National Assessment Governing Board
Committee on Standards, Design and Methodology
Friday, March 1, 2019
10:20 am – 12:50 pm
AGENDA

10:20 – 10:25 am	Welcome and Review of Agenda *Andrew Ho, COSDAM Chair*	
10:25 – 11:25 am	Next Steps for Developing a Statement on Intended Uses of NAEP (SV #3) *Sharyn Rosenberg, Assistant Director for Psychometrics* *Andrew Ho*	Attachment A
11:25 am – 12:00 pm	NAEP-ACT Linking Study Report (SV #2, #10) *Helena Jia, Educational Testing Service*	Attachment B
12:00 – 12:50 pm	**CLOSED SESSION**: Design and Analysis of NAEP Writing Assessment *Enis Dogan, National Center for Education Statistics*	
	Information Item Update on Implementing the Strategic Vision (SV #2-10)	Attachment C

Developing a Statement on the Intended Uses of NAEP
For March 2019 COSDAM Discussion

Over the past couple of years, COSDAM has been discussing the need to explicitly state how NAEP results are intended to be used, that is, to develop a statement of appropriate and inappropriate uses of NAEP scale scores and achievement levels. These discussions have not been limited to COSDAM as this topic is important to the Reporting and Dissemination Committee as well, especially to focus dissemination efforts on increasing the most appropriate and impactful uses of NAEP.

The importance of this discussion is emphasized in the guidance for test developers/users found in the current *Standards for Educational and Psychological Testing*. The relevant standard and other Governing Board documents related to the uses of NAEP are summarized briefly below.

The upcoming COSDAM discussion will focus on guidance provided in the existing Governing Board policy documents and research on how NAEP is actually used by various stakeholders. The goal of the discussion is to provide clear direction on the desirable outcome of this work and the path forward.

Background:

1. Standards. The very first standard (Standard 1.0) of the *Standards for Educational and Psychological Testing* states: "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA, APA, & NCME, 2014; p. 23). https://www.apa.org/science/about/psa/2014/09/educational-psychological-testing.aspx

2. Strategic Vision. The purpose of the Governing Board's Strategic Vision is to address the question, "How can NAEP provide information about how our students are doing in the most innovative, informative, and impactful ways?" The Strategic Vision includes a goal to expand the availability, utility and use of NAEP resources, in part by creating new resources to inform education policy and practice (SV #3). COSDAM activities to address this goal include: conducting research on how NAEP results are currently used (both appropriately and inappropriately) by various stakeholders; developing a statement of the intended and unintended uses of NAEP data; and working with NCES to identify and provide documentation of validity evidence in support of the appropriate uses of NAEP.

3. Evaluation of Achievement Levels. The recent evaluation of the NAEP achievement levels includes the following recommendation (Recommendation #5): *"Research is needed to articulate the intended interpretations and uses of the achievement levels and to collect validity evidence to support these interpretations and uses. In addition, research is needed to identify the actual interpretations and uses commonly made by NAEP's various audiences and evaluate the validity of each of them. This information*

should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations" (National Academies of Sciences, Engineering, and Medicine, 2017, p. 13).

The recently revised and adopted policy on Developing Student Achievement Levels for NAEP refers to an interpretative guide[1] that would accompany NAEP reports and include specific examples of appropriate and inappropriate interpretations and uses of the results (Principle 3h). As part of their work conducting the evaluation, the National Academies of Sciences, Engineering, and Medicine (2017) also conducted some research on how the achievement levels are being used. The evaluation report includes a summary of their findings about the uses, interpretations and actions various users have identified for the NAEP achievement levels (p. 192-193). https://www.nap.edu/catalog/6296/grading-the-nations-report-card-evaluating-naep-and-transforming-the

4. The NAEP Law. The current authorizing legislation for NAEP, Public Law 107-279, prohibits NAEP from serving certain functions and is relevant to the discussion of how NAEP should be used. Specific prohibitions are listed in the next section.

5. Governing Board Policy Documents. In addition to the authorizing legislation, the Governing Board has existing policy documents which speak to the intended uses of NAEP and which should inform this discussion. The documents are linked below and key takeaways are briefly described in the next section:

    a. Redesigning the National Assessment of Educational Progress, August 2, 1996

    b. Formal response to the No Child Left Behind proposal to use NAEP, May, 18, 2001

    c. General Policy: Conducting and Reporting the National Assessment of Educational Progress, August 3, 2013

6. Research on the Use of NAEP. As part of the Technical Support contract HumRRO has been conducting research on how NAEP results have been used by various audiences, including: federal, state, and local policymakers; educators; media; education researchers; and the general public. The first phase of this work, analyzing existing artifacts produced by these various audiences, is nearing completion. Using this work, their own knowledge of how NAEP is used, and research conducted as part of the recent evaluation of NAEP achievement levels, staff developed a list of actual uses of NAEP.

[1] Per the discussion at the November 2018 Board meeting, the Reporting and Dissemination (R&D) Committee is taking the lead on developing an interpretative guide for the NAEP achievement levels, with input from COSDAM. The R&D Committee is discussing preliminary plans for this guide during this quarterly Board meeting.

Policy Assertions: What NAEP Can and Cannot Do

The NAEP legislation and existing Governing Board policy documents provide some background and guidance on the appropriate uses of NAEP. Several documents are summarized briefly below with links provided for each full document. Following the summary, staff have developed a draft of what this legislation and relevant Board policies indicate that NAEP Can and Cannot Do. This draft will be used for discussion during the COSDAM meeting and as a springboard for planning next steps in developing additional guidance about the uses of NAEP.

A. The NAEP Law
https://www.nagb.gov/about-naep/the-naep-law.html

Prohibited Activities:

1. Use of assessment items or data to evaluate individual students or teachers, or to provide rewards or sanctions for individual students, teachers, schools, or LEAs
2. Use by agent(s) of Federal Government to establish, require, or influence the standards, assessments, curriculum, including lesson plans, textbooks, or classroom materials, or instructional practices of states or LEAs
3. Use of any assessment for student promotion or graduation purposes
4. Use of assessment to affect home schools, whether or not a home school is treated as a home school or a private school under state law, nor shall any home schooled student be required to participate in any assessment

B. Redesigning the National Assessment of Educational Progress, August 2, 1996: This policy describes the historical efforts to streamline the NAEP assessment and focus data collection on its principal audiences. "Because NAEP cannot do all that some would have it do …" the policy sets forth the elements that guide current NAEP operations.
https://nagb.gov/content/nagb/assets/documents/policies/Redesigning%20the%20National%20Assessment%20of%20Educational%20Progress.pdf

1. Purpose and objectives of NAEP
2. Audience for NAEP
3. Limitations: What NAEP is not
4. Assess all subjects specified by Congress
5. Provide NAEP results for states
6. Vary the amount of detail in testing and in reporting results
7. Use performance standards to report whether student achievement is "good enough"
8. Use international comparisons
9. Emphasize reporting for grades 4, 8, and 12
10. Use innovations in measurement and reporting
11. Keep test frameworks and specifications stable
12. Use an appropriate mix of multiple-choice and "performance" questions

C. Formal response to the No Child Left Behind plan to use NAEP, May, 18, 2001: This letter to Secretary of Education Rod Paige is from NAGB Chair Mark Musick and emphasizes the following. https://nagb.gov/content/nagb/assets/documents/policies/GP-NAEP-No-Child-Left-Behind-Act.pdf

1. The primary purpose of NAEP
2. NAEP should be used only to confirm state results
3. NAEP will provide results within six months
4. A "reasonable person" standard, not strict tests of statistical significance, should be applied to confirm state results
5. Using NAEP to confirm results on state tests will not lead to a national curriculum
6. The Federal Government should reduce the testing burden on states and provide resources for this purpose

D. General Policy: Conducting and Reporting the National Assessment of Educational Progress, August 3, 2013: This policy reaffirms the essential principles and values that are embodied in NAEP including the following. https://nagb.gov/content/nagb/assets/documents/policies/GP-Conducting-and-Reporting-National-Assessment-of-Educational-Progress.pdf

