National Assessment Governing Board
Committee on Standards, Design and Methodology

August 2, 2013
10:00 am – 12:30 pm

AGENDA

10:00 – 10:10 am	Introductions and Welcome to Sharyn Rosenberg, the New Governing Board Assistant Director for Psychometrics *Lou Fabrizio, COSDAM Chair*	Attachment A
10:10 – 10:20 am	Committee Questions on Information Items (see below) *Cornelia Orr, Executive Director, NAGB* *Michelle Blair, Senior Research Associate, NAGB*	
10:20 – 11:20 am	Interpreting NAEP Proficient using Preparedness Research Findings ▪ Summary of Recent Feedback on the Validity Argument ▪ Report Mock-up Presentation *Lou Fabrizio, COSDAM Chair* *Ray Fields, Assistant Director for Policy and Research, NAGB*	Attachment B
	Closed Session	
11:20 – 12:25	Discussion on Achievement Level Setting (ALS) on the 2014 NAEP Technology and Engineering Literacy (TEL) Assessment ▪ 2013 TEL Pilot Scaling Analyses and Plans ▪ Implications for Planning the TEL ALS *Cornelia Orr, Executive Director, NAGB* *Andreas Oranje, ETS*	Attachment C
	Open Session	
12:25 – 12:30 pm	Other Issues or Questions *COSDAM Members*	
Information Item	Update on Evaluation of NAEP Achievement Levels Procurement	Attachment D
Information Item	NAEP 12th Grade Academic Preparedness Research: Phase 2 Research Updates ▪ Course Content Analysis Research ▪ National and State Partnerships ▪ Research with Frameworks	Attachment E

Welcome to Sharyn Rosenberg,
New Governing Board Assistant Director for Psychometrics

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

Abbreviated Professional Biography for Sharyn Rosenberg

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

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

Interpreting NAEP Results Using Preparedness Research Findings

At the August 2013 meeting, the Governing Board will discuss the way forward on reporting the NAEP 12th grade results from the 2013 reading and mathematics assessments. As background for the discussion, included in this tab are:
- a draft of a prototype chapter for the report (Attachment B-1; new document)
- the independent technical reviews of the preparedness validity argument by Gregory Cizek and Mark Reckase (Attachments B-2 and B-3; new documents)
- the preparedness validity argument (Attachment B-4; included in the May 2013 COSDAM briefing materials, but changes were made to the proposed inferences as described below and indicated in highlighting on pages B32 and B65)

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

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

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

The Board staff drew two main conclusions from the Board discussion in May about the validity argument:
- While finding the validity argument supportive, the Board wanted to consider the independent technical reviews that were to be presented at the August 2013 meeting to inform its decision making.
- The Board found the inferences that were being proposed to be "not quite there yet."

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

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above Proficient on Grade 12 NAEP is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities that would make them academically prepared for college.

Mr. Ho's formulation for the preparedness inference was shared with Michael Kane, who is advising Board staff on the validity argument. Mr. Kane supported using this formulation in place of the one originally proposed and suggested adding "or reasonable" after "plausible." Board staff revised the validity argument accordingly and it was this formulation that was considered by the independent technical reviewers of the validity argument.

Question for Board Consideration:

With the understanding that additional work will be required in collaboration with NCES, along with additional guidance from the Board, is the general approach exemplified in the prototype chapter (Attachment B-1) an acceptable basis for moving forward with reporting on academic preparedness for college as a part of the reporting of the NAEP 12th grade reading and mathematics assessment results for 2013?

Towards NAEP as an Indicator of Academic Preparedness for College and Job Training
Ray Fields July 18, 2013

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

INTRODUCTION

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

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

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

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

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

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

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

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

NAEP and ACADEMIC PREPAREDNESS

The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would afford an invaluable public service: providing an indicator of the human capital potential of today's and future generations of the nation's population.

The Board began this initiative in 2004, after receiving recommendations from a distinguished blue-ribbon panel that had examined whether NAEP should continue assessing at the 12th grade.

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

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

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

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

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

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

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

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

Figure1 about here (see page 8)

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

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

Performance by the same students on NAEP and the SAT mathematics and reading tests was correlated at 0.91 and 0.74, respectively.

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

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

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

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

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

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

A TRANSITION TO REPORTING ON ACADEMIC PREPAREDNESS
The reporting of the 12th grade results for 2013 represents a transition point for NAEP.

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

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

The Context for Academic Preparedness for College
In the United States in 2013, there is no single, agreed upon definition of "academic preparedness for college" used by colleges for admission and placement (Fields and Parsad). Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.

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

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

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

The definition proposed in this report is intended to apply to the typical degree-seeking entry-level student at the typical college. For NAEP reporting, "academically prepared for college" refers to the reading and mathematics knowledge and skills needed for placement into entry-level, credit bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement,

without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

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

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

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

Proposed Inferences for NAEP Reporting
The NAEP preparedness research does not affect the NAEP results in any way. The distribution of student achievement is unchanged. That is, the average scores, the percentiles, and the achievement level results are not impacted by the NAEP preparedness research.

The independent technical reviewers confirmed that the research findings support inferences about performance on NAEP 12th grade results in reading and mathematics in relation to academic preparedness for college.

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

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

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

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

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

The study results support these inferences. However, there will be students scoring at or above Proficient who are not academically prepared and students scoring below Proficient who are academically prepared (i.e.,

there will be false positives and false negatives). This will be true for any assessment program that sets cut-scores for a similar purpose.

Figure 2. about here (see page 9)

Figure 3. about here (see page 9)

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

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

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships

between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

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

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

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

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

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

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

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

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

Limitations on Interpretation and Other Caveats

False Negatives and False Positives

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

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

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

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

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

Not a Preparedness Standard

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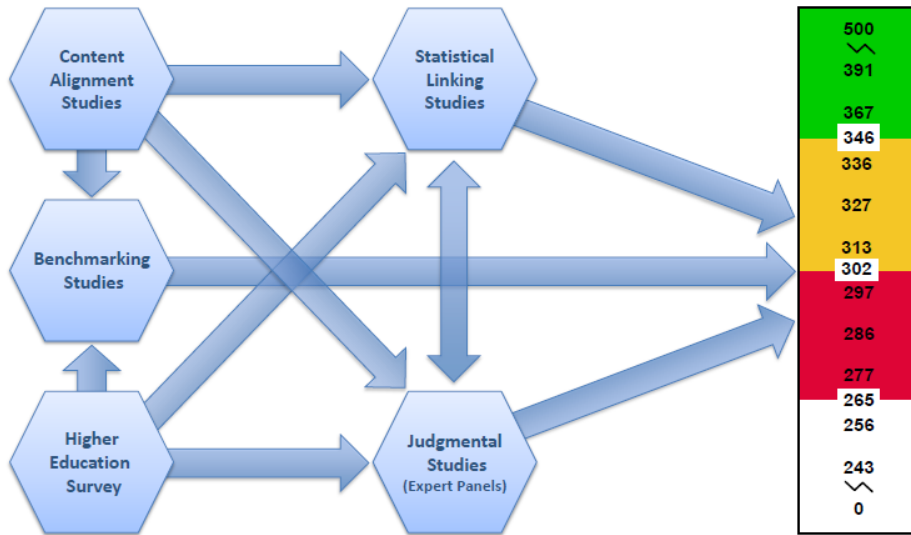
grade NAEP reading and mathematics results in NAEP reports.

GPA of B- or Better

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

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

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This

criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations

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

Preparedness for Job Training

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

Conclusion

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

The interpretations of NAEP results in relation to academic preparedness for college are being reported on a preliminary basis. They are provided to help foster public

understanding and policy discussions about defining, measuring, validating and reporting on academic preparedness for college by NAEP and more broadly.

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

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

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

Figure 1. Model of the Preparedness Research Program

Study Interrelationships

Figure 2.

Figure 3.

References to be added.

Review and Comment on

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

Prepared for:

National Assessment Governing Board

Prepared by:

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

June 30, 2013

Review and Comment on

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

Introduction

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

According to Kane (2013):

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

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

To produce this evaluation, I relied primarily on two documents that presented the NAEP preparedness validity argument and evidence (Fields, 2013a, 2013b). A draft response to *Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for Colleg*e (Fields,

2013a) was submitted to the National Assessment Governing Board on May 29, 2013 (Cizek, 2013). This paper is a response to a revision of *Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College* (Fields, 2013b)

The Proposed Interpretations and Claims

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

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

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

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

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

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

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

4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

Evaluation of Validity Evidence in Support of the Proposed Interpretations

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

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

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

1) To enhance the clarity of the proposed interpretations, I offer the following recommendation: *NAGB should consider making the score interpretations parallel by specifying the NAEP scale score associated with preparedness in Reading.*

As currently worded, a defensible and specific scale score associated with preparedness is offered for NAEP Mathematics score interpretations; however, the interpretation for Reading is phrased as an achievement level: "The percentage of students in the 12th grade NAEP distribution at or above (Proficient for reading and a score of 163 for mathematics) is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in (reading or mathematics) that would make them academically prepared for college."

