, DC: National Center for Education Statistics.	Proceed to Phase 3
Walberg, H. J., & Ethington, C. A. (1991). Correlates of writing performance and interest: A U.S. national assessment study. The Journal of Educational Research, 84(4), 198-203.	Proceed to Phase 3
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.	Proceed to Phase 3

Appendix D. Systematic Review Table

Figure 2: Systematic Review Table — Descriptive Characteristics

Identifying Information					Descriptive Characteristics								
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Sample Size	Participant Grade(s)	Assessment Type	Assessment Subject Area	Administration Mode	Motivation Construct*	Motivation Construct Categorization**	Number of Citations	Study Type
Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. Teachers College Record, 113(11), 2309-2344.	2011	Journal Article	Princeton University; NCES (U.S. ED); Educational Testing Service	not specified	2,612	12	NAEP	ELA	Paper & pencil	Engagement; Effort	Effort; Value	23	Intervention
Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.	2003	Journal Article	NCES	1992	9,499	12	NAEP	Mathematics	Paper & pencil	Motivation	Value; Expectancy	90	Observational
Craig, M. (2013). Attribution theory in science achievement. (Doctoral dissertation). St. John's University, New York, NY.	2013	Dissertation	N/A	2009	11,500	12	NAEP	Science	Paper & pencil	Effort; Self-Concept	N/A	Not available	Observational
Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Washington, DC: National Center for Education Statistics.	1993	Technical Report (NCES)	NCES	1992	26,669	4; 8; 12	NAEP	Mathematics	Paper & pencil	Motivation; Effort	Effort; Value; Expectancy	Not available	Descriptive
Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.	2003	Technical Report-NAEP	NAEP Validity Studies	1999	84	8	NAEP	Reading; Civics	Paper & pencil	Motivation	Value	13	Descriptive
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.	1992	Dissertation	N/A	1990	3,058	8	NAEP	Mathematics	Paper & pencil	Perception	Composite of Expectancy and Value	4	Observational
Kiplinger, V. L., & 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.	1993	Technical Report (NCES)	NCES	1990; 1992	80,836	8	NAEP	Mathematics	Paper & pencil	Motivation	N/A	53	Intervention
Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.	2013	Journal Article	none reported	1998; 2007	160,486	8	NAEP	ELA (Writing)	Paper & pencil	Attitude; Self-concept	Expectancy; Value	11	Observational

Identifying Information					Descriptive Characteristics								
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Sample Size	Participant Grade(s)	Assessment Type	Assessment Subject Area	Administration Mode	Motivation Construct*	Motivation Construct Categorization**	Number of Citations	Study Type
O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). Final report of experimental studies on motivation and NAEP test performance (CSE Tech. Rep. No. 427). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.***	1997	Technical Report (NCES)	NCES	1992	1,468	8; 12	NAEP	Mathematics	Paper & pencil	Motivation	Effort; Expectancy	N/A	Intervention
O'Neil, Jr., H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.	1995	Journal Article	NCES	1992	1,468	8; 12	NAEP	Mathematics	Paper & pencil	Motivation	Effort; Expectancy	89	Intervention
O'Sullivan, C. Y., & Weiss, A. R. (1999). Student work and teacher practices in science: A report on what students know and can do. Washington, DC: National Center for Education Statistics.	1999	Technical Report (NCES)	NCES	1996	22,116	4; 8; 12	NAEP	Science	Paper & pencil	Motivation	Effort; Value; Expectancy	21	Descriptive
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.	2009	Technical Report	U.S. Department of Education	2005	11,642	12	NAEP	ELA	Paper & pencil	Motivation	Effort	Not available	Observational
The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. (1991). Washington, DC: National Center for Education Statistics.	1991	Technical Report	NCES	1990	8,902	4, 8, 12	NAEP	Mathematics	Paper & pencil	Attitude; Perception	Effort; Expectancy; Value	110	Descriptive
Walberg, H. J., & Ethington, C. A. (1991). Correlates of writing performance and interest: A U.S. national assessment study. The Journal of Educational Research, 84(4), 198-203.	1991	Journal Article	none reported	N/A	288	12	NAEP	ELA (Writing)	Paper & pencil	Motivation	Value; Expectancy	10	Observational
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.	1996	Technical Report	NAEP (administered by the Office of Educational Research and Improvement, U.S. ED)	1995	16	8	NAEP	Science	Paper & pencil	Motivation	Value	11	Observational

Figure 3: Systematic Review Table — Study Characteristics

Identifying Information					Study Characteristics									
Reference	**Year Published**	**Source of Study**	**Funding Entity**	**Year(s) Data Collected**	**Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/ PIRLS/PISA**	**Direction of Treatment Effect on Motivation**	**Magnitude of Relationship Between Motivation and Achievement on NAEP/ TIMSS/PIRLS/ PISA**	**Magnitude of Treatment Effect/Effect Size**	**p-Value of Relationship/ Effect**	**Statistically Significant Relationship/Effect (α=0.05)?**	**Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?**	**Low Attrition (RCT Intervention Studies Only)?**	**Baseline Equivalence Established (QED Intervention) Studies Only)?**	**Alignment with Research Question(s)**
Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. Teachers College Record, 113(11), 2309-2344.	2011	Journal Article	Princeton University; NCES (U.S. ED); Educational Testing Service	not specified	N/A	Positive	N/A	0.15	N/A	Yes	Yes	No	No	Q1, Q2, and Q3
Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.	2003	Journal Article	NCES	1992	Positive	N/A	"correlation (ability-liking math) = .37 correlation (utility-relevance of math) = .03 correlation (math is fact-learning belief) = -.36"	N/A	p < .001	"Yes (ability; math is fact) No (utility)"	Yes	N/A	N/A	Q1 and Q2
Craig, M. (2013). Attribution theory in science achievement. (Doctoral dissertation). St. John's University, New York, NY.	2013	Dissertation	N/A	2009	Positive	N/A	beta= -7.153 (students who reported they did not exert effort were likely to score 7 units lower than the mean science score)	N/A	p <0.001	Yes	Yes	N/A	N/A	Q1 and Q2