1. Purpose and characteristics of NAEP
2. Limitations: What NAEP is not
3. The Audiences for NAEP
4. Goals for Conducting and Reporting the National Assessment
    - To serve as a consistent external, independent measure of student achievement by which results across education systems can be compared
    - To develop technically sound, relevant assessments designed to measure what students know and can do
    - To set and report achievement levels for NAEP results
    - To bring attention to achievement gaps where they exist among demographic groups and the urgency of closing those achievement gaps
    - To disseminate timely NAEP reports and to make NAEP data and information useful and easily accessible to various audiences, including educators, policymakers, parent leaders and the public
    - To innovate in NAEP framework development, item development, test administration, data collection, test security, scoring, analysis and reporting

E. Actual Uses of NAEP By Various Audiences: The following initial list of how NAEP scores and achievement levels have been used was compiled by Governing Board staff based on their own experience, research being performed by HumRRO under the Technical Support contract, and research conducted by the National Academies of Sciences, Engineering, and Medicine as part of its evaluation of NAEP achievement

levels. This list contains a combination of results, claims or interpretations, and actions. **The elements of this list are numbered to facilitate the discussion at the COSDAM meeting, but the numbering does not represent a hierarchy of uses.**

How NAEP is Used: Results, Interpretation, and Actions

1. Compare NAEP scale scores and/or achievement levels at a single point in time across states, districts (TUDA), and/or student groups
2. Compare NAEP scale scores and/or achievement levels over time (trends) for the nation, states, districts (TUDA), and/or student groups
3. Rank order states or districts in terms of NAEP scale scores and/or achievement levels overall and/or for a specific student group
4. Analyze performance gaps in NAEP scale scores and/or achievement levels between two student groups at a single point in time
5. Analyze changes in performance gaps in NAEP scale scores and/or achievement levels between two student groups over time (gap trends)
6. Validate performance or changes in performance on state tests
7. Analyze the relationship between contextual variables and NAEP scale scores and/or achievement levels
8. Describe the context in which students learn from information gathered by student, teacher, and school questionnaires
9. Compare NAEP scale scores and/or achievement levels across subject areas
10. Compare NAEP scale scores and/or achievement levels across grades
11. Compare NAEP scale scores and/or achievement levels before and after a program or policy is implemented
12. Estimate the percentage of students who are academically prepared for college by the end of high school
13. Show examples of what students know and can do through sample items and item maps
14. Establish a common scale for linking state tests and comparing results across all school districts (e.g., Stanford Education Data Archive)
15. Link other assessments to NAEP to provide state-level results on other assessments that were not administered at the state level (e.g., TIMSS)
16. Establish a common scale for comparing the rigor of state standards to each other and to NAEP Proficient
17. Compare the percentage of students at or above each achievement level on NAEP and on other assessments, including state and international assessments
18. Serve as a benchmark of performance at NAEP Proficient to inform standard settings on other assessments
19. To evaluate whether current programs and policies are effective
20. To support the need for new programs and policies
21. To influence decisions about funding for educational policies and programs
22. To influence legislation

23. To determine whether the nation, states, and/or TUDAs are making progress for students overall and/or selected student groups
24. To evaluate the quality of education at a single point in time and/or over time
25. To claim that some states and/or districts are doing a better job educating students based on their rankings on NAEP
26. To identify where there are large performance gaps and/or interventions are needed
27. To identify states and/or TUDAs who are doing something extraordinary so that best practices can be shared
28. To criticize states for lying about the percentage of students at or above Proficient if it varies substantially from NAEP
29. To generate and test hypotheses about factors related to student achievement (education research)
30. To claim that students should do more of X because X is correlated with higher performance
31. To determine whether U.S. students will be internationally competitive
32. To call for higher standards
33. To call for more accountability systems
34. To claim that the majority of students lack basic skills (or are faring well)
35. To make claims about the percentage of students who are performing “on grade level”
36. To inform the development of state content standards

Using the resources described above, Governing Board staff drafted Table 1 below to generate discussion at the upcoming COSDAM meeting. The elements of the table are numbered to facilitate the discussion at the COSDAM meeting. The table highlights prominent examples but is not intended to be comprehensive.

Table 1. What NAEP Can and Cannot Do (Draft as of 1/30/2019)		
	NAEP Can DO	**NAEP Cannot DO**
1	Adopt an independent assessment framework developed through a national consensus approach, but which is free of specific ideologies	Develop an assessment framework that duplicates local or published curriculum frameworks
2	Administer NAEP independently of specific educational initiatives	Modify NAEP to address specific educational initiatives
3	Make judgements about what students should know and be able to do in relation to the NAEP frameworks	Make judgements about what students should know and be able do in relation to state and/or local curricula or in relation to specific ideologies or teaching techniques
4	Make judgements about how well students are performing in relation to the NAEP Frameworks and ALDs	Make judgments about how well students are performing in relation to state and local curricula and state ALDs
5	Assess statistically representative samples of students (large group results only)	Assess all students
6	Assess only a few grade levels in several designated subject areas	Assess many subjects at many grade levels
7	Report on the status of and trends in US student achievement using a common yardstick for: • The nation • States • Select large urban districts (TUDAs) on voluntary basis	Report on how individual students, schools, or districts are achieving Make claims about why some states or TUDAs are doing better or worse than others
8	Provide external evidence of student achievement in states and large districts	Provide data on local policy changes and their success or lack thereof
9	Provide information for a few primary audiences: • General public • Federal and State-level policymakers and educators • Selected large urban districts	Answer all questions of concern to the education community
10	Identify achievement gaps between student groups	Make judgments about what factors caused the achievement gaps
11	Provide insight into which contexts are related to (statistically significant) student achievement	Make claims about which contextual variables caused higher or lower achievement
12	Make comparisons of NAEP to other assessments which assess similar content and a similar population of students	Make comparisons of NAEP to all other assessments regardless of the content and students assessed

Maximizing the Impact and Usefulness of NAEP

There is a tension between the limitations of what NAEP is allowed and was designed to do, and the Governing Board's goal of maximizing the impact of NAEP data (as articulated in the Strategic Vision). This tension was surfaced during the August 2018 COSDAM meeting, and minutes from that meeting are excerpted below:

> *COSDAM members agreed that the Governing Board's role should be to assert intended uses but not try to be the arbiter of all possible uses. This is not an exercise in compliance, but rather an effort to provide a model of appropriate uses. NAEP is in a unique position because most tests are developed by organizations that are primary users of the data, but NAEP is used primarily by people and organizations who are not the developers of the assessment. The Governing Board and NCES cannot be responsible for all possible uses of NAEP, rather the goal of this activity is to make recommendations about general uses.*
>
> *There was disagreement about the extent to which the statement of intended uses of NAEP should strive to be either bold or comprehensive. Some committee members argued that the statement should focus on a small number of consensus uses and interpretations that are clearly supported by the Governing Board and the NAEP legislation, in addition to articulating misinterpretations and misuses that are obviously unsubstantiated or logically and technically flawed.*
>
> *Other committee members expressed a desire to be bold, stating that the ultimate purpose of NAEP should be to improve education in America. It is important to make sure that NAEP results are useful to states and TUDAs. If a state or TUDA is not happy with their NAEP results, how can the data help them figure out how to improve? Since education takes place at the state and local level, what are the most important ways that governors and educators can use NAEP to affect higher outcomes? That is, the most important question should be not how NAEP can be used, but rather how NAEP can be useful.*
>
> *The discussion also noted that the authorizing legislation for NAEP was written in a very different time and context, when not all states administered assessments and those that did often used norm-referenced tests. At that time, NAEP provided for the first time a picture of how the nation was doing, albeit with clear restrictions on individual results and influencing curriculum. At the current time, there is interest in using NAEP at the state and local levels as an agent of change and improvement. This takes place through legislators and educators as the agents of change.*
>
> *Mr. Ho closed the discussion by challenging the committee to consider how the Board can strategically partner with actors who can use NAEP as an agent of change in ways that Board and NCES cannot accomplish directly. He noted that his suggestion to stick with just a few straightforward and easily supportable*

notions for a statement of intended uses and interpretations is meant to describe the Board's grasp, not to circumscribe its reach.