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

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

2) To enhance the coherence of the proposed interpretations, I offer the following recommendation: *NAGB should consider conducting additional research into the content coverage of the NAEP and the alignment of NAEP with traditional college admissions measures.*

In its present form, it is argued that, in essence, the content of NAEP assessments in Reading and Mathematics covers everything that traditional college admissions measures (e.g., ACT, SAT, etc.) do, but also more. It is claimed that NAEP content coverage is "broader." The Venn diagram below illustrates this claim:[1]
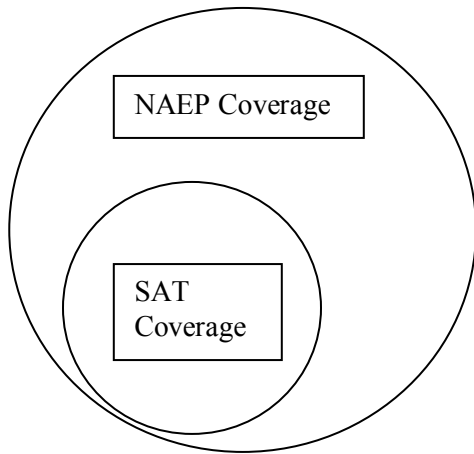
Figure 1
Hypothetical content coverage between NAEP Assessment and College Admissions Assessment

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

---

[1] The Venn diagram and the reference to the SAT are presented only illustrate content relationships between content coverage on assessments. The diagram is not intended to represent the actual proportional content coverage between NAEP and college admissions assessments, nor that of the SAT in particular.

NAEP Reading (and be deemed prepared for college) by getting very little of the "SAT-like" content correct on NAEP (that is, content deemed necessary for college success) and getting a lot of the "other" NAEP content correct (that is, the additional/broader content that may or may not necessarily be relevant to college preparedness).

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

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

Conclusion

The logical and empirical evidence gathered to date provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear, coherent, and comprehensive. Recommendations were presented for future strategies to strengthen the validity

case. Nonetheless, based on the empirical evidence and logical rationales to date, there appear to be strong warrants for the intended interpretations regarding NAEP reporting and indications of college preparedness in Reading and Mathematics.

References

Cizek, G. J. (2013, May). Response to *Draft validity argument for NAEP reporting on 12th grade academic preparedness for college*. Report prepared for the National Assessment Governing Board, Washington, DC.

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

Fields, R. (2013, May 9). *Draft validity argument for NAEP reporting on 12th grade academic preparedness for college*. Washington, DC: National Assessment Governing Board.

Fields, R. (2013, June 17). *Draft validity argument for NAEP reporting on 12th grade academic preparedness for college*. Washington, DC: National Assessment Governing Board.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

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

**Mark D. Reckase
Michigan State University
July 10, 2013**

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

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

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

A More Detailed Analysis

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

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

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

This statement is very rich in meaning. To fully understand it, some background information is assumed to be known by the reader. Some of this background is listed here, but the list may not be comprehensive.
1. NAEP produces an accurate representation of the distribution of achievement of students in the areas of reading and mathematics.
2. The estimate of the proportion of students above a cut score on the NAEP reporting score scale is fairly accurate.
3. Students who are estimated to be above the specified cut score are likely to have high school grades and college admissions test scores that will make them eligible for admission to college.
4. Those students who are eligible for admission attend college and enroll in entry-level, credit-bearing courses.
5. The skills and knowledge in reading and mathematics are prerequisite to learning the content presented in the entry-level, credit-bearing courses.

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

A Slight Variation on the Validity Argument

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

with the validity argument in the NAGB report, but rather to add some commentary on that argument.

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

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

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

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

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

6. A cut score on NAEP is consistent with the threshold.
    a. There is a good process for identifying a reasonable cut score on NAEP to correspond to #5.
    b. The combination of information from different tests results in strong support for parts of the argument.

7. The proportion of students estimated to be above the cut score on NAEP gives a good estimate of the proportion who exceed the threshold for admission into entry level courses.
   a. This is well supported by the statistical analysis procedures if the argument for an appropriate cut score is supported. In this case, there is reasonable support for the cut score from the connection to placement and admissions tests.

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

Reference

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1 – 73.

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

Ray Fields – July 7, 2013

Introduction

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

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

The Governing Board's initiative on 12th grade academic preparedness began in March 2004, with the report of a blue-ribbon panel.[1] The panel was composed of K-12 education leaders—the "producers" of high school graduates—and leaders in business, postsecondary education, and the military—the "consumers" of high school graduates.

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

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

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

---

[1] The blue-ribbon panel was known officially as the National Commission on NAEP 12th Grade Assessment and Reporting.

"Entrance into the military" was subsumed by "job training" with the intention of identifying occupations with civilian and military counterparts and utilizing the military's experience as the world's largest occupational training organization and its extensive research on the relationship between performance on the Armed Service Vocational Aptitude Battery (ASVAB) and job training outcomes.

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

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

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

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

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

Twelfth grade is the end of mandatory schooling for most students and represents the transition point to adult postsecondary pursuits. If it is essential for students to graduate from high school

academically prepared for college and job training, it is essential for the public and policymakers to know the degree to which this is occurring.

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

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

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

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

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

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

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

> … the academic knowledge and skill levels in reading and mathematics necessary to be qualified for placement…into a credit-bearing entry-level general education course that fulfills requirements toward a two-year transfer degree or four-year undergraduate degree

at a postsecondary institution [without the need for remedial coursework in those subjects]. (National Assessment Governing Board, 2009; p.3.)

This definition was intended to apply to the "typical" college, not to highly selective institutions, and thus, to the vast majority of prospective students, or about 80% of the college freshmen who enrolled in 2-year and 4-year institutions within 2 years following high school graduation (Ross, Kena, Rathbun, KewalRamani, Zhang, Kristapovich, and Manning, p 175). To make this clear, the definition is further elaborated as follows.

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

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

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

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

The Central Issue: Validity
Having made the decision to determine the feasibility of NAEP reporting on 12th grade academic preparedness, the Governing Board recognized that the central concern would be establishing the validity of inferences about 12th grade academic preparedness that are to be made from NAEP scores and used in NAEP reports. The Governing Board would need to ensure that the content of NAEP 12th grade reading and mathematics assessments was appropriate for measuring academic preparedness and that research was conducted to collect evidence by which the validity of

proposed inferences could be evaluated. Finally, a formal validity argument would need to be developed, specifying the proposed inference(s) for NAEP reporting, the underlying assumptions or propositions, and the evidence related to the assumptions or propositions.

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

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

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

Validity Argument

Overview
The National Assessment of Educational Progress (NAEP) program is designed to provide information about student achievement in reading, mathematics and other content areas at the 4th, 8th, and 12th grades. The items for the assessments are developed according to content frameworks and test specifications developed by the National Assessment Governing Board. Scientific sampling procedures are used to produce estimates of score distributions representative of the national population of students at each grade level, as well as estimates representative of public school students in individual states and in 21 urban school districts. The NAEP results do not produce scores for individual students, but rather, group estimates. The NAEP results are reported, based on the estimated score distributions, by average score, percentiles, and in terms of the percentages of students at or above three performance standards used for NAEP reporting, called achievement levels, that are designated Basic, Proficient, and Advanced.

The purpose of the research reported here was to examine whether the interpretation of 12th grade NAEP results in reading and mathematics could be extended to include statements about the percentage of U.S. 12th graders who are academically prepared for college and, if such an interpretation were found to be defensible, to determine the specific statements about academic preparedness that were supportable by the research evidence. The specific statements would be based on the following general definition for academic preparedness, used in relation to the NAEP preparedness research:

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

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

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

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

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

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

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

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

4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

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

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

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

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

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

Proposed Inferences

<mark>**For reading:**</mark>

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

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

<mark>**For mathematics:**</mark>

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

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

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

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

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

Limitations on Interpretation and Other Caveats

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

- likely to be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and
- not likely to need remedial/developmental courses in reading or mathematics in college,

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

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

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

Not a Preparedness Standard

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

Academically Prepared for College

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

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

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

The proposed inferences are focused on the first year of college; they do not support conclusions about college persistence beyond the first year or completion of a degree. The inferences will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.

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

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

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

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

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

Discussion of the Claims and Evidence

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

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

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

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

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

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

In designing the NAEP preparedness research program, the Governing Board determined that it would be essential to examine how performance on NAEP relates to performance on other

measures and outcomes associated with academic preparedness for college. The research program studied the relationship between performance on NAEP and performance on the SAT and ACT college admission tests, including the respective college readiness benchmarks that had been established by these testing programs.