Identifying Information					Study Characteristics									
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Direction of Treatment Effect on Motivation	Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Magnitude of Treatment Effect/Effect Size	p-Value of Relationship/Effect	Statistically Significant Relationship/Effect ($\alpha=0.05$)?	Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?	Low Attrition (RCT Intervention Studies Only)?	Baseline Equivalence Established (QED Intervention) Studies Only)?	Alignment with Research Question(s)
Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Washington, DC: National Center for Education Statistics.	1993	Technical Report (NCES)	NCES	1992	N/A	N/A	N/A	N/A	N/A	N/A	Yes	N/A	N/A	Q2
Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. (Working Paper No. 2003-12). NAEP Validity Studies.	2003	Technical Report-NAEP	NAEP Validity Studies	1999	N/A	N/A	N/A	N/A	N/A	N/A	Yes	N/A	N/A	Q2
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.	1992	Dissertation	N/A	1990	Positive	N/A	0.22	N/A	$p < 0.05$	Yes	Yes	N/A	N/A	Q1 and Q2

Identifying Information					Study Characteristics									
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Direction of Treatment Effect on Motivation	Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Magnitude of Treatment Effect/Effect Size	p-Value of Relationship/Effect	Statistically Significant Relationship/Effect (α=0.05)?	Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?	Low Attrition (RCT Intervention Studies Only)?	Baseline Equivalence Established (QED Intervention) Studies Only)?	Alignment with Research Question(s)
Kiplinger, V. L., & 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.	1993	Technical Report (NCES)	NCES	1990; 1992	N/A	Positive (for items 1-9)	N/A	0.18 (for items 1-9)	Not available	"Yes (for items 1-9) No (for items 10-17)"	Yes	Cannot tell	No	Q1, Q2, and Q3
Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.	2013	Journal Article	none reported	1998; 2007	Positive (but significant variation by gender)	N/A	"Like to write (Cohen's d = .8) Good at writing (Cohen's d = .9) Writing is one of my favorite activities (Cohen's d = .5)"	N/A	Not available	Yes	Yes	N/A	N/A	Q1 and Q2
O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). Experimental studies on motivation and NAEP test performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.	1997	Technical Report (NCES)	NCES	1992	N/A	Positive (for 8th grade)	N/A	"0.22 ($1 incentive) 0.14 (Ego incentive) 0.00 (Task incentive)"	Not available	Yes (for 8th grade)	Cannot tell	Cannot tell	Cannot tell	Q1, Q2, and Q3

Identifying Information					Study Characteristics									
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Direction of Treatment Effect on Motivation	Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Magnitude of Treatment Effect/Effect Size	p-Value of Relationship/Effect	Statistically Significant Relationship/Effect (α=0.05)?	Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?	Low Attrition (RCT Intervention Studies Only)?	Baseline Equivalence Established (QED Intervention) Studies Only)?	Alignment with Research Question(s)
O'Neil, Jr., H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.	1995	Journal Article	NCES	1992	N/A	Positive (for 8th grade)	N/A	"0.17 ($1 incentive) 0.04 (Ego incentive) -0.03 (Task incentive)"	Not available	Yes (for 8th grade)	Cannot tell	Cannot tell	Cannot tell	Q1, Q2, and Q3
O'Sullivan, C. Y., & Weiss, A. R. (1999). Student work and teacher practices in science: A report on what students know and can do. Washington, DC: National Center for Education Statistics.	1999	Technical Report (NCES)	NCES	1996	N/A	N/A	N/A	N/A	N/A	N/A	Yes	N/A	N/A	Q2
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.	2009	Technical Report	U.S. Department of Education	2005	Positive	N/A	N/A	N/A	p = ~ 0	Yes	Yes	N/A	N/A	Q1 and Q2

Identifying Information					Study Characteristics									
Reference	Year Published	Source of Study	Funding Entity	Year(s) Data Collected	Nature of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Direction of Treatment Effect on Motivation	Magnitude of Relationship Between Motivation and Achievement on NAEP/TIMSS/PIRLS/PISA	Magnitude of Treatment Effect/Effect Size	p-Value of Relationship/Effect	Statistically Significant Relationship/Effect (α=0.05)?	Met Minimum Level of Criteria for Either Osborn or WWC Frameworks?	Low Attrition (RCT Intervention Studies Only)?	Baseline Equivalence Established (QED Intervention) Studies Only)?	Alignment with Research Question(s)
The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. (1991). Washington, DC: National Center for Education Statistics.	1991	Technical Report	NCES	1990	N/A	N/A	N/A	N/A	N/A	N/A	Yes	N/A	N/A	Q2
Walberg, H. J., & Ethington, C. A. (1991). Correlates of writing performance and interest: A U.S. national assessment study. The Journal of Educational Research, 84(4), 198-203.	1991	Journal Article	none reported	N/A	No correlation	N/A	correlation = 0	N/A	Not available	No	Yes	N/A	N/A	Q1 and Q2
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.	1996	Technical Report	NAEP (administered by the Office of Educational Research and Improvement, U.S. ED)	1995	"Standard block scores and perceived ability: positive and weak; Block scores and perceived block difficulty: positive and weak"	N/A	"Standard block score and perceived ability: .601 (Pearson's); .501 (Spearman) Standard block score and perceived block difficulty: .136 (Pearson's) .280 (Spearman)"	N/A	"Standard block score and perceived ability: p = .014 (Pearson's); p = .048 (Spearman) Standard block score and perceived block difficulty: p = .615 (Pearson's); p = .293 (Spearman)"	"Yes (perceived ability) No (perceived block difficulty)"	Yes	N/A	N/A	Q1 and Q2

* This study is a duplicate of O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). Experimental studies on motivation and NAEP test performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

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

Braun, H., Kirsch, I, & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. Teachers College Record, 113(11), 2309-2344.

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.

Byrnes, J. P. (2003). Factors predictive of mathematics achievement in white, black, and Hispanic 12th graders. Journal of Educational Psychology, 95(2), 316-326.

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.

Craig, M. (2013). Attribution Theory in Science Achievement. (Ed.D. Dissertation, St. John's University (New York), School of Education and Human Services).

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).[1] 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."

Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States. (1993). Educational Testing Service, Washington, DC.; National Assessment of Educational Progress, Princeton, NJ.

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.

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An Investigation of Why Students Do Not Respond to Questions. (Working Paper No. 2003-12). NAEP Validity Studies.

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.

Kim, L. Y. (1992). Factors Affecting Student Learning Outcomes: A School-Level Analysis of the 1990 NAEP Mathematics Trial State Assessment. (Ed.D. Dissertation, University of Southern California).