Following the August 2018 COSDAM meeting, one suggestion was to consider developing a statement of intended uses which first delineated a small number of intended uses based on the NAEP law and other policy documents, along with an additional section to discuss how NAEP data can be used by other stakeholders to accomplish broader aims. To explore this idea, Governing Board staff drafted Table 2 to generate discussion at the upcoming COSDAM meeting.

Table 2. What NAEP Can and Cannot Do Directly and Indirectly (Draft as of 1/30/2019)

	NAEP Can DO	**NAEP Cannot DO Directly**	**How Others May Use NAEP Data in Impactful Ways**
1	Adopt an independent assessment framework developed through a national consensus approach, but which is free of specific ideologies	Develop an assessment framework that duplicates local or published curriculum frameworks	Use NAEP frameworks and items as a resource when developing state standards and/or curricula Use NAEP frameworks to stimulate discussions about what students should be expected to know and be able to do
2	Administer NAEP independently of specific educational initiatives	Modify NAEP to address specific educational initiatives	
3	Make judgements about what students should know and be able to do in relation to the NAEP frameworks	Make judgements about what students should know and be able do in relation to state and/or local curricula or in relation to specific ideologies or teaching techniques	Use NAEP results in conjunction with state assessment results to understand why there may be differences in how students are performing on the different assessments
4	Make judgements about how well students are performing in relation to the NAEP Frameworks and ALDs	Make judgments about how well students are performing in relation to state and local curricula and state ALDs	Use NAEP as external reference in standard settings on other tests to provide additional context
5	Assess statistically representative samples of students (large group results only)	Assess all students	
6	Assess only a few grade levels in several designated subject areas	Assess many subjects at many grade levels	

Table 2. What NAEP Can and Cannot Do Directly and Indirectly (Draft as of 1/30/2019)			
	NAEP Can DO	**NAEP Cannot DO <u>Directly</u>**	**How <u>Others</u> May Use NAEP Data in Impactful Ways**
7	Report on the status of and trends in US student achievement using a common yardstick for: • The nation • States • Select large urban districts (TUDAs) on voluntary basis	Report on how individual students, schools, or districts are achieving Make claims about why some states or TUDAs are doing better or worse than others	Use NAEP results to pose hypotheses about why certain states/TUDAs may have higher scores, to stimulate further research Compare performance with similar student groups to understand how similar students are doing and pose hypotheses about performance
8	Provide external evidence of student achievement in states and large districts	Provide data on local policy changes and their success or lack thereof	Use NAEP results to understand whether there is a coherent narrative of performance, and if not, to pose hypotheses about what might account for differences
9	Provide information for a few primary audiences: • General public • Federal and State-level policymakers and educators • Selected large urban districts	Answer all questions of concern to the education community	Use NAEP results to call attention to issues that may warrant a need for further understanding or action
10	Identify achievement gaps between student groups	Make judgments about what factors caused the achievement gaps	Use data to uncover opportunity and achievement gaps that merit attention, efforts, and resources to narrow Use data to pose hypotheses about the reasons for gaps and also what might account for smaller gaps among some states/TUDAs
11	Provide insight into which contexts are related to (statistically significant) student achievement	Make claims about which contextual variables caused higher or lower achievement	Use data and contextual variables to speak with other states/TUDAs to understand their theories for what practices may be contributing to higher achievement

Table 2. What NAEP Can and Cannot Do Directly and Indirectly (Draft as of 1/30/2019)			
	NAEP Can DO	**NAEP Cannot DO Directly**	**How Others May Use NAEP Data in Impactful Ways**
12	Make comparisons of NAEP to other assessments which assess similar content and a similar population of students	Make comparisons of NAEP to all other assessments regardless of the content and students assessed	

Proposed Next Steps

Following the March 2019 COSDAM discussion, Governing Board staff propose to develop a draft document of intended uses of NAEP to support discussions during the May 2019 quarterly Board meeting. This document is intended to be brief (approximately 2-3 pages) and may include the following sections:

- Introduction/Background
- Primary intended uses of NAEP clearly supported by NAEP law and relevant policies (based on Table 1)
- A few prominent examples of common uses of NAEP that are clearly inappropriate and should be avoided (based on research of how NAEP is actually used)
- How NAEP data can be useful and impactful via educators and policymakers (based on Table 2)
- Conclusion

Discussion Questions

1. What is your reaction to the proposed next steps?

2. What is your reaction to the information presented in Tables 1 and 2?

3. Is additional research needed at this time to inform the statement of intended uses of NAEP?

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress.* Washington, DC: The National Academies Press.

Patelis, Thanos (2019). *How Different Types of Audiences Use NAEP*. Human Resources Research Organization (HumRRO) under contract to the National Assessment Governing Board (ED-NAGB-17-C-0002).

NAEP-ACT Linking Studies in Reading and Mathematics (SV #2, SV #10)

Since 2003, the Governing Board has been interested in using NAEP 12th grade assessments in reading and mathematics to support statements about students' academic preparedness for college. During the first phase of this work several research studies were conducted between 2008 and 2013. This research included statistical linking studies using data from students who were sampled and assessed by NAEP 12th grade reading or mathematics in 2009 and had also taken the SAT by June 2009. Parallel linking studies were not able to be conducted in 2009 with NAEP and the ACT assessments.

In conjunction with the administration of the 2013 NAEP 12th grade assessments in reading and mathematics, Governing Board staff approached ACT senior leadership about the possibility of conducting such linking studies. Data sharing agreements were established between Governing Board staff, NCES staff, NCES contractors Westat and ETS, and ACT. The process of finalizing all necessary agreements and conducting the work took several years, and a draft report is now available for COSDAM review and discussion.

In addition to bearing on the Board's interest in the knowledge and skills relevant for postsecondary preparedness (SV #10), this work is an example of how NAEP can be connected to other student assessments (SV #2). Understanding how NAEP performance relates to the ACT can help provide additional meaning to the NAEP 12th grade assessment results in reading and mathematics in the context of other indicators that are familiar to various stakeholders.

During the March 2019 COSDAM meeting, Helena Jia of Educational Testing Service will provide a brief overview of the research and will answer questions from COSDAM members.

January 29, 2019

Draft for COSDAM Discussion

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

Nuo Xi
Mei-Jang Lin
Laura Jerry
David Freund
Helena Jia

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

Prepared by Educational Testing Service for the National Center for Education Statistics under contract ED-IES-12-R-0021, Task 9.3.1 at the request of the National Assessment Governing Board.

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. The commission recommended that a state program similar to 4th and 8th grade should be implemented 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*.

According to the Governing Board[1], to be "academically prepared for college", 12th graders should have the knowledge and skills in reading and mathematics to qualify for placement into entry-level college credit courses that meet general education requirements, without the need for remedial coursework in reading or mathematics. Statistical linking studies were conducted to examine performance on NAEP in relation to the college readiness benchmark adopted by the College Board for the SAT critical reading and mathematics tests (Moran, Oranje, & Freund, 2012). The statistical linking studies used data from students who were sampled and assessed in the NAEP 12th grade reading or mathematics in 2009 and had also taken the SAT by June 2009.

After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research (National Assessment Governing Board, 2009), potential benchmarks were identified on the 12th grade NAEP reading and mathematics scales to indicate what level of performance would correspond to a reasonable probability of being academically prepared for placement into first-year, general education college coursework without the need for remediation. Beginning with the 2013 12th grade NAEP assessments, the reporting of results included the percentage of students academically prepared for college, as indicated by a score of 302 on the grade 12 NAEP reading assessment (same as the *NAEP Proficient* cut score) and a score of 163 on the grade 12 NAEP mathematics assessment (between the cut scores for *NAEP Basic* and *NAEP Proficient*). The Governing Board is currently working to expand beyond academic preparedness for college, to report on preparedness for postsecondary endeavors more broadly using a dashboard of indicators.