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

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

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

Indicators: College Board and ACT College Readiness Benchmarks

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

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

The Governing Board has a partnership with the state of Florida as a part of the Board's program of preparedness research. Florida was one of 11 states that volunteered to provide state-

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

Analysis of Results for Mathematics

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

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

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

Analysis of Results for Reading

The statistical linking study examining performance on the NAEP 12th grade reading assessment and the SAT critical reading test resulted in a correlation of .74. Although it may not be high enough to predict the performance of individual students from one test to another (which is not required to support the proposed inference for reading), it is sufficient to support the group-level inferences reported by NAEP.

Performance on NAEP was also examined in relation to the SAT College Readiness Benchmark for critical reading (i.e., a score on the SAT critical reading test of 500). The SAT benchmark provides "an indication of college readiness at a typical college (College Board)." This is consistent with the Governing Board's definition of academic preparedness discussed in the results for mathematics above.

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

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

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

Analysis of Results for Mathematics

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

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

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) because it is the most recent of the national studies. The average NAEP score is 163 for students with an SAT score **at** the College Readiness Benchmark for mathematics of 500.

The other study results are consistently convergent with the NSLS results. The average NAEP mathematics scores for 12th grade students scoring **at** the SAT College Readiness Benchmark of 500 for mathematics are compared first for the 2005 HSTS and the 2009 NSLS. The average scores are 161 and 163 respectively.

Figure 1

NAEP 12th Grade Preparedness Research: Mathematics

Average Scores and Inter-quartile Ranges For Selected Variables, 2005 High School Transcript Study, 2009 High School Transcript Study, 2009 NAEP/SAT Linking Study, 2009 Florida Longitudinal Study
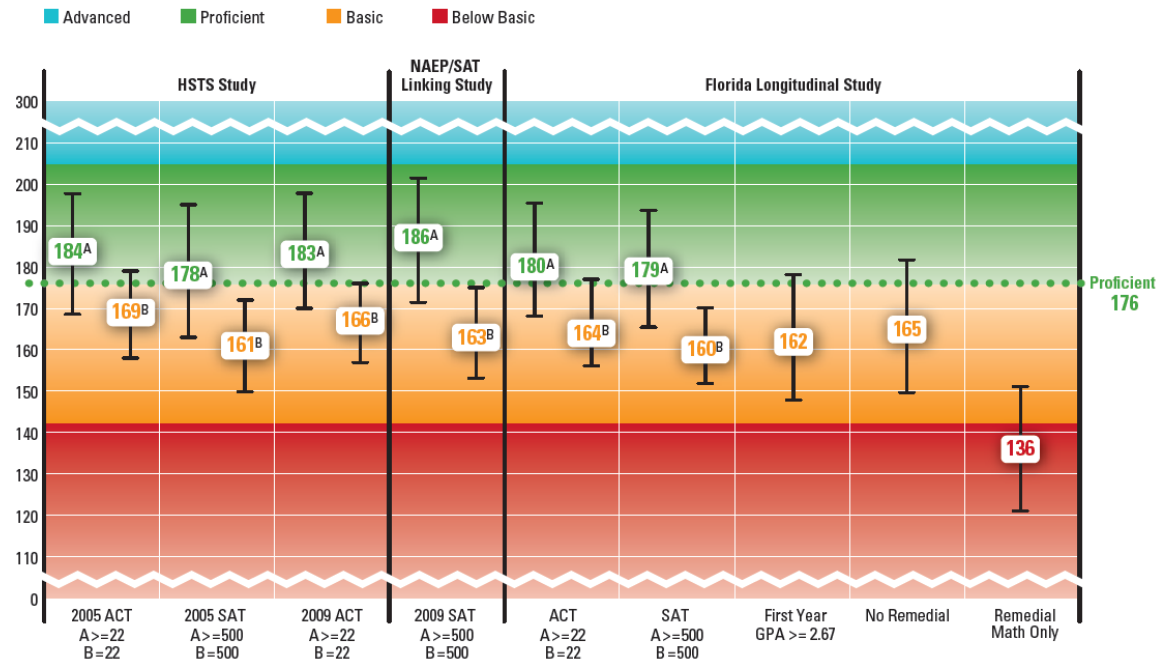
These results are confirmed by the FLS. The average NAEP mathematics score for the 12th grade Florida NAEP test takers who scored at the SAT College Readiness Benchmark of 500 was 160, much like the 2009 NSLS results and the 2005 HSTS results.

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

To answer the question, "What is the relationship between performance on NAEP and actual student outcomes?", we look to the FLS results. First we examine the average NAEP mathematics score for the 12th grade Florida NAEP test takers who attained a first-year GPA of B- or better. The average NAEP score for these students was 162. This is consistent with the SAT College Readiness Benchmark analyses and further supports the inference that students at

or above 163 on the 12th grade NAEP mathematics scale are likely to be academically prepared and attain a first-year GPA of B- or better. It follows, of course, that students who are academically prepared will not require remedial courses.

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

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

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

Analysis of Results for Reading

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

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

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) for reading, because it is the most recent of the national studies. The average NAEP score is 301 for students with an SAT score **at** the College Readiness Benchmark for critical reading of 500. A NAEP score of 301 in 12th grade reading is not significantly different from the cut-score for the 12th grade Proficient achievement level (302).

The FLS results are consistently convergent with the NSLS results. The average NAEP reading score was 299 for the 12th grade Florida NAEP test takers who were not placed into remedial courses in their first year. The average score was 298 for those who had a first year overall GPA of a B- or better. These data, which show the relationship between performance on NAEP and actual student outcomes, provide strong confirmation that students scoring at or above Proficient on the NAEP 12th grade reading assessment are likely to be academically prepared for ECNRG college courses.

Figure 2

NAEP 12th Grade Preparedness Research: Reading

Average Scores and Inter-quartile Ranges For Selected Variables for the 2009 NAEP SAT Linking Study and 2009 Florida Longitudinal Study
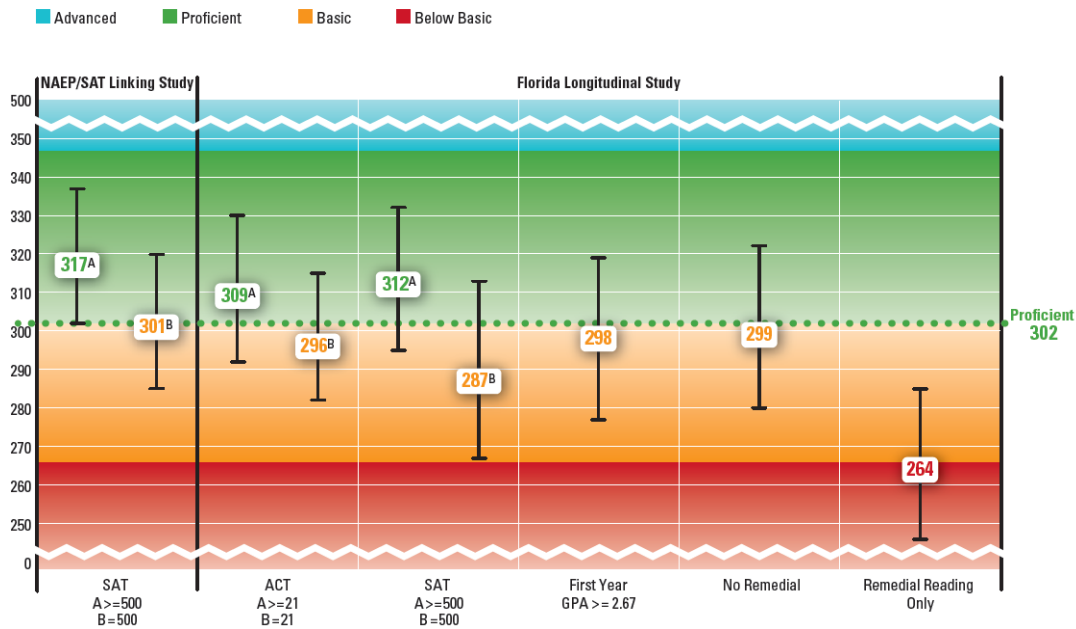
As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for reading is defined differently than the SAT College Readiness Benchmark for reading. However, it is noteworthy that even with this different definition, the ACT results from the 2009 FLS analysis are similar to the NSLS analysis and the FLS outcome data.

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

In conclusion, these results suggest that the percentage of students at or above the Proficient level in reading on 12th grade NAEP would provide a plausible (or reasonable) estimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.

4.   The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

> ➢ **NAEP Assessment Frameworks Were Revised to Measure Academic Preparedness**

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

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

For NAEP to report on 12th grade academic preparedness for college, it must measure relevant content at the 12th grade. The content of each assessment is determined by the NAEP assessment frameworks, which the Governing Board is responsible for developing and approving. Accordingly, the Governing Board decided that the extant NAEP frameworks intended for the 2009 for reading and mathematics at the 12th grade would be reviewed. The review would identify changes needed to measure 12th grade academic preparedness for college.[2] Examples of the changes made are described in the next two subsections.