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 (.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 though 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.

Kiplinger, V. L., & 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.

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, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. Written Communication, 30(2), 164-193.

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.

National Center for Education Statistics (NCES). (1991). The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

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.

O'Neil, Jr, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. Educational Assessment, 3(2), 135-157.

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.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). Final report of experimental studies on motivation and NAEP test performance (CSE Tech. Rep. No. 427). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.[2]

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

[2] Also cited as O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). Experimental studies on motivation and NAEP test performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

(mean = 3.53) 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.

O'Sullivan, C. Y., & Weiss, A. R. (1999). Student Work and Teacher Practices in Science: A Report on What Students Know and Can Do. National Center for Education Statistics.

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.

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

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 ($p = \sim 0$). 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.

Walberg, H. J., & Ethington, C. A. (1991). Correlates of Writing Performance and Interest: A US National Assessment Study. The Journal of Educational Research, 84(4), 198-203.

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.

Yepes-Baraya, M. (1996). A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment. National Assessment of Educational Progress.

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.

IES NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

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.

Attachment C

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

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[1] 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[2] 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.

August 2016

DISCUSSION DRAFT

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.

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

				Reading				
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	35%	4%	3%	3%	#[1]	#	1%	**46%**
Female	39%	5%	5%	4%	#	#	1%	**54%**
***Total*[2]**	**75%**	**8%**	**8%**	**7%**	**#**	**#**	**2%**	**100%**
							Overall Match Rate	**74%**
				Mathematics				
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	36%	4%	3%	3%	#	#	1%	**47%**
Female	38%	5%	5%	3%	#	#	1%	**53%**
***Total*[2]**	**74%**	**9%**	**9%**	**6%**	**#**	**#**	**2%**	**100%**
							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 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).

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.

[2] 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 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

Subject	**SAT Benchmark**	**Mean**	**Percentage**	**SD**	**Percentile**		**IQR[1]**
					25th	***75th***	
Reading	*Below*	282	48%	29	263	301	38
	At	304	4%	23	289	319	30
	At or Above	323	52%	27	304	340	36
Mathematics	*Below*	147	42%	21	134	161	27
	At	166	4%	13	157	175	18
	At or Above	187	58%	21	172	200	28

NOTES: [1]IQR 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.

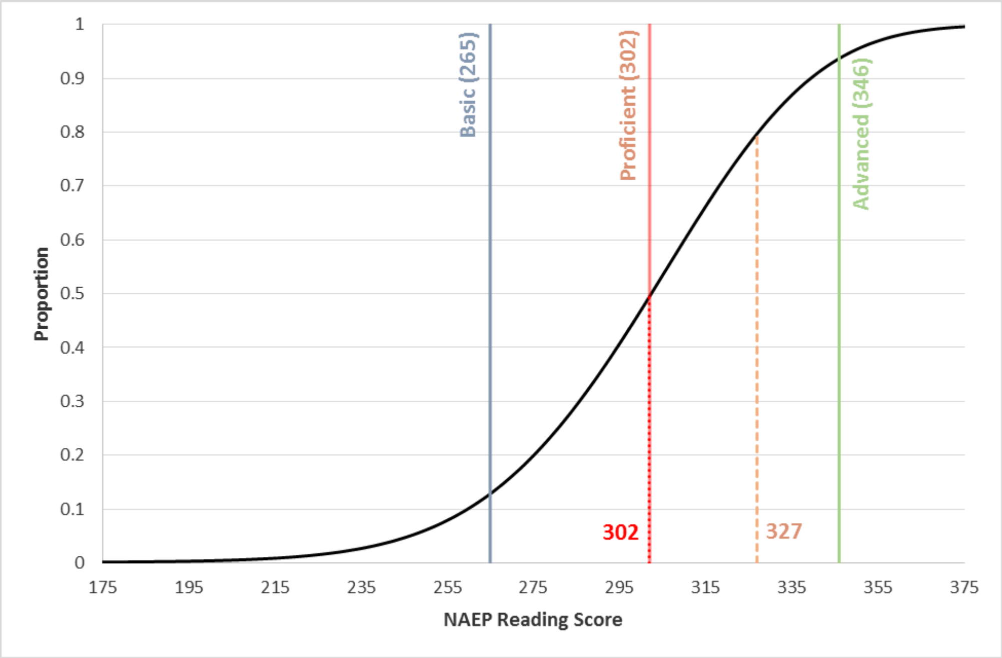


Figure 1: Proportion of students meeting the SAT critical reading benchmark of 500 in Massachusetts for NAEP reading scores

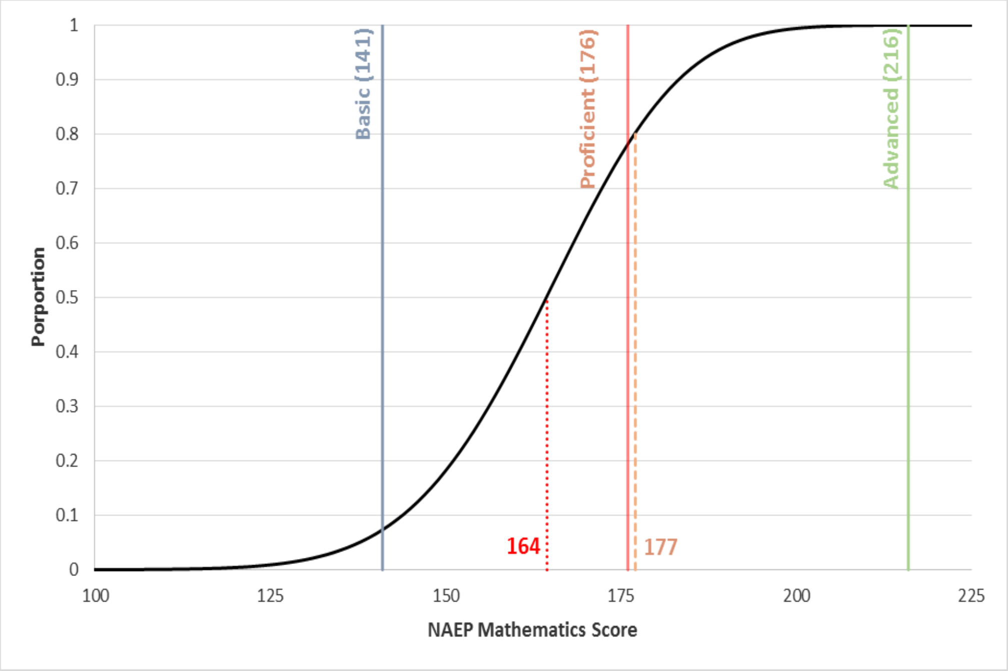


Figure 2: Proportion of students meeting the SAT mathematics benchmark of 500 in Massachusetts for NAEP mathematics scores

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.