As part of the second phase of the Governing Board's academic preparedness research, Michigan and Tennessee participated in the state-level statistical linking research[2] connecting NAEP and ACT and provided data on students who were part of the NAEP grade 12 sample during the 2012–2013 school year including their ACT data. The state-level NAEP-ACT linking study results can be found on the Governing Board's website. Although NAEP-ACT linking studies were performed using special samples of students in these two states (MI and TN), it is important to understand how national results on the two assessments are related to each other. The objective of the current study is to enable interpretation of the NAEP results in reference to the ACT score scale and the ACT college readiness benchmarks. This study also compares the newly identified potential benchmarks

[1] Governing Board preparedness research website: https://www.nagb.gov/focus-areas/reports/preparedness-research.html

[2] Massachusetts also participated in the state-level statistical linking research. However, the linking study conducted in MA was to connect NAEP and SAT, at the state level.

on NAEP to those established in the NAEP-SAT linking study, to evaluate the consistency in linking through different assessments.

In this report, we will describe the NAEP and ACT tests in reading and mathematics that were used for this study, discuss the linking methodology (and refer the interested readers to more technical references), and provide the results. A summary concludes this report.

Linking Assessments

The ACT Test

The ACT test is a U.S. college admission test measuring what students learn in high school to determine their academic readiness for college (https://www.act.org/content/act/en/products-and-services/the-act.html). The assessment contains four multiple-choice tests and an optional writing test that does not affect students' composite scores. Each multiple-choice test measures student achievement in one of the following four areas: English, mathematics, reading, and science. The testing time and number of items vary by subject. For ACT reading, students have 35 minutes to finish 40 multiple-choice questions. For ACT mathematics, students have 60 minutes to finish 60 multiple-choice questions. The ACT scores provide evidence about the knowledge and skills that students are likely to have in each of the four aforementioned areas. The distribution of item difficulties was selected so that the tests will effectively differentiate among students who vary widely in the level of achievement. 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, the ACT reading score was used to link with the NAEP reading assessment and the ACT mathematics score was used to link with the NAEP math assessment. The data set included the sample that participated in the 2013 NAEP grade 12 reading and mathematics assessments and had also taken the ACT by June 2013.

The ACT test is designed to assess a student's academic achievement in high school, especially with respect to college readiness. To help students translate test scores into clear indicators of their current levels of college readiness, ACT provides college readiness benchmarks reflecting the minimum ACT scores required for students to have a reasonable chance of success in credit-bearing college courses. Using data from ACT's Research Services, including the Course Placement Service and Prediction Service, ACT derived their college readiness benchmarks to reflect the ACT score associated with a 50% chance of earning a B or higher (or approximately a 75% chance of earning a C or higher) in a related first-year college course at a typical postsecondary institution (including two-year and four-year colleges; ACT, 2013). The ACT reading benchmark is a 22 on the ACT reading test and is derived based on course success in social science courses (ACT Technical Manual, Table 8.4). The ACT mathematics benchmark is a 22 on the ACT mathematics test and is derived based on course success in college algebra (ACT Technical Manual, Table 8.4). These two benchmarks were used in this study.

The National Assessment of Educational Progress (NAEP)

The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board and measures what U.S. students know and can do in various subjects across the nation, states, and in some urban districts. 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 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. Selected students had 50 minutes[3] 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 mathematics, each booklet contains two blocks of about 15 items each. A mix of multiple-choice (MC) and constructed-response (CR) items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design).

NAEP reports assessment results reflecting student performance in two ways: average scores on the NAEP subject scale and percentages of students attaining NAEP achievement levels. The three NAEP achievement levels—*NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*—describe what students should know and be able to do. In particular, the *NAEP Proficient* achievement level is a benchmark for solid academic performance (i.e., students reaching this level have demonstrated competency over challenging subject matter; https://nagb.gov/focus-areas/NAEP-achievement-levels.html). For grade 12 reading, the NAEP performance is reported on a 0–500 scale, and the *NAEP Proficient* achievement level cut score is 302. For grade 12 mathematics, the NAEP results are reported on a 0–300 scale, and the *NAEP Proficient* achievement level cut score is 176.

By law, no student or school results are estimated or reported using the NAEP 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. For linking studies involving NAEP, this requires that the relationship between NAEP and other measures (e.g., the ACT) be directly estimated using this latent regression methodology, since there are no appropriate student-level NAEP scores available. In the methodology section we will discuss some of the steps required to complete this part of the research. For a comprehensive description of NAEP estimation procedures, interested readers are referred to Mislevy, Beaton, Kaplan, and Sheehan (1992).

Linking

When linking scales of different assessments, construct similarity between the two assessments plays an important role in determining the degree of linkage that can be achieved (Dorans, 2004). 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 the NAEP and ACT reading and mathematics assessments were

[3] The NAEP reading and mathematics assessments transitioned to a digital delivery platform in 2017. Under the digitally-based assessment (DBA) design, each selected student took two blocks of cognitive items, each with 30 minutes of allotted testing time, making the total testing time 60 minutes.

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 the 2009 NAEP and ACT, and the content overlap was more extensive in mathematics than in reading (https://www.nagb.gov/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf).

In addition, statistical indices such as correlations between the two sets of scale scores to be linked are also needed when determining the most appropriate type of linking method. In the current study, two types of test score links are considered: concordance and projection. *Concordance* occurs when scores are linked across tests built to different specifications, measure similar constructs, and are administered to similar populations (Holland & Dorans, 2006). A generally accepted minimum correlation for concordance is $r = 0.866$ (Dorans, 1999; Dorans & Walker, 2007), which corresponds to a 50 percent reduction in uncertainty (Dorans, 1997). *Projection* addresses assessments constructed around different conceptions of students' competence, or around the same conceptions but with tasks that differ in format or content (Mislevy, 1992). Neither concordance nor projection is a perfect prediction, but concordance in general assumes and requires a much stronger relationship than projection. Additionally, concordant scores have matched distributions (Dorans, 2004) and are provided in the form of a concordance table, while for projection there is usually no simple "one-to-one" correspondence tables generated. In the case of the NAEP-ACT linking study, a moderately strong relationship was expected based on experience with past studies, and linking the two assessments with projection appeared to be more appropriate. We elaborate further on this in subsequent sections.

Methodology

In this section we discuss the data and the linking methodology. The purpose is to give readers some insight into the procedures followed and therefore the opportunity to evaluate the results within the context of this NAEP-ACT linking study.

Data

The NAEP-ACT linking study used data from students attending public schools who were sampled and assessed in the NAEP 12th grade reading or mathematics in 2013 and had also taken the ACT test by June 2013.

NAEP Samples

From late January through early March of 2013, the NAEP assessments in reading and mathematics were administered to samples of 12th grade students that were representative of the nation. Whereas grade 12 NAEP assessments in reading and mathematics are in general administered to only nationally representative samples, 13 states (Arkansas, Connecticut, Florida, Idaho, Illinois, Iowa, Massachusetts, Michigan, New Hampshire, New Jersey, South Dakota, Tennessee, and West Virginia) volunteered to participate in a twelfth-grade state pilot program in 2013. As a result, larger samples of students from public schools in each of the 13 pilot states (roughly 2,600 students per subject) were drawn and augmented the nationally-representative samples of public school students. Overall, approximately 44,300 twelfth-grade public school students were assessed in NAEP reading and 44,900 twelfth-grade public school students were assessed in NAEP

mathematics in 2013. Sampling weights were used to ensure appropriate representation of the larger state samples in national-level analyses.

ACT Scores

For each student in the ACT data set, scores were available from one or more ACT administrations, which included separate scores for English, mathematics, reading, and science. The scale scores for each section, as well as the composite score, range from 1 to 36 in 1-point increments. The reading and mathematics scores from each student's highest ACT composite score were used in this study because these were the ACT scores most likely considered in college admissions. English and science scores were not used in this study.