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

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

---

[2] The review also addressed academic preparedness for job training, but that part of the NAEP preparedness initiative is not being addressed in this validity argument.

community colleges, as well as four-year state institutions, but generally not the more highly selective "flagship" campuses." (Achieve, 2004, p. 107)

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

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

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

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

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

The NAEP reading framework used for the 2009, 2011, and 2013 assessments contains the following statement

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

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

Mathematics

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

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

Conclusion

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

Examples of Objectives added to the 2009 Grade 12 Mathematics Framework

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

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

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

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

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

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

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

h) * Represent situations and solve problems involving polar coordinates.
Data Analysis, Statistics, and Probability
c) * Draw inferences from samples, such as estimates of proportions in a population, estimates of population means, or decisions about differences in means for two ""treatments"."

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

k) * Use the binomial theorem to solve problems.

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

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

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

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

g) * Determine the sum of finite and infinite arithmetic and geometric series.

> ➢ **Content Alignment Studies Found Significant Overlap between NAEP and the ACT, SAT and ACCUPLACER**

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

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

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

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

The ACCUPLACER has the purpose of "… [determining] which course placements are appropriate for [incoming college] students and whether or not remedial work is needed." (ACCUPLACER, p. A-2)
The SAT, ACT and ACCUPLACER in reading and mathematics are widely used for these purposes by admissions and placement professionals in postsecondary education institutions. These testing programs regularly conduct curriculum surveys, validity studies and other research to support their claims that the content measured is directly related to the reading and mathematics knowledge and skills needed to qualify for entry-level credit-bearing courses (e.g., see the ACT curriculum studies for 2012, 2009, 2005, and 2002 at http://www.act.org/research-policy/national-curriculum-survey/, and the College Board *National Curriculum Survey on English and Mathematics* at http://research.collegeboard.org/publications/content/2012/05/national-curriculum-survey-english-and-mathematics .

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

The second reason for conducting the content alignment studies was to provide information for interpreting the results of planned statistical linking studies between NAEP and the other tests, which measure academic preparedness for college. The linking studies were designed to examine how performance on NAEP compares with performance on the other tests, with the

purpose of supporting inferences about academic preparedness for college. For NAEP to support inferences about academic preparedness for college based on the linking studies, a sufficient content match would be needed between NAEP and the other tests, not just a statistical relationship.

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

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

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

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

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

The Content Alignment Studies: Reading Results

Reading: ACT

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

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

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

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

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

> "A total of 152 Reading items (comprising 17 blocks) were classified in [the reading] study. Of these, 97 were multiple-choice (MC). Nine were dichotomously-scored ("incorrect" or "correct") short constructed-response (DSCR) items. Thirty-three were polytomously-scored short constructed-response (PSCR) items, each scored using a three-point scoring rubric. Thirteen were extended constructed-response (ECR) items,

each scored using a four-point rubric. Each DSCR had one creditable score category, each PSCR had two, and each ECR had three. Each Reading panelist, therefore, assigned a total of 211 classifications to the NAEP Reading items [and rubric scoring categories]." (p.54)

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

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

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

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

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

Reading: SAT

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the SAT critical reading test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found

at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf

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

> ### What is the correspondence between the reading content domain assessed by NAEP and that assessed by SAT?
> The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, Integrate/Interpret," including many to Goal 2.1, "Make complex inferences within and across *both literary and informational texts*."
>
> Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.
>
> ### To what extent is the emphasis of reading content on NAEP proportionally equal to that on SAT?
> Both tests had many of their item alignments to the same NAEP "Integrate/Interpret" objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:
> * the "main idea" and "author's purpose" (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
> * the "tone and attitude" of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
> * the use of "rhetorical strategies" (NAEP Objective 2.1.d and SAT Objective B.1.2); and
> * connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).
>
> Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP "Integrate/Interpret" and aligned to SAT "Passage-Based Reading" Objectives B.1.1– B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency…
>
> Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP's emphasis on the understanding of

word meaning in context; the assessment is not intended to measure students' prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students' success on the items almost certainly depends on their prior knowledge of word definitions.

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

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

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

Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire reading subdomains are missing or not aligned?
With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards "Locate/Recall" and "Critique/Evaluate." This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.

With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

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

Reading: ACCUPLACER

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

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

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

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

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

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

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

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

Are there systematic differences in content and complexity between the NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between the NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire reading subdomains are missing or not aligned?
In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications [review]… ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area of alignment of ACCUPLACER items to the NAEP framework, NAEP

objectives in "Locate/Recall" and "Integrate/Interpret," applied to informational text only or to both informational and literary text.

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

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

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

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

NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

The Content Alignment Studies: Summary Discussion for Reading

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

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

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

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

longer reading passages and the inclusion of both short and extended constructed-response items."

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

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

The Content Alignment Studies: Mathematics Results

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

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

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

> "… the two assessments have much of their content domains in common. However, in the NAEP-to-ACT comparison, the difference in specificity with which the domains are articulated in the assessment documents left the panel uncertain as to whether a number

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

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

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

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

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

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

Finally, although the study results support the conclusion that the 12th grade NAEP mathematics assessment measures content that is also covered by other assessments designed and used to make judgments about academic preparedness for college, it is noted that the study was

conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in mathematics.

Mathematics: SAT

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

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

> ***"What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by SAT?***
> At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.
>
> Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions.
>
> ***To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on SAT?***
> The greatest commonality between the two tests is their emphasis at the standard level. This is evident in the distribution of percentages of total hits from both assessments matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP item pool's greater proportion of alignment to SAT "Data analysis, statistics, and probability," and the SAT short-version's greater proportion of alignment to SAT "Geometry and measurement," the proportions of alignments to "Algebra and functions" and "Number and operations" are comparable. There is also considerable overlap among some specific skills, with both assessments addressing many of the same NAEP "Number properties and operations" objectives and SAT objectives…

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both assessments emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

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

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

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

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

Mathematics: ACCUPLACER

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test. The ACCUPLACER is used specifically to determine whether entry-level students have the mathematic knowledge and skills necessary for college level work or require remedial mathematics courses.

WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf.

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

> ***"What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?***
> The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP's "Number properties and operations" and "Algebra" standards and in ACCUPLACER's "Arithmetic," "Elementary algebra," and "College level math" standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced). This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.
>
> ***To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?***
> Regarding alignment to the NAEP framework, within the "Number properties and operations" and "Algebra" standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in "Number properties and operations" and "Algebra"). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP "Algebra" objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.
>
> Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—"Arithmetic," "Elementary algebra," and "College level math"—with a relatively even distribution, although

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

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

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

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

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

The Content Alignment Studies: Summary Discussion for Mathematics

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

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

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

The ACCUPLACER study found that "The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra…although their respective degrees of alignment and focus in these subdomains vary… the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from "Number properties and operations" in the NAEP framework, and four of the five goals in "Algebra." Conversely, the

NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives…Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives…"

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

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

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

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

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

Currently, there are either no or very low stakes consequences associated with the use of NAEP results. NAEP is not used as a basis for evaluating or diagnosing individual students, classroom or school performance, the effectiveness of individual teachers or administrators, or for any other accountability purpose. This will not change as a consequence of the inferences proposed in this validity argument.

Although the uses and consequences of NAEP will not change, employing the proposed inferences for NAEP reporting will bring a potential for misinterpretation. NAEP reports should include text explaining the limitations on interpretation and other caveats that were discussed in detail on pages 8-10 above.

Summary and Conclusion

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

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

The results of these analyses were consistent across studies and across years. They support the conclusions that students in the NAEP 12th grade distribution at or above the Proficient achievement level in reading and at or above 163 on the NAEP score scale for mathematics are
- likely to be academically prepared for entry-level, credit-bearing non-remedial courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions, and
- not likely to need remedial/developmental courses in reading or mathematics in college

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

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

For reading:

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

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

For mathematics:

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

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

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

References

Achieve (2004). *Ready or Not: Creating a High School Diploma that Counts* Washington, DC: Author.

ACT Technical Manual; http://www.act.org/aap/pdf/ACT_Technical_Manual.pdf

ACCUPLACER on-line technical manual; http://isp.southtexascollege.edu/ras/research/pdf/ACCUPLACER_OnLine_Technical_Manual.pdf

AERA/APA/NCME *Standards for Educational and Psychological Testing (1999)*

Allen, Jeff, Sconing, Jim (2005). *Using ACT Assessment Scores to Set Benchmarks for College Readiness (ACT Research Series 2005-3).* Iowa City, IA: ACT, Inc,

Bozick, R., and Lauff, E. (2007). *Education Longitudinal Study of 2002 (ELS:2002): A First Look at the Initial Postsecondary Experiences of the Sophomore Class of 2002* (NCES 2008-308). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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

Conley, D.T. (2007). *Toward a More Comprehensive Conception of College Readiness*. Eugene, OR: Educational Policy Improvement Center.

Fields, R. & Parsad, B. (2012). *Tests and Cut Scores Used for Student Placement in Postsecondary Education: Fall 2011*. Washington, DC: National Assessment Governing Board.