Subject	NAEP *Proficient*	Mean	Percentage	SD	Percentile		IQR[1]
					25th	*75th*	
Critical Reading	*Below*	431	47%	89	370	490	120
	At or Above[2]	565	53%	92	500	620	120
Mathematics	*Below*	452	57%	78	400	500	100
	At or Above	610	43%	78	550	660	110

NOTES: [1]IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2]The "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.

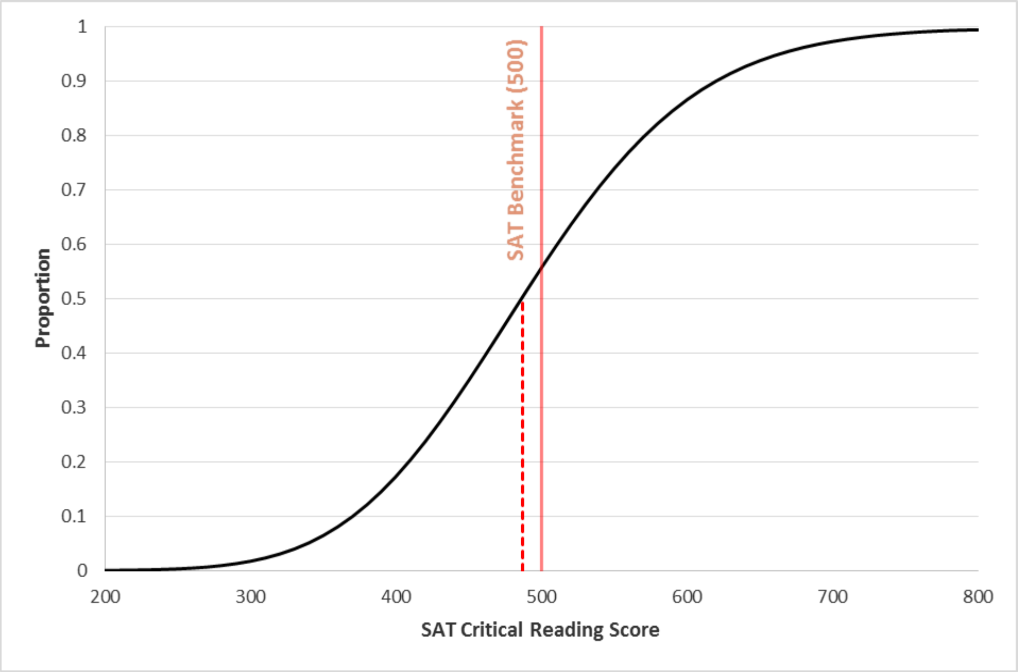


Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Massachusetts for SAT critical reading scores

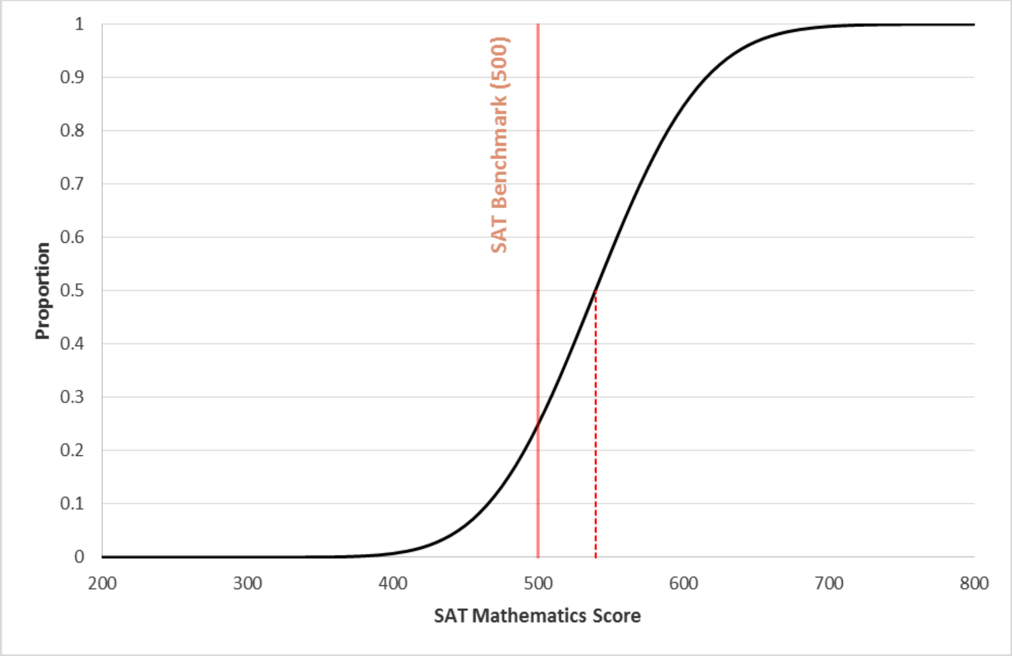


Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Massachusetts for SAT mathematics scores

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

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moran, R., Oranje, A. H., & Freund, D. S. (2011). *NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between NAEP and SAT* (Technical Report for NAEP 12th Grade Preparedness Research). Retrieved from https://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/SAT-NAEP_Linking_Study.pdf

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report.*

The SAT® College and Career Readiness Benchmark User Guidelines (http://media.collegeboard.com/digitalServices/pdf/sat/12b_6661_SAT_Benchmarks_PR_120914.pdf)

Wyatt, J., Kobrin, J., Wiley, A., Camara, W. J., & Proestler, N. (2011). *SAT Benchmarks: Development of a College Readiness Benchmark and its Relationship to Secondary and Postsecondary School Performance* (Research Report 2011-5). Newtown, PA: College Board.