Matching NAEP and ACT Test Takers

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. A data sharing agreement was established between all parties including ACT. This agreement involved the NAEP contractors working with ACT to match the needed ACT scores for students in the NAEP samples. A process for matching the student records was developed to protect students' identity. Confidentiality of ACT scores was assured through the assignment of a pseudo ID for students taking the ACT and using that pseudo ID as a way to transfer ACT scores from ACT to ETS. Similarly, the pseudo ID was appended to NAEP files by Westat, which then provided that file to ETS. Via the pseudo ID, ETS matched ACT scores to NAEP files without requiring access to any personally identifiable information (PII) data from ACT. The final student data were limited to questionnaire responses, ACT scores, and the pseudo ID. The NAEP scores were matched at a rate of 41% for the weighted reading sample and 42% for the weighted mathematics sample, resulting in 19,900 students for reading and 20,300 students for mathematics. These match rates are lower than the national ACT participation rate of approximately 54% of high school graduates in 2013. Table 1 provides weighted[4] percentages by gender and race/ethnicity for the matched sample and overall match rates.

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

	Reading							
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	30%	7%	5%	2%	#[1]	#	1%	**46%**
Female	33%	9%	7%	2%	#	#	1%	**54%**
***Total*[2]**	**63%**	**16%**	**13%**	**5%**	**1%**	**#**	**2%**	**100%**
							Overall Match Rate	**41%**

[4] For all NAEP assessments, sample weights are applied at the student level to ensure the representativeness of the jurisdictions from which they are selected. For more information on NAEP assessment weighting procedures, refer to https://nces.ed.gov/nationsreportcard/tdw/weighting/.

Mathematics								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total[2]***
Male	29%	7%	5%	2%	1%	#	1%	**45%**
Female	34%	9%	7%	3%	1%	#	1%	**55%**
Total[2]	**63%**	**16%**	**13%**	**5%**	**1%**	**#**	**2%**	**100%**
							Overall Match Rate	**42%**

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

Identification and Removal of Outliers

Given the fact that the two assessments to be linked have different purposes, reporting goals, and stakes, an outlier analysis is in order. For instance, if there are participants who scored very high on a higher stakes test (i.e., the ACT test) and very low on a lower stakes test (i.e., NAEP), the low performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered 'outliers' and removed from further analysis. An initial examination of the joint distribution of NAEP and ACT revealed very few potential outlier cases. After this cursory inspection, standardized residuals from robust regression (Huber, 1973) were used to identify approximately 0.8% of cases in reading and approximately 0.7% of cases in mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). These outliers were excluded from the final linking sample 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 five specific analysis questions:

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 college readiness benchmarks?

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

4) What scores on the ACT reading and mathematics scales correspond to the grade 12 *NAEP Proficient* cut scores in reading and mathematics?

5) What are the average ACT reading and mathematics scores and IQRs for students below, at, and at or above the *NAEP Proficient* cut scores?

We 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 the linking functions.

Estimating Correlations between NAEP and ACT

The main objective of NAEP is to report on the achievement of policy-relevant population subgroups, estimated directly using the latent regression estimation method. 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. To unbiasedly estimate a subgroup's proficiency on NAEP, the subgroup membership needs to be specified in the latent regression model. This means that to correctly estimate the correlation between NAEP and ACT scores, a separate latent regression model was defined to include the relationship between NAEP and ACT in the estimation process. Again, readers interested in learning details of the NAEP latent regression estimation process are referred to Mislevy et al. (1992).

In this study, the ACT scores were included as linear main and interaction effects in the latent regression model. The model included three interaction terms: ACT × gender, ACT × race/ethnicity, and ACT × gender × race/ethnicity. The estimation results indicated that the true score correlation between NAEP and ACT was 0.75 for reading and 0.87 for mathematics. While the correlation for mathematics met the minimum requirement for concordance (Dorans, 2004), the correlation for reading was only moderately strong, suggesting that there was enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points was 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, when both scales are continuous. 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 result in matched distributions between the two assessments or a one-to-one correspondence. Projection analyses support claims like "A score of X on NAEP corresponds to a 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. Consequently, a relatively high NAEP score level is identified, and likely, the percentage of students who actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable.

The relationship between NAEP and ACT mathematics (r=0.87) just met the minimum correlation requirement of 0.866 for concordance (Dorans, 1999; Dorans & Walker, 2007), but the relationship between NAEP and ACT reading (r =0.75) was not sufficiently strong to support concordance. Additionally, a technically sound concordance linking requests the two assessments to have similar testing population (Holland and Dorans, 2006), which was likely not met as NAEP is a low-stakes survey assessment and ACT is a high-stakes college admission test. Therefore, projection was selected for both mathematics and reading 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. For the current study, a bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-ACT distributions[5]. Based on the smoothed joint distributions between NAEP and ACT, projection tables containing conditional cumulative distributions of NAEP proficiencies for ACT scores were created. The range of possible NAEP scores below, at, and at or above the ACT college readiness 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 for a selected p. We discuss the results of the linking study in the following section.

Results

ACT college readiness benchmarks projected on the NAEP scale

The second and the third analysis questions ask what scores on the NAEP reading and mathematics scales correspond to the ACT college readiness 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?

Table 2 provides descriptive statistics to provide an initial sense of where the ACT benchmark is likely located on the NAEP scales as well as some distributional properties as context for these results. The average scores and percentile estimates for students below, at, and at or above the ACT benchmarks are spread out. 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 proficiency estimates from the latent regression model including the relationship between NAEP and ACT.

[5] As part of the loglinear smoothing procedure for reading, we preserved the first 3 moments for the NAEP distribution, 5 moments for the ACT distribution, and 4 cross-moments. For mathematics, we preserved the first 4 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 were also taken into consideration.

Table 2: Descriptive NAEP statistics for students below, at, and at or above the ACT college readiness benchmarks

Subject	ACT Benchmark	Percentage	Mean	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	53%	276	28	257	295	38
	At	6%	299	24	283	315	32
	At or Above	47%	319	28	300	338	38
Mathematics	*Below*	55%	142	22	127	157	30
	At	5%	168	15	158	178	20
	At or Above	45%	188	21	173	201	28

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

A graphical representation of the relationship between NAEP scores and the proportion of students meeting the ACT benchmark helps illustrate the process of identifying the NAEP scale scores that most reasonably corresponds to the ACT benchmarks. 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 college readiness benchmark for score levels on NAEP. Colored vertical lines indicate where the NAEP performance standards 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 the NAEP score above which students are more likely to have reached the benchmark than not (i.e., the conditional proportion *p* is set at 0.50). 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 college readiness benchmark.

From Figure 1, it can be deduced that 301 is the location on the NAEP reading scale where students have a 0.50 probability of meeting the ACT reading benchmark. Note that 301 is only 1 point below the *NAEP Proficient* achievement level cut score for NAEP reading at grade 12. Figure 2 shows that the corresponding location on the NAEP mathematics scale is 167, about 9 points below the *NAEP Proficient* achievement level cut score for NAEP mathematics at grade 12.