Kane, Michael T. (2013, Spring). *Validating the Interpretations and Uses of Test Scores.* Journal of Educational Measurement, Vol. 50, No. 1, pp. 1-73.

Kim, Y.K., Wiley, A., and Packman, S., *National Curriculum Survey on English and Mathematics*, College Board Research Report 2011-13 (New York: The College Board, 2011) http://professionals.collegeboard.com/data-reports-research/cb/RR2011-13

National Assessment Governing Board. (2008, September). *Mathematics Framework for the 2009 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board. (2008, September). *Reading Framework for the 2009 National Assessment of Educational Progress*. Washington, DC: Author

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

National Center for Education Statistics (2010).*The Nation's Report Card: Grade 12 Reading and Mathematics 2009 National and Pilot State Results* (NCES 2011–455). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Commission on NAEP 12th Grade Assessment and Reporting. (2004, March 5). *12th Grade Student Achievement in America: A New Vision for NAEP*. Washington, DC: Author.

Ross, T., Kena, G., Rathbun, A., KewalRamani, A., Zhang, J., Kristapovich, P., and Manning, E. (2012). *Higher Education: Gaps in Access and Persistence Study* (NCES 2012-046). U.S. Department of Education, National Center for Education Statistics. Washington, DC: Government Printing Office.

WestEd ACCUPLACER Mathematics Report; http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Mathematics_Content_Comparison.pdf.

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

WestEd SAT Mathematics Report; http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pd.f

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

Wyatt, Jeffrey, Kobrin, Jennifer, Wiley, Andrew, Camara, Wayne J., and Proestler, Nina (2011). *SAT Benchmarks: Development of a College Readiness Benchmark and its Relationship to Secondary and Postsecondary School Performance (Research Report 2011-5)*. New York, NY: College Board.
http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2011-5-sat-college-readiness-benchmark-secondary-performance.pdf

Setting Achievement Levels on the NAEP 2014
Technology and Engineering Literacy (TEL) Assessment

Status: Information and discussion

Objective: To discuss issues that are being addressed and that should be addressed in TEL scaling analyses.

Attachment: C-1 NCES description of current issues being reviewed in TEL Scaling.

C-2 Overview of the Evidence-Centered Design Method in the article "Evidence-Centered Design for Certification and Licensure" (Williamson, Mislevy, and Almond, 2004)

Background

At the March 1, 2013 meeting, the Committee began discussion on setting achievement levels for the 2014 NAEP TEL assessment. For the May 17, 2013 meeting, an issues paper was developed to support procurement and project planning for developing recommended achievement levels for TEL. In the Committee's May 2013 discussion, the Committee expressed a need for more information before proceeding with procurement plans, particularly regarding TEL scaling issues that could hinder a strong TEL Achievement Level Setting (ALS) effort.

Timeline

The following timeline provides a preliminary list of key dates and activities related to TEL assessment development and achievement level setting.

Date	Activity	Responsibility
2008 - 2010	TEL Framework development	ADC, Board, WestEd (contractor)
2010 - 2012	Assessment development for 2013 pilot test	NCES, NAEP contractors
2010 - 2012	Item review for 2013 pilot test	NCES, NAEP contractors, TEL Standing Committee, ADC
Early 2013	Pilot test – national sample, grade 8	NCES, NAEP contractors
May 2013	TEL ALS issues paper	COSDAM, consultant
Late 2013	ALS procurement and contract award	Board staff, COSDAM
Early 2014	Operational administration – national sample, grade 8	NCES, NAEP contractors
2014 - 2015	Final phase of ALS process and Board action on TEL	COSDAM, ALS contractor, Board
2015	Reporting TEL results	Board, NCES, contractors

TEL Assessment Design

The 2014 Technology and Engineering Literacy (TEL) assessment is based on the Board-adopted Framework and Specifications (see Abridged TEL Framework in Attachment D-2; complete documents are at www.nagb.org, Publications).

The TEL assessment is composed of three major areas:
- Design and Systems
- Information and Communication Technology
- Technology and Society

Another key dimension of the TEL assessment is the three practices, each of which is applicable to the three major areas noted above:
- Understanding Technological Principles
- Developing Solutions and Achieving Goals
- Communicating and Collaborating

The TEL assessment was developed using an evidence-centered design (ECD) approach (see Attachment C-2). From the beginning, all TEL tasks and items were designed using an evidential chain of reasoning that links what is to be measured, the evidence used to make inferences, and the tasks used to collect the desired evidence. In addition to student responses to complex tasks and discrete items, the computer-based TEL assessment allows NAEP to capture a wide array of data on student performance. For example, NAEP will collect information on how students interact with the TEL simulations and experiments. Such data may include the number of experimental trials run and the number and types of variables controlled. These observable data on "strategies and processes" may also contribute to the scoring of student performance.

TEL Reporting

Based on the ECD approach, TEL reporting will be expanded beyond the traditional NAEP scores. It is expected that data from the complex performance tasks and discrete items will be reported in a number of ways:
- A composite scale score on which the achievement levels will be set
- Subscores for the content areas (Design and Systems; Information Communication Technology; Technology and Society)
- Reporting on the practices (Understanding Technological Principles; Developing Solutions and Achieving Goals; Communicating and Collaborating)
- Information on students' processes and strategies, related to the ECD model, captured as observable data from their work on the TEL scenario-based tasks.

Ongoing Potential Discussion Questions for COSDAM
- Given the emerging field of setting achievement levels on ECD-based complex performance assessments, what additional background materials are needed to inform the COSDAM/Board decision on an appropriate method for ALS on the TEL assessment?
- To what extent should research studies be built into the TEL ALS project?

Technology and Engineering Literacy Field Trial Analyses

As part of the 2013 NAEP administration, a large field trial (also often referred to as 'pilot' in similar contexts) was conducted for the Technology and Engineering Literacy assessment. The field trial was designed to be similar to other field trials in subjects for which an entirely new framework is used. Specifically, the trial was designed to support both a detailed evaluation of the items and tasks individually as well as how they relate to each other (e.g., through scaling and otherwise correlation-based analyses) using a partial balanced incomplete block design. The analysis of the field trial data focuses on three goals:

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

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

3. Further development of extended reporting goals. Extended reporting refers to results based on task-use patterns and other process data (e.g., strategies students use to solve a particular problem, consistency and efficiency in running a particular simulation), is exploratory in nature, is generally task-dependent, and could serve to provide additional context to the broader, generalizable scaled results (i.e., domains and competencies). Specifically, these indicators are related to TEL, but not measurements of TEL.

In addition, the field trial data could be used to further an achievement level setting effort associated with the TEL assessment. Particularly, the scaling analyses in combination with item maps can provide an approximate grouping of items and levels of performance. There are a number of subtleties that will be addressed during this presentation, including the extent to which the field trial results and the probe results are sufficiently comparable, how context and positioning effects may play a role, and how administration design differences could affect outcomes. In addition, operationalizing a new construct in NAEP suggests the need for some dimensionality analyses to determine at what level (e.g., overall, by domain) meaningful standards can and ought to be set.

Besides discussing scaling goals and challenges, in this presentation we will also discuss the kinds of data and reports that could be generated based on the field trial data. In principle, the TEL field trial data are very close *in design* to a regular assessment and should, therefore, provide the similar kinds of summary performance results that have been used in other achievement level setting activities. However, discrete item and survey question selection does result in comparability issues. In addition to item responses, we also collected time-stamped process data and can explore whether there are other types of data (e.g., behavioral data) that could potentially be useful as corroborating information to standard setting efforts.

EVIDENCE-CENTERED DESIGN FOR CERTIFICATION AND LICENSURE

David M. Williamson, Robert J. Mislevy, Russell G. Almond
Educational Testing Service

What is Evidence-Centered Design?

Evidence-Centered Design (ECD) (Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003) is a methodology applied at Educational Testing Service that emphasizes an evidentiary chain of reasoning for assessment design. This approach results in a more complete representation of the design rationale for an assessment, better targeting of the assessment for its intended purpose, and a more substantial basis for a construct-representation validity argument supporting use of the assessment. The approach encourages test developers to design with intent and provides several advantages:

Clarity of purpose – representation of assessment goals and the relevance of design decisions to those goals.

Interrelated design – modeling the interactions of design decisions and how changes in one aspect of design affect other design elements.

Evidentiary requirements – explication of what constitutes relevant evidence of ability and how such evidence bears on assessment-based decision-making.

Validity – a documented chain of reasoning and rationale underlying design decisions and their relevance to the criterion of interest.

Innovation – a guide for developing assessments targeting elusive domain constructs or using emerging technologies and new item types.

The foundations of ECD stem from validity theory (Messick, 1989), psychometrics (Mislevy, 1994), philosophy (Toulmin, 1958), and jurisprudence (Wigmore, 1937). They adapt the evidence-oriented approach to evaluating the degree to which conclusions about people can be made on the basis of collected evidence. The ECD process centers around four key questions:

1. *Claims*: Who is being assessed and what will be declared about them as a result?