August 2016

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). 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). 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

Reading

	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total[2]***
Male	38%	6%	2%	2%	#[1]	#	1%	**49%**
Female	39%	7%	2%	1%	#	#	1%	**51%**
Total[2]	**77%**	**13%**	**5%**	**3%**	**1%**	**#**	**1%**	**100%**
							Overall Match Rate	**95%**

Mathematics

	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total[2]***
Male	39%	6%	2%	1%	#	#	1%	**50%**
Female	38%	7%	2%	2%	#	#	1%	**50%**
Total[2]	**77%**	**13%**	**5%**	**3%**	**1%**	**#**	**1%**	**100%**
							Overall Match Rate	**95%**

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., 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)[2]. 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[3].

[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**	**Percentile**		**IQR**[1]
					25th	***75th***	
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: [1]IQR 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.

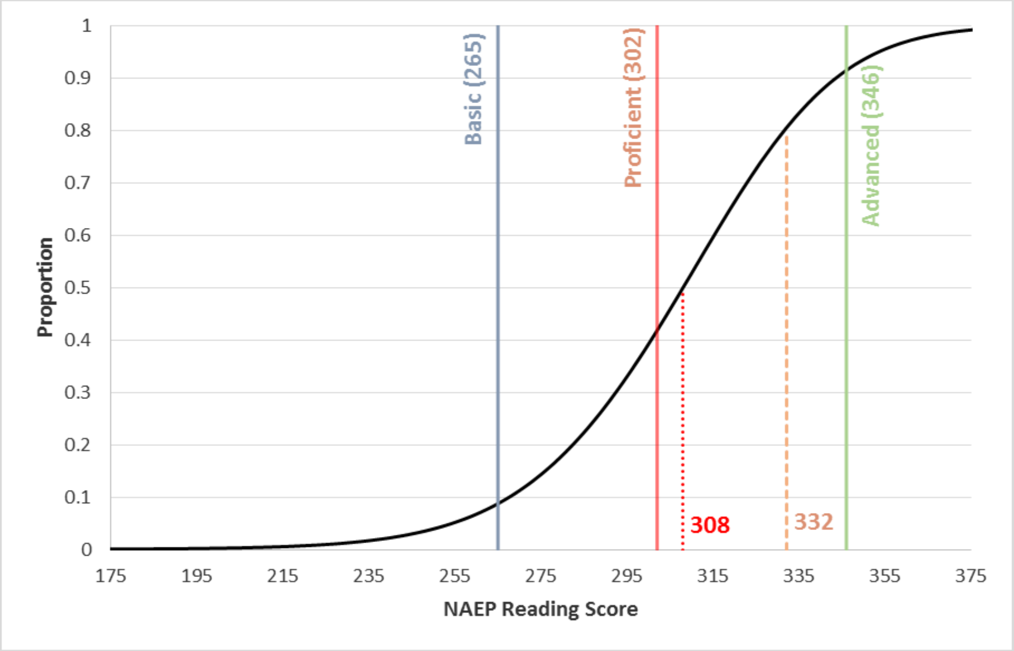


Figure 1: Proportion of students meeting the ACT reading benchmark of 22 in Michigan for NAEP reading scores

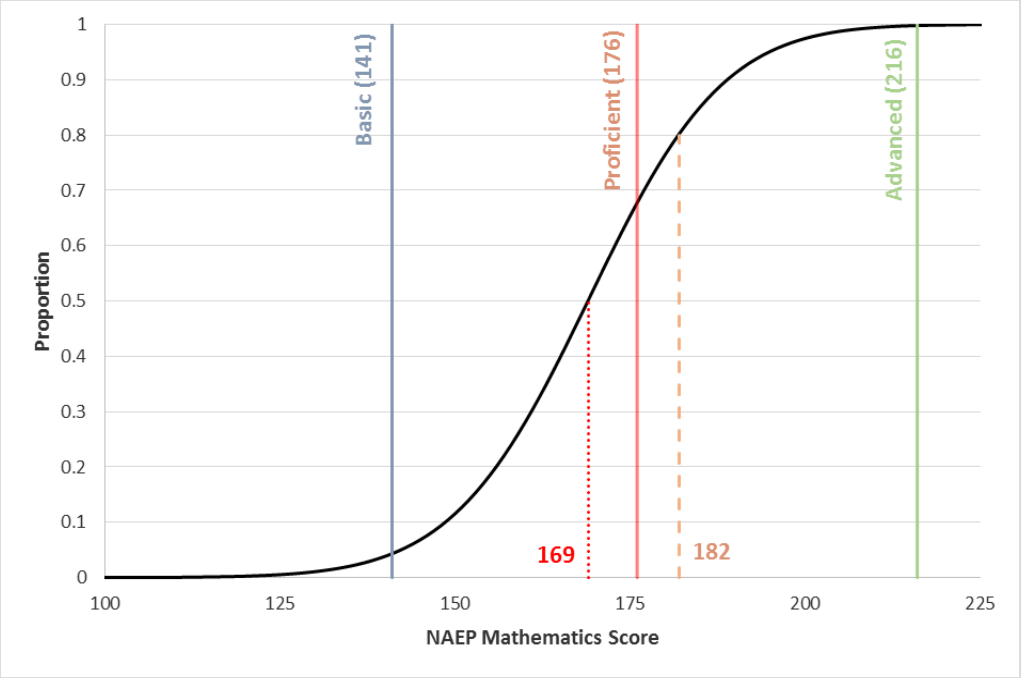


Figure 2: Proportion of students meeting the ACT mathematics benchmark of 22 in Michigan for NAEP mathematics scores

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

Subject	NAEP *Proficient*	Mean	Percentage	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	17	61%	4	13	19	6
	At or Above[2]	25	39%	5	20	28	8
Mathematics	*Below*	18	72%	3	15	20	5
	At or Above	26	28%	4	23	28	5

NOTES: [1]IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2]The "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.