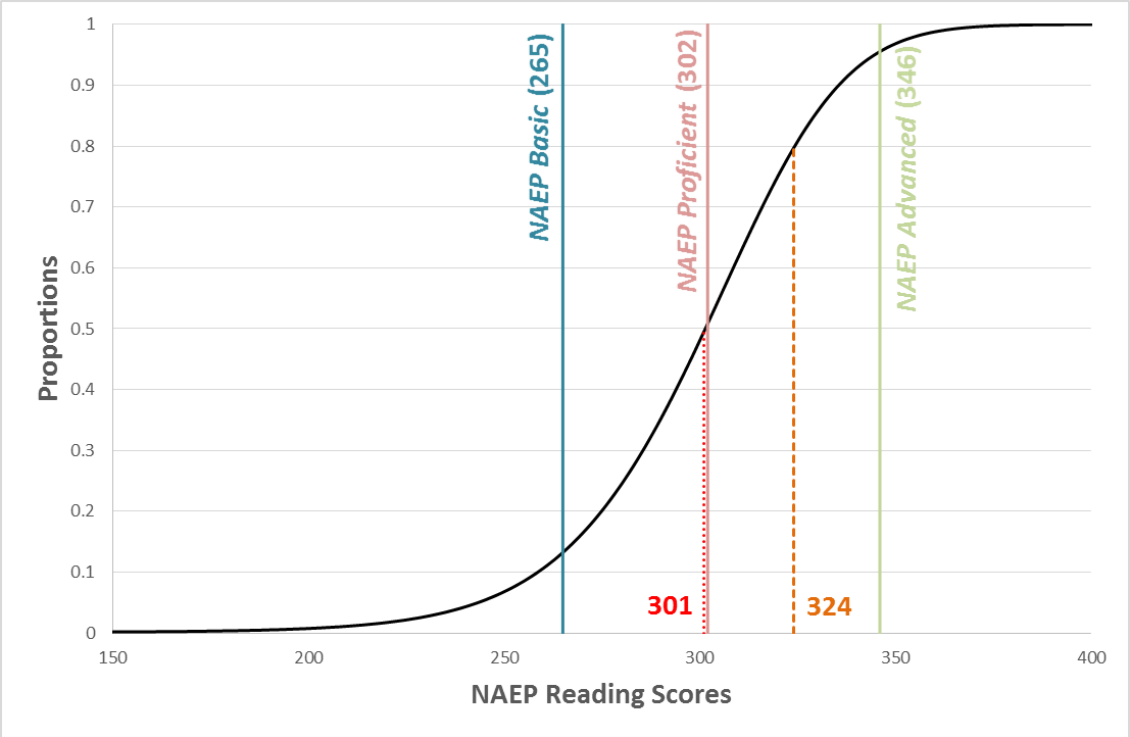

Figure 1: Proportion of students meeting the ACT reading college readiness benchmark of 22 for NAEP reading scores

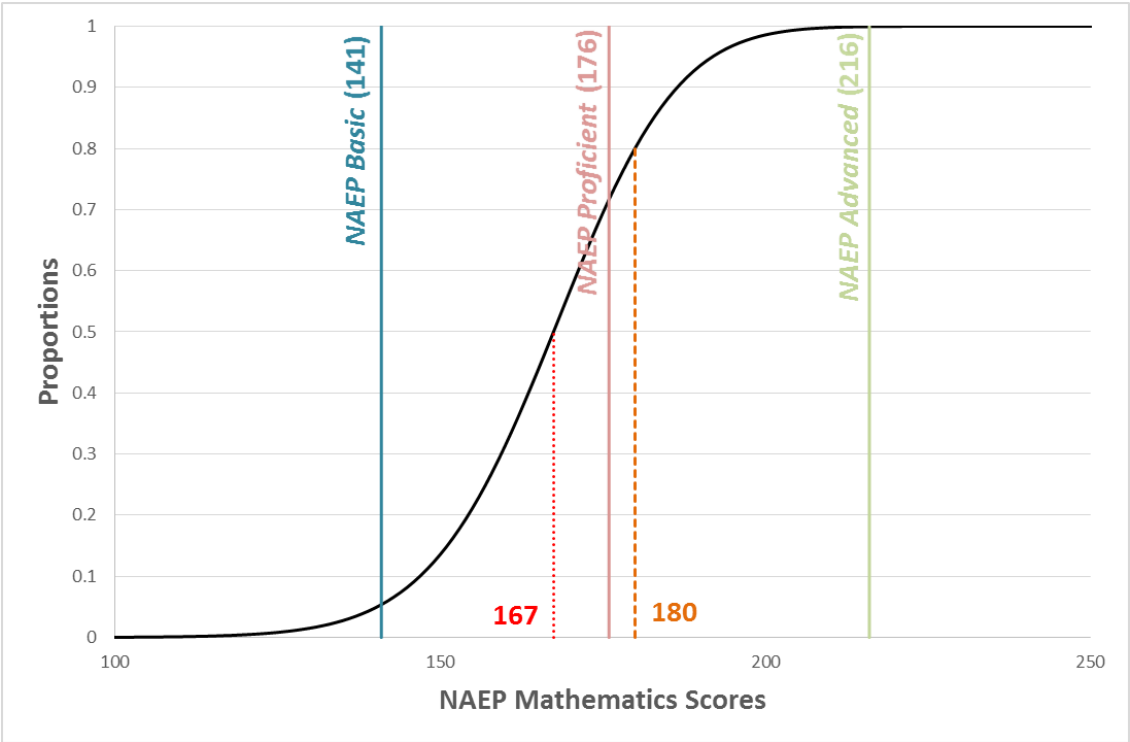

Figure 2: Proportion of students meeting the ACT mathematics college readiness benchmark of 22 for NAEP mathematics scores

Sensitivity Analysis

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[6] in order to produce reliable comparisons. For the final linking samples, both gender groups met that criterion. For the race/ethnicity groups, White, Black, Hispanic, and Asian student subgroups met the criterion. Separate linking functions were established for these subgroups.

The comparison results showed some variation across the six identified subgroups for reading but much less so for mathematics. For reading, the linking functions for Male, Female, White, and Asian student subgroups were close to the overall linking function, and the linking functions for Black and Hispanic students were slightly lower than the overall linking function. For mathematics, the linking functions for Male, Female, and White student subgroups were very close to the overall linking function, the linking function for Asian student subgroup was slightly higher than the overall linking function, and the linking functions for Black and Hispanic student subgroups were slightly lower. Thus, the location on the NAEP scale that corresponds to a 0.50 probability of meeting the ACT college readiness benchmark potentially varies by subgroup. However, it should be emphasized that some subgroups considered here had much smaller sample sizes than the overall linking sample, and therefore the difference observed between the linking functions might be due to a greater sampling variance and should be interpreted with great caution.

Impact

Given that potential NAEP benchmarks have been identified, it is important to show what percentage of students at the national level are deemed to have a reasonable probability (i.e., the probability set at 0.50) of meeting the ACT college readiness benchmarks in grade 12 across various student subgroups. Table 3 provides those percentages, based on the potential benchmarks identified on the NAEP scales, as well as the ACT college readiness benchmarks. Table 3 indicates that between 45 and 46 percent of students met the newly identified potential benchmarks on NAEP, but the results differ across subgroups. The percentages of students meeting the ACT college readiness benchmarks are similar to those meeting the NAEP potential benchmarks. No significance testing has been conducted to compare these percentages and, therefore, no comparative statements will be made.

[6] The minimum was set at 500 as a rule of thumb, 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 3: Percentage of the final linking samples that have a reasonable probability of meeting the ACT college readiness benchmarks based on the potential NAEP benchmarks, compared to the actual percentage of the same sample meeting the ACT reading benchmark of 22 and ACT mathematics benchmark of 22.

	Reading		Mathematics	
Student Group	**NAEP ≥ 301**	**ACT ≥ 22**	**NAEP ≥ 167**	**ACT ≥ 22**
Total	46%	47%	45%	47%
Male	42%	46%	49%	51%
Female	49%	48%	41%	44%
White	56%	58%	54%	56%
Black	13%	19%	12%	17%
Hispanic	34%	29%	29%	32%
Asian	54%	55%	70%	74%

NAEP *Proficient* cut scores projected on the ACT scale

To conduct the complementary analyses, we identified the point on the ACT scale that corresponds most closely to the NAEP *Proficient* cut score, essentially reversing the direction of the linking relative to the previous analyses. Table 4 provides descriptive statistics on the ACT reading and mathematics scores for students below, at, and at or above the grade 12 *NAEP Proficient* cut score. The grade 12 *NAEP Proficient* achievement level cut score was set at 302 for reading and 176 for mathematics.

Table 4: 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	55%	5	15	21	6
	At	22	1%	4	18	24	6
	At or Above	26	45%	5	22	29	7
Mathematics	*Below*	18	65%	3	15	20	5
	At	23	1%	3	21	25	4
	At or Above	27	35%	4	24	28	4

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

Following the same methodology of statistical projection (see Figures 3 and 4) we identified an ACT reading score of 22.41, rounding to 22, and a mathematics score of 23.38, rounding to 23, as cut points. Students attaining these newly identified cut points on ACT reading and mathematics have a reasonable chance (p=0.50) to meet the corresponding *NAEP Proficient* achievement levels.