2. *Proficiencies*: What proficiencies must be measured to make appropriate decisions?

3. *Evidence*: How will we target, recognize, and interpret evidence of these proficiencies?

4. *Tasks*: Given practical constraints, what situations will elicit the kind of evidence needed?

Addressing these questions results in three fundamental assessment design models, represented here as Figure 1. These ECD models include:

- *Proficiency Model* – defines the claims and constructs of interest for the assessment and their interrelationships.

- *Evidence Models* – define how observations of behavior are considered as evidence of proficiency.

- *Task Models* – describe how assessment tasks must be structured to ensure opportunities to observe behaviors constituting evidence.

These interrelated models comprise a chain of reasoning for an assessment design that connects the design of assessment tasks to evidence of proficiencies targeted by the assessment, which in turn are formally associated with claims made on the basis of assessment results.

**Figure 1: Fundamental Models of
Evidence-Centered Design**
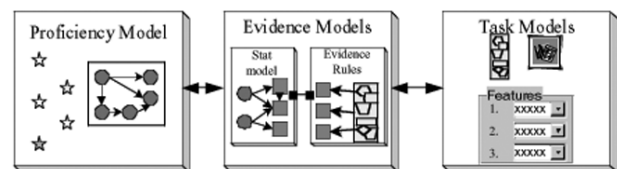
David M. Williamson, Robert J. Mislevey and Russell G. Almond are _____ with Educational Testing Service.

The following presents each of these models in turn with some discussion of their implications in the context of certification and licensure testing.

Proficiency Model

The proficiency model is really a combination of the formal assessment *claims* to be made on the basis of assessment and the *proficiencies* measured by the test. *Claims* are the specific arguments being made about people on the basis of assessment results. *Proficiencies* are measured knowledge, skills and abilities of people that provide the basis for making claims.

In order to make such claims or to identify important proficiencies, one must first have a good understanding of the population being served. Therefore, a precursor to claim specification is a definition of the examinee population, the users of the test results, and the intended use of test results in decision-making by these users. In certification and licensure testing the decision being made on the basis of the assessment is typically straightforward: either to issue or withhold the credential in question. Based on this definition, the sole users of test results are the issuing body of the credential.[1] However, since the credential itself represents a claim about the examinee made by the credentialing organization, it is typical to consider the interests of the users of these credentials (e.g., potential employers, the general public selecting their services, state licensure boards, etc.) when establishing claims. The examinee population is typically defined as individuals who have met some educational and/or practice prerequisites and are seeking the credential in question. Implicit in this definition is the perceived value of the credential and how it benefits the personal and professional interests of the examinee.

The understanding of assessment use and population being served drives the specification of claims being made on the basis of assessment results. These claims are represented as stars in the Proficiency Model portion of Figure 1. For example, in licensure testing a common global claim made on the basis of assessment might be something like, "Can engage in professional practice without representing a risk to the health, safety or well-being of the public." Often, such a global claim about ability is supported by a number of subclaims that make explicit statements intended to directly support the overall claim of the assessment. Often these are based on elements of the domain of practice that are ultimately reflected in test content. In this way, it is typical for the claims associated with an assessment design to be organized as a claim hierarchy that elaborates the various arguments that a test score represents about individual ability. As such, the specific claims chosen for an assessment design are often directly related to needs of score reporting or delivery of instruction.

The proficiencies of individuals being measured by the assessment follow from the claims. The claims express the goals of assessment design as states of knowledge about aspects of proficiency and represent the declarations that must be supported by test results. In order to support these arguments, certain levels of ability must be demonstrated during the assessment. It is these proficiencies and the levels required to make certain claims that are specified in the proficiency structure. Assume, for example, that for a certification of computer network engineers there is a claim that such persons are adept at troubleshooting technical problems in network connectivity. It might be reasonable to expect that supporting this claim would require declarative knowledge (recall) of computer network hardware and their technical capabilities and interconnectivity protocols. It might also be reasonable to expect that supporting this claim requires an ability to employ a logical and efficient cognitive strategy to determine the cause of common network problems. Therefore, two proficiencies that might be implied by such a claim could include "hardware connectivity knowledge" and "strategic troubleshooting." These proficiency variables are inherently latent (not directly observable) and are therefore the target of the inference process of the assessment. These various proficiencies of interest are represented symbolically in Figure 1 by the set of circles and arrows in the Proficiency Model section. The circles represent various proficiency variables of interest and the arrows reflect known relationships between proficiencies (e.g., correlations or prerequisite relationships) and conditional independence relationships between variables.

The specification of claims and the description of proficiencies that one must possess to support these claims are related to traditional approaches to professional domain analysis. Often this is conducted through traditional job analyses, or in the case of assessments emphasizing strategic problem-solving, cognitive task analysis (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

Evidence Models (Conceptual)

Operationally, the evidence model specifies the manner in which observations during assessments are used to update estimates of ability. However, during the initial phases of assessment design, the evidence models are specified from a purely optimal domain perspective in order to drive task model development. This conceptual specification begins by imagining that there are no constraints or limitations to the ability to observe and track behaviors in naturalistic settings for a domain of interest. The task is to specify the situations and observable behaviors that are most revealing in terms of distinguishing among levels of ability in the proficiency model. The specification of what these crucial

---

[1] Some organizations also define the unsuccessful examinee as a user of results when diagnostic information is provided in order to guide further study.

situations are and what important behaviors can be observed will drive both the evidence model development for scoring and the specification of task models (discussed below). We will revisit the Evidence Model from the scoring perspective after the section on Task Models below.

Task Models

Task models are detailed descriptions of families of tasks with similar characteristics. These task models establish the framework, or the blueprint, for producing tasks (or items) that address particular targeted areas of the overall test blueprint. The conceptual evidence model helps to specify the characteristics of these task models that best distinguish among levels of ability. The task models, as pictured in Figure 1, consist of several variable elements: task design *features* (symbolized by the set of drop-down menu variables in the lower portion of the task model figure); *presentation material* (symbolized by the video screen icon in the upper right portion of the task model figure); and *work products* (symbolized by the jumble of shapes in the upper left of the task model figure). Task features describe the intent, construction, and associated design elements and options for a task. Presentation material defines what is presented to an examinee as part of a particular task (e.g., any graphics, any text, a question prompt, options to select from among, etc.). The work products are the resultant examinee data captured as a result of the examinee's interaction with the task, regardless of whether that data is directly used in scoring or not.

As an example of a portion of a task model, assume that for a test of basic math there was a claim of "Can add two integers" and an associated proficiency called "basic addition" (both very fine-grained examples). An item model (one of many) targeting such an ability might have elements such as those that appear as Table 1.

Table 1: Example Portion of a Task Model

Task Model Variable	*Possible Values (implication)*
Ability target	basic addition, sum of two integers
Difficulty factors	• single-digit integers with single-digit outcome (easier) • single digit integers with two-digit outcome (moderate difficulty) • two-digit integers with two-digit outcome (harder)
Reading load	• none (easier) • few simple words (moderate) • word problem (harder)
Presentation material	• Equation form of a problem (easier) • Word problem embedding problem (harder)

Work product	• multiple-choice (easier)
	• free response (harder)
	  – show work (complex scoring)
	  – response only (simple scoring)

Note that along with specification of various aspects of the task, it also indicates how different potential specifications can be expected to impact the difficulty of the item. Figure 2 and Figure 3 present the presentation material (in this case assuming paper presentation) for two items, both of which could be produced from the task model excerpted above. Note that while each is consistent with the task model above, each has characteristics that would tend to make it more or less difficult for the examinee, as well as to deliver (particularly assuming computerized presentation material rather than paper) and to score. These design decisions have implications for how a set of tasks discriminates among different levels of ability targeted by the assessment.

Figure 2

$$4 + 5 = \_\_\_$$

a) 1
b) 5
c) 9
d) 45

Figure 3

If you place a box that is 12 inches high on top of another box that is 23 inches high, how high are the two boxes together? Show your work and write your answer below.

Another characteristic to note is that for both multiple-choice and the free-response task there is an implication for a need to extend the level of detail of the task model. For the multiple-choice item, this extension would address how the distracters are developed (note, for example, that option (a) in Figure 2 is the answer someone would obtain if they subtracted 4 from 5 instead of adding) and their order in presentation. For a free-response story problem task, any use of word problems operationally would require further specification of the permissible vocabulary, sentence structures, topics, and representation of actors in the text (as well as considering the impact of the potential confound of reading ability with pure math ability on the measurement of proficiency model variables).

Obviously, the work products for these two examples in Figures 2 and 3 differ in that the former consists of an indication of which option is selected, while the latter consists of an indicated response and the calculations the examinee executed to determine the answer. Part of the consideration of such work products includes the medium used to collect the response. The item in Figure 2 is almost equally viable in paper and computerized format, while for the item in Figure 3, it is more difficult to capture the work products in computerized administration than in paper-and-pencil administration.