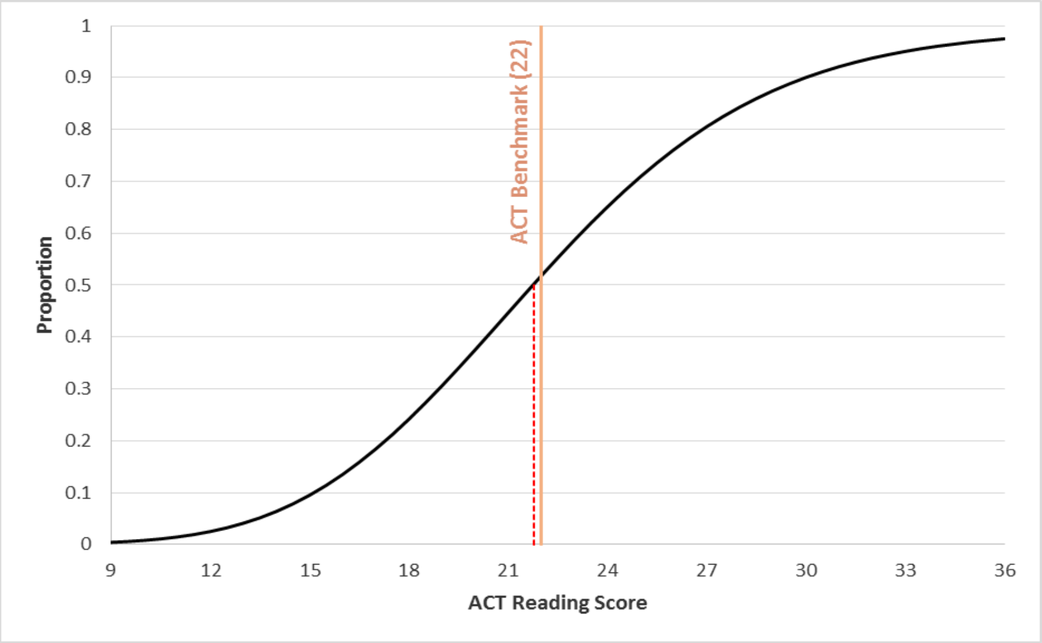


Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Michigan for ACT reading scores

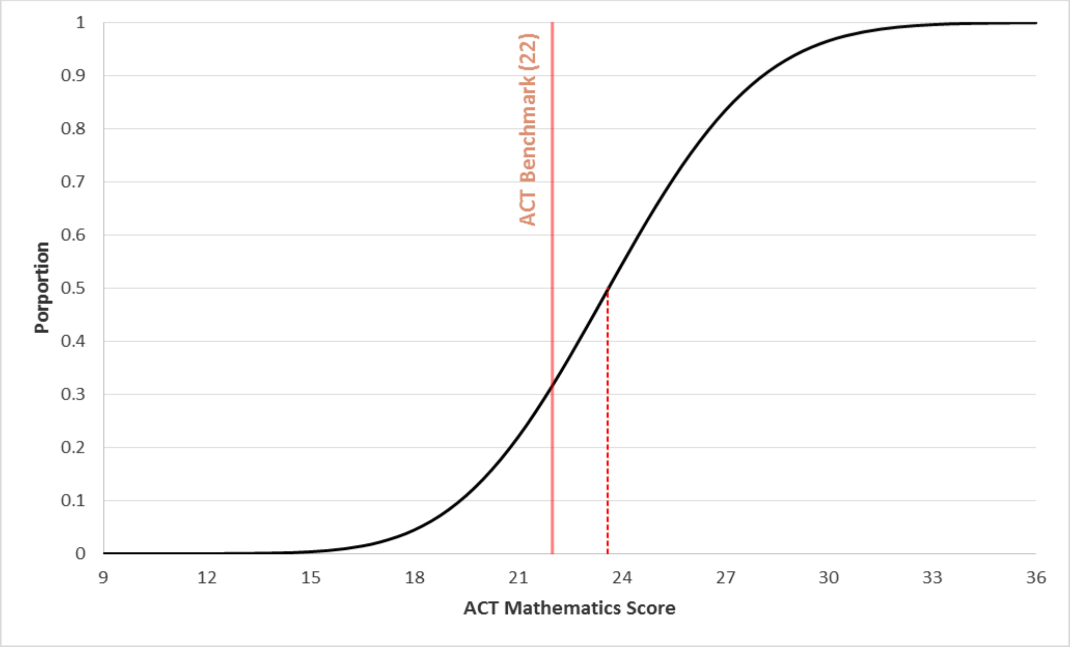


Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Michigan for ACT mathematics scores

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

ACT (2013). *What are the ACT College Readiness Benchmarks?* (http://www.act.org/content/dam/act/unsecured/documents/benchmarks.pdf).

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report.*

August 2016

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 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 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 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 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

Reading

	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	36%	10%	2%	1%	#[1]	#	#	**49%**
Female	35%	13%	2%	1%	#	#	#	**51%**
***Total*[2]**	**70%**	**23%**	**4%**	**2%**	**#**	**#**	**1%**	**100%**
							Overall Match Rate	**89%**

Mathematics

	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	36%	10%	2%	1%	#	#	#	**50%**
Female	35%	12%	2%	1%	#	#	#	**50%**
***Total*[2]**	**71%**	**22%**	**4%**	**2%**	**#**	**#**	**1%**	**100%**
							Overall Match Rate	**90%**

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., 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)[1]. 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].

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[3] 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

Subject	ACT Benchmark	Mean	Percentage	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	269	63%	29	250	289	39
	At	295	5%	23	280	310	30
	At or Above	311	37%	25	294	328	34
Mathematics	*Below*	135	73%	23	120	151	31
	At	164	4%	14	154	173	19
	At or Above	181	27%	18	167	192	25

NOTES: [1]IQR 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.