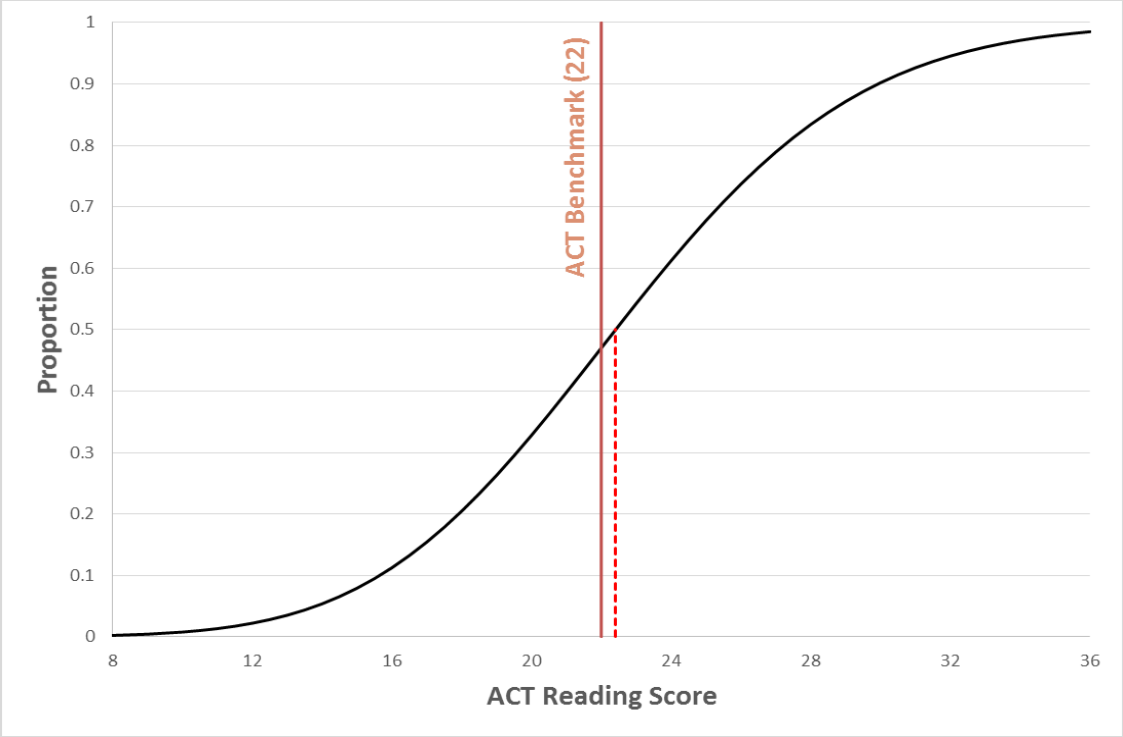

Figure 3: Proportion of students meeting the NAEP reading Proficient achievement level of 302 for ACT reading scores

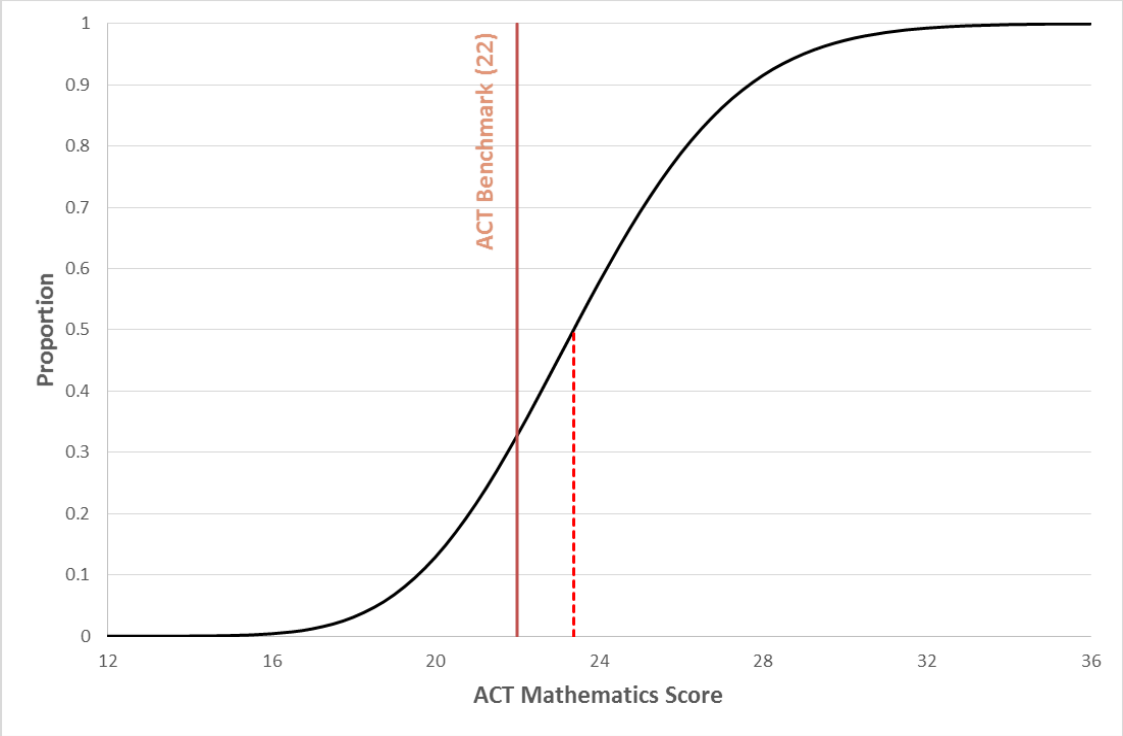

Figure 4*: Proportion of students meeting the NAEP mathematics Proficient achievement level of 176 for ACT mathematics scores*

Summary and Discussion

The objective of the study was to statistically relate NAEP and ACT assessments and use that relationship to identify a reference point or range on the NAEP 12th grade reading and mathematics scales reasonably associated with the ACT college readiness benchmarks for reading and mathematics. Identifying such points would potentially allow NAEP to report on the percentage of students at the 12th grade who are likely to achieve the ACT college readiness benchmarks associated with success in first-year college courses at typical post-secondary institutions. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential benchmarks on the NAEP scales that could form the basis for reporting preparedness at grade 12 (see Figures 1 and 2 for examples of how the markers were determined).

In addition, we identified the point on the ACT scale that corresponds most closely to the *NAEP Proficient* achievement level cut score, for grade 12 reading and mathematics, 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 moderately strong, meaning that the kind of relationship statements that can be made need to be presented in notions of probability rather than one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct and the test taking populations could be different due to the nature of the two tests. However, it does make interpretation somewhat more challenging. The results showed that 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 they differed slightly for mathematics. In particular, the reading *NAEP Proficient* achievement level cut score of 302 could form a reasonable basis for reporting on meeting the ACT college readiness benchmarks and therefore potentially academic preparedness for college at grade 12. However, the mathematics counterpart is 167 on the NAEP scale, about 9 points lower than the *NAEP Proficient* achievement level cut score for grade 12 mathematics. Going the other direction, the projection of the *NAEP Proficient* reading cut score on the ACT scale coincides with the existing ACT college readiness benchmark for reading, and is about 1 point higher than the ACT benchmark for mathematics.

The current NAEP-ACT linking study is closely related to the statistical linking study that connected NAEP and SAT on the national level (Moran et al., 2012). The national NAEP-SAT linking study used data from students who were sampled and assessed in NAEP 12th grade reading or mathematics in 2009 and had also taken 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 mathematics scales. These numbers are very close to the correlations calculated in the current study (i.e., 0.75 for reading, and 0.87 for mathematics). The projection results obtained from the national NAEP-SAT linking study (see Table 1 of Moran et al., 2012, $p = 0.5$) indicated that a NAEP reading scale score of 302 and a NAEP mathematics scale score of 164 could potentially be used as the thresholds on the NAEP scales to indicate academic preparedness for college. These reference points also correspond well with the newly identified thresholds on the NAEP scales through this NAEP-ACT linking study.