The set of task models for an assessment design can be organized hierarchically to facilitate the degree to which the test developer must exercise control over the types of tasks used. For example, the math test item of the general form:

{single-digit integer} + {single-digit integer} = [single-digit integer]

is a sub-category of the more general form:

{integer} {operation} {integer} = [integer]

Depending on the degree of control which must be exercised in test authoring (based on the test blueprint) and the intent of the item usage, a hierarchy of task models can be developed with varying degrees of specificity of the model design. For example, in cases where the prediction of the specific difficulty of an item is important, the test designer may wish to exercise a relatively high degree of control. This is often the case in efforts that use task modeling as the basis for automatic item generation (Embretson, 1998; Embretson, 1999; Williamson, Johnson, Sinharay, & Bejar, 2002; Newstead, Bradon, Handley, Evans, & Dennis, 2002), in which a computer generates items according to a given task model with no human intervention.

Evidence Models (Scoring)

Evidence Models specify how the evidence contained in task data informs belief about Proficiency Model variables. Evidence Models for scoring rely on the Proficiency Model as the fixed target for inferences and the Task Models and tasks authored from them, as the mechanism for producing data to be used in scoring. The Evidence Model for scoring, as presented in Figure 1, consists of two subcomponents:

- *evidence rules* – determine what elements of the task performance constitute evidence and summarizes their values

- *statistical model* – aggregates evidence to update estimates of ability in the proficiency model

Evidence rules transform elements of the work product (the record of examinee task performance) into observables; summary representations of work used by the statistical model to update estimates of proficiency. This process is called *evidence identification* in ECD terminology. In Figure 1, this process is represented in the Evidence Rules portion of the Evidence Models diagram. This figure illustrates how work products from the Task Model are parsed to produce observables, symbolized by the three squares (for three observables). With most multiple-choice questions this process seems almost trivial. For each question there is only one observable, the value of which is determined by comparing the response indicated by the examinee with a predetermined key and representing the observable as a simple 1, for correct, when the response is the same as the key, or 0, for incorrect, when the response does not match the key. In other situations the determination of observables requires more effort, such as for the task presented in Figure 3. Some evidence rules would be required to establish both the correctness of the final answer and for representing the degree of adequacy of shown work in computing the answer. Note that the determination of the value of observable variables also implies using some elements of the work product and ignoring other elements. For example, most scoring of multiple-choice items ignores the particular choice the examinee made if the choice was incorrect, while others might infer the nature of misunderstandings examinees may have when they select particular incorrect answers. Also, in computerized testing environments it is common to collect information on the amount of elapsed time an examinee took to respond to a question despite the fact that this is seldom used in the evidence rules for scoring. The representation in Figure 1 implies three observables obtained from this particular work product.

The statistical model portion of the Evidence Model uses the values of observables to update estimates of ability. In number-right scoring this statistical model is a simple summation function in which the prior value plus the value of the observable (1 or 0) equals the new value. In models using item response theory (IRT) this updating is controlled by the parameters associated with the item for which the observable is being used as evidence and by the fundamental statistical relationships for updating ability estimates from observations under the IRT model being applied. In most common applications (e.g., number right, IRT, etc.) there is a single proficiency variable for ability and a single observable variable from each item. In Figure 1, however, we illustrate the case where three observables are produced from an item and these are used to update two proficiency variables. These two proficiency variables, in turn, represent two of the five proficiency variables that make up the Proficiency Model.

Such a representation illustrates the value of such models for more complex assessment designs, such as for computerized simulations that use automated scoring, while still representing the fundamental structure and critical models for design of traditional assessments.

Summary of ECD Model Interactions

In review, the ECD process provides a framework for assessment design that emphasizes a systematic consideration of multiple models for design and their interaction. These begin with the fundamentals of assessment purpose (specification of populations being served, decisions being made, known assessment constraints, etc.) from which formal claims are developed. These claims drive the specification of a Proficiency Model. The implications of the Proficiency Model and claims in combination drive the evidential needs of the assessment, formally represented as the Evidence Model. These needs are actualized in the design of assessment tasks, the blueprints for which are expressed as Task Models.

Once tasks from these models are developed and fielded, the scoring process is essentially a reversal of the development process. The administered tasks result in work products with pre-established properties. These work products are parsed according to the evidence rules of the Evidence Model to produce observables. The statistical model of the Evidence Models is applied to draw inferences about proficiencies on the basis of these observables. Finally, the ultimate values of proficiency variables establish what assessment claims can be supported on the basis of the assessment. These reported claims, in turn, are used by the consumers of score reports to make informed decisions.

Conclusion

This work has presented the basic concepts of ECD and made an argument for the relevance and value of such an approach for any assessment design process, whether for a paper-and-pencil assessment using multiple-choice tasks or a computerized assessment using complex simulations and automated scoring. It is hoped that through wide adoption of such a process, the process of assessment design can be improved, both by formalizing processes that good assessment designers perform implicitly, and by encouraging consideration of issues not previously addressed in formal assessment design. It is also hoped that such resultant design rationales strengthen the quality and the validity arguments for use of such measures for their intended purpose.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, ι*(5). Available from http://www.jtla.org.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods*, 3 (3), 380 – 396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407–433.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439-483.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

Mislevy, R.J., Steinberg, L.S., Breyer, F. J., Almond, R.G. & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and human behavior*, 15, 335-374.

Newstead, S.E., Bradon, P., Handley, S., Evans, J., and Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In Kyllonen, P. and Irvine, S.H. (Eds.) *Item Generation for Test Development.* Lawrence Erlbaum Associates: Mahwah, NJ.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Wigmore, J.H. (1937). *The science of judicial proof* (3rd Ed.). Boston: Little, Brown, & Co.

Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. (2002, April). *Applying Hierarchical model calibration to automatically generated items.* Paper presented at the American Educational Research Association, New Orleans, LA.

Update on Evaluation of NAEP Achievement Levels Procurement

Objective   To receive a brief informational update from NCES on the current status of the procurement being planned to evaluate NAEP achievement levels. Ongoing updates will be provided at each COSDAM meeting.

Background

The NAEP legislation states:

> The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public.

In providing further detail, the aforementioned subsection (f) outlines:

> (1) REVIEW-
>
> A. IN GENERAL- The Secretary shall provide for continuing review of any assessment authorized under this section, and student achievement levels, by one or more professional assessment evaluation organizations.
>
> B. ISSUES ADDRESSED- Such continuing review shall address--
>
> (i)   whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
>
> (ii)   whether student achievement levels are reasonable, valid, reliable, and informative to the public;-
>
> (iii) whether any authorized assessment is being administered as a random sample and is reporting the trends in academic achievement in a valid and reliable manner in the subject areas being assessed;
>
> (iv) whether any of the test questions are biased, as described in section 302(e)(4); and
>
> (v)   whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.
>
> (2) REPORT- The Secretary shall report to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health,

Education, Labor, and Pensions of the Senate, the President, and the Nation on the findings and recommendations of such reviews.

(3) USE OF FINDINGS AND RECOMMENDATIONS- The Commissioner for Education Statistics and the National Assessment Governing Board shall consider the findings and recommendations of such reviews in designing the competition to select the organization, or organizations, through which the Commissioner for Education Statistics carries out the National Assessment.

Responsively, a procurement has been planned to administer an evaluation of NAEP achievement levels. The last update COSDAM reviewed on this topic was in May 2013.

In this brief written update, NCES provides the Committee with a summary of the status of this procurement.

Evaluation of NAEP Achievement Levels

The National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES) will administer the Evaluation of the NAEP Achievement Levels. The Department's Contracts and Acquisitions Management office posted a Request for Information (RFI) on FedBizOpps.gov on May 22, 2013. We anticipate that the Department will issue a Request for Proposals (RFP) this summer, with an award announced later this fall.

NAEP 12th Grade Preparedness Research

Based on the Program of Preparedness Research adopted by the Governing Board in March 2009, four categories of research studies were conducted to produce evidence to develop and support the validity of statements for NAEP reporting on the academic preparedness in reading and mathematics of 12th grade students for college and job training.

- content alignment studies;
- statistical relationship studies;
- judgmental standard setting studies; and
- surveys

Additionally, the Texas Commissioner of Higher Education offered the opportunity to conduct a benchmarking study with Texas higher education institutions, and a pilot study to examine the feasibility was conducted.

The research studies completed to date are available in an online technical report. In addition, the NAEP 12th Grade Preparedness Commission conducted a symposium in Washington, DC on July 9, 2013 focused on the Board's preparedness research results and the Phase 2 research plans.

The following informational attachments are provided:

Additionally, the following attachments are provided for reference:

Attachment E-1
College Course Content Analysis Progress Update

In September 2012, the Governing Board awarded a contract to the Education Policy Improvement Center (EPIC) to conduct research on entry level non-remedial college course content in order to (1) identify the prerequisite knowledge and skills in reading and mathematics for entry-level college courses and (2) determine the extent to which there is a match with the content of grade 12 NAEP reading and mathematics assessments. This project addresses academic preparedness for college only—a separate parallel research project addresses preparedness for job training (described below).