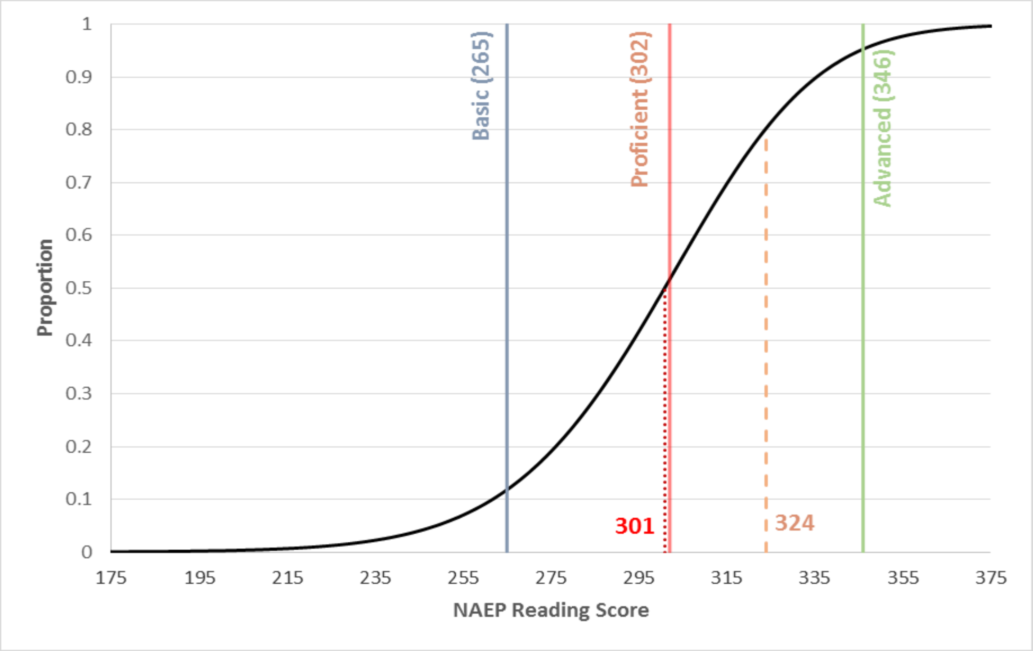


Figure 1: Proportion of students meeting the ACT reading benchmark of 22 in Tennessee for NAEP reading scores

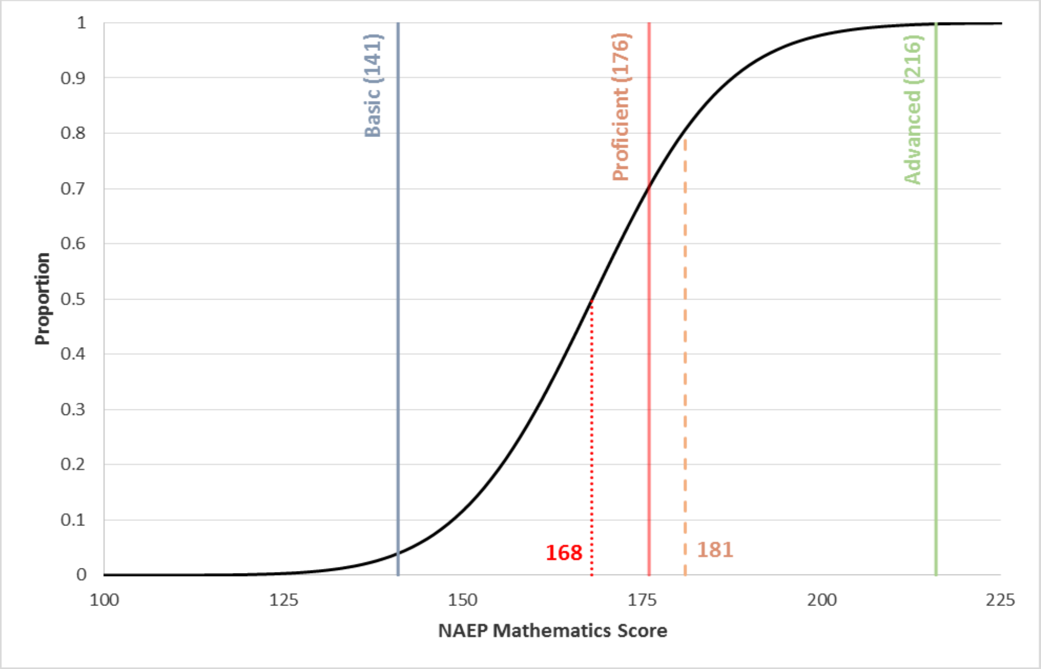


Figure 2: Proportion of students meeting the ACT mathematics benchmark of 22 in Tennessee for NAEP mathematics scores

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.

Subject	NAEP *Proficient*	Mean	Percentage	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	18	68%	5	14	20	6
	At or Above[2]	25	32%	5	21	28	7
Mathematics	*Below*	18	82%	3	15	19	4
	At or Above	26	18%	4	23	28	5

NOTES: [1] IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2] The "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.

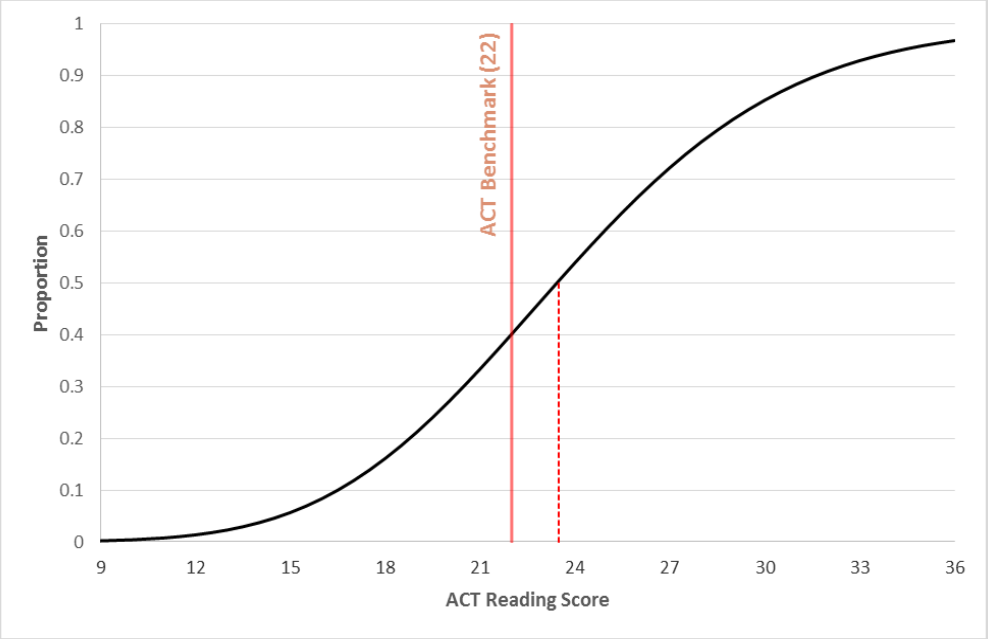


Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 in Tennessee for ACT reading scores