Important limitations must be kept in mind when interpreting the results from the current linking study. The correlation between the NAEP and ACT scores is moderately high, indicating considerable variance associated with the relationship between the two assessments. As a consequence, a statistical linking methodology was selected to express the ACT college readiness benchmarks on the NAEP scales. Such a relationship is not symmetric, and the two linked scales do not correspond to each other on a one-on-one level. Besides, the statistical relationship established between NAEP and ACT is not invariant across major population subgroups. Additional investigation and longitudinal studies are needed to evaluate the validity and meaningfulness of the identified benchmark points on NAEP as indicators of students being academically prepared for college.

References

ACT, Inc. (2013). *What Are the ACT College Readiness Benchmarks?* Iowa City, IA: ACT, Inc. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/benchmarks.pdf.

ACT, Inc. (2017). *The ACT Technical Manual*. Iowa City, IA: ACT, Inc. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.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 (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*, 227-246.

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. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Educational Testing Service, Policy Information Center.

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., & Freund, D. (2012). *NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between NAEP and SAT*. Research report prepared for National Assessment Governing Board.

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.* Washington, DC: Technical Panel on 12th Grade Preparedness Research.

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.

Strategic Vision Activities Led by COSDAM

During the November 2016 Board meeting, a Strategic Vision was formally adopted to guide the Board's work over the next several years. For each activity led by COSDAM, information is provided below to describe the current status and recent work, planned next steps, and the ultimate desired outcomes. Please note that many of the Strategic Vision activities require collaboration across committees and with NCES, but the specific opportunities for collaboration are not explicitly referenced in the table below. In addition, the activities that include contributions from COSDAM but are primarily assigned to another standing committee (e.g., framework update processes) or ad hoc committee (i.e., exploring new approaches to postsecondary preparedness) also have not been included below.

The Governing Board's Assistant Director for Psychometrics, Sharyn Rosenberg, will answer any questions that COSDAM members have about ongoing or planned activities.

Strategic Vision Activity	**Current Status and Recent Work**	**Planned Next Steps**	**Desired Outcome**
SV #2: Increase opportunities to connect NAEP to administrative data and state, national, and international student assessments *Incorporate ongoing linking studies to external measures of current and future achievement in order to evaluate the NAEP scale and add meaning to the NAEP achievement levels in reporting. Consider how additional work could be pursued across multiple subject areas, grades, national and international assessments, and longitudinal outcomes*	Ongoing linking studies include: national NAEP-ACT linking study; longitudinal studies at grade 12 in MA, MI, TN; longitudinal studies at grade 8 in NC, TN; NAEP-TIMSS linking study; NAEP-HSLS linking study; NAEP Validity Studies (NVS) studies Informational update on current studies was provided in the March 2018 COSDAM materials Results from the national NAEP-ACT linking study will be presented to COSDAM at the upcoming Board meeting	Complete ongoing studies Decide what new studies to take on Decide how to use and report existing and future results Complete additional studies	NAEP scale scores and achievement levels may be reported and are better understood in terms of how they relate to other important indicators of interest (i.e., other assessments and milestones)

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV #3: Expand the availability, utility, and use of NAEP resources, in part by creating new resources to inform education policy and practice *Research when and how NAEP results are currently used (both appropriately and inappropriately) by researchers, think tanks, and local, state and national education leaders, policymakers, business leaders, and others, with the intent to support the appropriate use of NAEP results (COSDAM with R&D and ADC)* *Develop a statement of the intended and unintended uses of NAEP data using an anticipated NAEP Validity Studies Panel (NVS) paper and the Governing Board's research as a resource (COSDAM with NCES)*	Ina Mullis of the NVS panel spoke with COSDAM at the March 2017 Board meeting and is working on a white paper about the history and uses of NAEP During the August 2018 Board meeting, COSDAM discussed how to use information from an ongoing study to inform a policy statement on intended and appropriate uses of NAEP At the upcoming Board meeting, COSDAM will continue discussing next steps for developing a policy statement on intended uses of NAEP	Use research to draft short document of intended and appropriate uses for COSDAM discussion NCES produces documentation of validity evidence for intended uses of NAEP scale scores Governing Board produces documentation of validity evidence for intended uses of NAEP achievement levels	Board adopts formal statement or policy about intended uses of NAEP. The goal is to increase appropriate uses and decrease inappropriate uses (in conjunction with dissemination activities to promote awareness of the policy statement)
Disseminate information on technical best practices and NAEP methodologies, such as training item writers and setting achievement levels	This idea was generated during the August 2017 COSDAM discussion of the Strategic Vision activities	Work with NCES and R&D to refine list of technical topics for dissemination efforts	Stakeholders benefit from NAEP technical expertise

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV# 5: Develop new approaches to update NAEP subject area frameworks to support the Board's responsibility to measure evolving expectations for students, while maintaining rigorous methods that support reporting student achievement trends *Consider new approaches to creating and updating the achievement level descriptors and update the Board policy on achievement levels*	Input for the policy revision was provided through a panel of standard setting experts, a literature review on considerations for creating and updating achievement level descriptors (ALDs), and a technical memo on developing a validity argument for the NAEP achievement levels (early 2018) COSDAM discussed the policy revision during the May and March 2018 Board meetings Full Board discussed the draft revised policy during the August 2018 Board meeting Public comment was sought from August 30 – October 15, 2018; Board calls to discuss the comments took place in October The revised policy was unanimously adopted during the November 2018 Board meeting	Board staff and COSDAM will work on implementing the revised policy on NAEP achievement level setting, including reviewing and updating achievement level descriptions	Board has updated policy on achievement levels that meets current best practices in standard setting and is useful for guiding the Board's achievement levels setting work

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV# 7: Research policy and technical implications related to the future of NAEP Long-Term Trend assessments in reading and mathematics *Support development and publication of multiple papers exploring policy and technical issues related to NAEP Long-Term Trend. In addition to the papers, support symposia to engage researchers and policymakers to provide stakeholder input into the Board's recommendation*	White papers commissioned, symposium held (March 2017), and follow-up event held at American Educational Research Association (AERA) conference (April 2017) Several Board discussions took place during 2017 and 2018 The NAEP budget in Fiscal Year 2019 was increased by $2 million with the goal of moving up the next administration of LTT Following discussion at the November 2018 Board meeting, Chair Bev Perdue sent a response to Congress indicating that the Board would add a paper-based 2020 LTT administration to the NAEP Assessment Schedule and	Board action to add a 2020 paper-and-pencil administration of LTT to the NAEP Assessment Schedule is planned for the upcoming Board meeting At a future Board meeting, NCES will discuss with COSDAM design considerations for transiting the LTT to a digital-based assessment in advance of the 2024 LTT administration	Determine whether changes to the NAEP LTT schedule, design and administration are needed (led by Executive Committee and NCES)
SV# 9: Develop policy approaches to revise the NAEP assessment subjects and schedule based on the nation's evolving needs, the Board's priorities, and NAEP funding *Pending outcomes of stakeholder input (ADC activity), evaluate the technical implications of combining assessments, including the impact on scaling and trends*	COSDAM presentation and discussion on initial considerations for combining assessments During the past 2 years, there have been several full Board presentations and discussions on the NAEP Assessment Schedule	Board action on the NAEP Assessment Schedule scheduled for May 2019	Determine whether new assessment schedule should include any consolidated frameworks or coordinated administrations

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV# 10: Develop new approaches to measure the complex skills required for transition to postsecondary education and career *Continue research to gather validity evidence for using 12th grade NAEP reading and math results to estimate the percentage of grade 12 students academically prepared for college*	Several studies are ongoing (see activities under SV# 2) During the November 2018 Board meeting, the Board took action to explore the creation of a postsecondary preparedness dashboard	Decide whether additional research should be pursued at grade 8 to learn more about the percentage of students "on track" to being academically prepared for college by the end of high school or whether additional research should be conducted with more recent administrations of NAEP and other tests Decide whether Board should make stronger statement and/or set "benchmarks" rather than using "plausible estimates"	Statements about using NAEP as an indicator of academic preparedness for college continue to be defensible and to have appropriate validity evidence