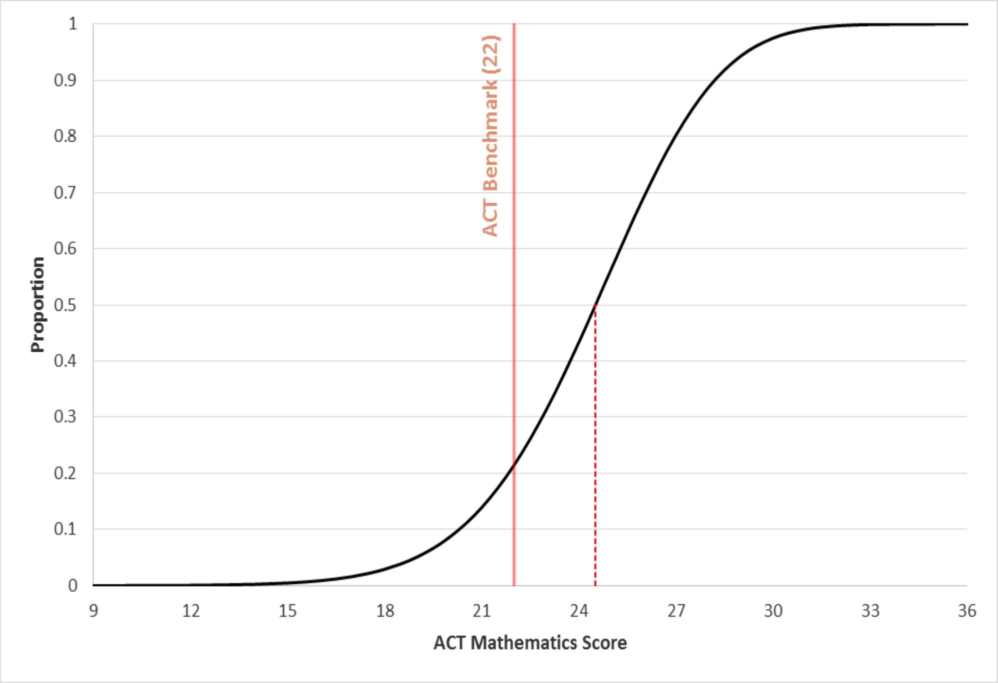


Figure 4: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 in Tennessee for ACT mathematics scores

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

ACT (2013). *What are the ACT College Readiness Benchmarks?* (http://www.act.org/content/dam/act/unsecured/documents/benchmarks.pdf).

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report.*

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?

NATIONAL ASSESSMENT GOVERNING BOARD

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

National Assessment Governing Board

BOARD MEMBERSHIP (2014-2015)

www.nagb.org | www.nagb.gov

Suggested Citation:

The National Assessment Governing Board. (2015). *The National Assessment of Educational Progress, research on academic preparedness for job training programs.* Washington, DC: National Assessment Governing Board.

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.

Type of Research Study	Status	Reports
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*).

Limitations for Other Research Designs

Additional research plans to examine statistical relationships or benchmarking of results against a reference group, such as program recruits, could not be pursued because of a lack of available data and settings that could support these plans. Few standardized assessments across employers exist that explicitly address preparedness for job training. The WorkKeys assessment was considered for this purpose, however, performance results for WorkKeys examinees are not usually sufficiently available to conduct statistical linking with other assessments. One potential data opportunity was explored in Florida, but the sample was not large enough for analysis. (See the *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 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.

References

ACT, Inc. (2010a). *The alignment of the NAEP grade 12 mathematics assessment and the WorkKeys applied mathematics assessment*. Iowa City, IA: Author.

ACT, Inc. (2010b). *The alignment of the NAEP grade 12 reading assessment and the WorkKeys reading for information assessment*. Iowa City, IA: Author.

Becker, D.E., Dickson, E.R., McCloy, R.A., Sinclair, A.L., & Thacker, A.A. (2015). *Evaluation of NAEP 12th grade reading and mathematics frameworks and item pools as measures of academic preparedness for college and job training comprehensive report* (No. 2015 004). Alexandria, VA: Human Resources Research Organization.

Dickson, E.R., Smith, E., Deatz, R., Thacker, A.A., Sinclair, A.L., & Johnston-Fisher, J. (2014). *The content alignment between the NAEP and WorkKeys assessments final report* (No. 2014 054). Alexandria, VA: Human Resources Research Organization.

Fields, R. (2014). *Towards the National Assessment of Educational Progress (NAEP) as an indicator of academic preparedness for college and job training*. Washington, DC: National Assessment Governing Board.

Kamil, M.L. (2012, April). *Reading preparedness for college and technical professions*. Paper presented in the Setting Academic Preparedness Standards for Job Training Programs: Are We Prepared? symposium at the annual meetings of the National Council on Measurement in Education, April 14, 2012, Vancouver, British Columbia, Canada.

Kilpatrick, J. (2012, April). *The standard for minimal academic preparedness in mathematics to enter a job-training program*. Paper presented in the Setting Academic Preparedness Standards for Job Training Programs: Are We Prepared? symposium at the annual meetings of the National Council on Measurement in Education, April 14, 2012, Vancouver, British Columbia, Canada.

Loomis, S. C. (2012, April). *A study of "irrelevant" items: Impact on bookmark placement and implications for college and career readiness*. Paper presented in the Setting Academic Preparedness Standards for Job Training Programs: Are We Prepared? symposium at the annual meetings of the National Council on Measurement in Education, April 14, 2012, Vancouver, British Columbia, Canada.

McCloy, R.A., & Day, T.C. (2015). *NAEP technical advisory panel proceedings of the symposium on academic preparedness research* (No. 2014 062). Alexandria, VA: Human Resources Research Organization.

Sinclair, A. L., Becker, D. E., McCloy, R. A., & Thacker, A. A. (2014). *Comparisons between NAEP and O*NET on academic preparedness for job training for five target occupations* (No. 2014 012). Alexandria, VA: Human Resources Research Organization.

U.S. Department of Education, Technical Panel on 12th Grade Preparedness Research. (2009). *Making new links: 12th grade and beyond*. Washington, DC: National Assessment Governing Board.

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
800 North Capitol Street, N.W., Suite 825
Washington, DC 20002–4233

www.nagb.org / www.nagb.gov