In this project, EPIC will determine the entry-level (introductory) credit-bearing courses most frequently taken by entering students that are reflective of college-level reading and mathematics demands and that satisfy general education requirements. These introductory courses should have no college-level prerequisite course requirements, and only non-remedial courses that satisfy general education requirements should be included in the analysis. Further, in cases where multiple versions of a course are offered for majors and non-majors, only the course for non-majors should be included.

Using course artifacts for a generally representative sample of institutions, EPIC will analyze the introductory course artifacts for commonalities and differences in the reading and mathematics prerequisites needed to qualify for placement into the course. From these analyses, EPIC will develop descriptions of the knowledge, skills, and abilities (i.e., the prerequisite KSAs) needed for students to qualify for placement into the introductory course, based on an analysis of the course artifacts. And as part of a set of comparative analyses, EPIC will then use these descriptions to review:
- the description of minimal requirements for placement into college-level coursework as developed in the NAEP preparedness judgmental standard setting (JSS) research
- KSAs represented by 2009 grade 12 items that map to the NAEP scale with a response probability of .67 and fall within the range of cut scores set by the two replicate panels in the JSS research
- 2009 and 2013 grade 12 NAEP items
- the KSAs represented by 2009 items that map in the range of the NAEP score scale from the the Basic level through the Proficient level; and
- the NAEP achievement level descriptions.

A new progress report is attached with more details on the project and a description of work completed to date.

College Course Content Analysis Study for NAEP Preparedness Research
Progress Update
Submitted by
Educational Policy Improvement Center (EPIC)

INTRODUCTION AND BACKGROUND

The College Course Content Analysis (CCCA) study is one of a series of studies contributing to National Assessment of Educational Progress' (NAEP) Program of 12th Grade Preparedness Research conducted by the National Assessment Governing Board (NAGB). The purpose of the CCCA study is to identify a comprehensive list of the reading and mathematics knowledge, skills, and abilities (KSAs) that are pre-requisite to entry-level college mathematics courses and courses that require college level reading based on information from a representative sample of U.S. colleges. The Educational Policy Improvement Center (EPIC) is the contractor working for the Board to conduct this study.

Another goal of the CCCA study is to extend the work of the two previous preparedness studies—the Judgmental Standards Setting (JSS)[1] study, implemented in 2011 and the Job Training Program Curriculum (JTPC) study, implemented in 2012. The CCCA study is designed so the results can be compared to the JSS and JTPC studies, reporting on how this new information confirms or extends interpretations of those earlier studies. The design of the CCCA study is based on the JTPC study but with modifications based on the lessons learned.

The CCCA study will answer four core research questions.

1. What are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements?
2. How do these prerequisite KSAs compare with the 2009 and 2013 NAEP reading and mathematics frameworks and item pools?
3. How do these prerequisite KSAs compare with previous NAEP preparedness research (i.e., the descriptions of minimal academic preparedness requirements produced in the JSS research)?
4. How can these prerequisites inform future NAEP preparedness research (i.e., planning and analysis efforts relative to the 2013 grade 12 NAEP reading and mathematics assessments)?

---

1 National Assessment Governing Board. (2010). *Work Statement for Judgmental Standard Setting Workshops for the 2009 Grade 12 Reading and Mathematics National Assessment of Educational Progress to Reference Academic Preparedness for College Course Placement.* (Higher Education Solicitation number ED-R-10-0005).

The final report is due May 2014, and until then COSDAM will receive detailed reports at each Board meeting.

METHODOLOGY

The Design Document for the CCCA study is complete. It provides guidance for the study by describing:

- Criteria for collecting courses and artifacts;
- A sampling plan to comprise a representative sample of institutions;
- Review and rating processes, including a training plan and process for ensuring reviewer effectiveness and consistency; and
- The process for ensuring reliability across reviewers providing artifact analysis.

This study comprises three primary phases:

1. Identification and collection of course artifacts,
2. Review of course artifacts by Review Teams, and
3. Analysis and reporting.

The first phase of the study is complete. The course artifacts have been identified, all artifacts have been collected, review packets have been created from those artifacts, and all of the data collection surveys are programmed and ready for use. The second and third phases of the study have begun.

Most notably NAEP Advisory Panels, in both reading and mathematics, were conducted in June of 2013 and the guidance from those panels is being integrated into the implementation of the next phases of the study. Content reviewer training sessions were conducted in early July and the independent content reviews will occur during July 2013 and August 2013. Preparation for data analysis and final reporting has also begun.

OVERVIEW OF ACTIVITIES BY PHASE
Phase 1: Identification and collection of course artifacts

In the CCCA study, a *course artifact* is defined as a syllabus, a non-textbook based assignment or assessment, and textbook excerpt. In mathematics, there are some instances where the only specifically identified assignments were listed in the syllabus and were from the textbook. In those cases, a textbook based assignment or assessment was allowed. The CCCA sample of artifacts is derived from extant artifacts and combined with newly gathered course artifacts. Extant artifacts contributing to the CCCA sample were extracted from EPIC's repository of extant artifacts compiled during previous research on entry-level curricula at postsecondary educational institutions. Project staff has solicited new course artifacts as needed to create a complete and representative sample.

EPIC identified a set of inclusion criteria that courses must meet to be included in the CCCA study as well as a set of institutional characteristics of which the final CCCA Artifact Bank must be representative. The final CCCA Artifact Bank will comprise a set of courses and artifacts that will be used as the basis for the content review.

At the conclusion of artifact collection, the CCCA Artifact Bank will include all relevant artifacts compiled into course packets to be reviewed by mathematics and reading content review teams in the second phase of the study.

Phase 1 preparatory work also included the convening of NAEP Advisory Panels, for reading and mathematics respectively, to obtain content-based guidance and recommendations. In these meetings, preliminary coding schemas, training materials and decision rules were reviewed. NAEP advisors also reviewed all of the course packets that will serve in validation data analyses, training sessions, and determining sufficient reviewer competence (qualifying). Guidance from this NAEP Advisory Panels is being integrated into the implementation of the study.

Phase 2: Review of course artifacts by Review Teams

In Phase 2, content reviewers are recruited and training materials are developed. Content reviewers will first be trained to review the course packets from a holistic perspective and identify prerequisite mathematics and reading KSAs. In the second independent review training, the NAEP frameworks for grade 12 reading and mathematics will be used as a basis for coding the packets. If additional KSAs are identified during either review sessions, the new KSAs will be documented and included in all successive reviews, comparisons and data analyses. The overarching goal of the CCCA study is to identify all prerequisite KSAs, not just those KSAs associated with the NAEP frameworks.

The CCCA design has embedded validity checks within the process to evaluate the reliability of the review team coding. Two validation packets were created for each of the four course titles in reading and mathematics. The validation packets look like any other course packet and will be mixed in with the others during the independent and group reviews. The content reviewers will not know which packets are the validation packets. The NAEP experts have coded the validation packets and their coding will serve as a reference for determining how well the content reviewers are coding. The percent agreement between the four review teams' group consensus coding on the validation packets and the reference coding as reliability evidence will be calculated. Project staff will also report the agreement of group consensus coding by the four review teams within each course title. The agreement statistic will be calculated using the same method.

The CCCA study's methodology combines independent individual judgments with panel consensus processes. The first independent review is focused on applying the conceptual

understanding of the mathematics or reading knowledge and skills required in entry-level college courses by content reviewers with experience in teaching these types of courses and training and experience in the EPIC methodology of coding artifacts. The goal of the second, or group, review is to focus on adjudicating differences in coding of the packets completed during the independent review and the confirming the identification of exclusions in the NAEP framework objective statements. An additional focus is to review all KSAs that were identified in the packets but were not found in the NAEP frameworks.

The final result of this two-part review process will be a comprehensive list of prerequisite KSAs, answering the Board's research question: what are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements.

Finally, a review is conducted by NAEP content experts to address the remaining research questions.

Phase 3: Analysis and reporting

Phase 3 includes processing and analyzing the judgments collected during the review of course artifacts by review teams, and preparing the data to be reported in ways that are directly responsive to research questions in accordance with the analysis plan specified within the Design Document. Standard statistical methods and metrics necessary will be employed to monitor and demonstrate validity and reliability, and both conceptual (information processing/document analysis) and technical (quantitative) analyses will be conducted. The CCCA study is structured to provide a fully crossed, three factor design to ensure that results can be reviewed in statistical generalizability analyses, which will allow us to evaluate the reliability of the study design.

Final results will include narrative summaries of the prerequisite knowledge, skills, and abilities in mathematics and reading. Summary analyses will also address all aspects of the CCCA study design (see Illustration 1). As project elements are completed, the appropriate sections of Illustration 1 are shaded in dark gray. Project elements that have begun and are in progress are shaded in a lighter gray. Those project elements that have just begun have no shading in the diagram.

Illustration 1: Project Design
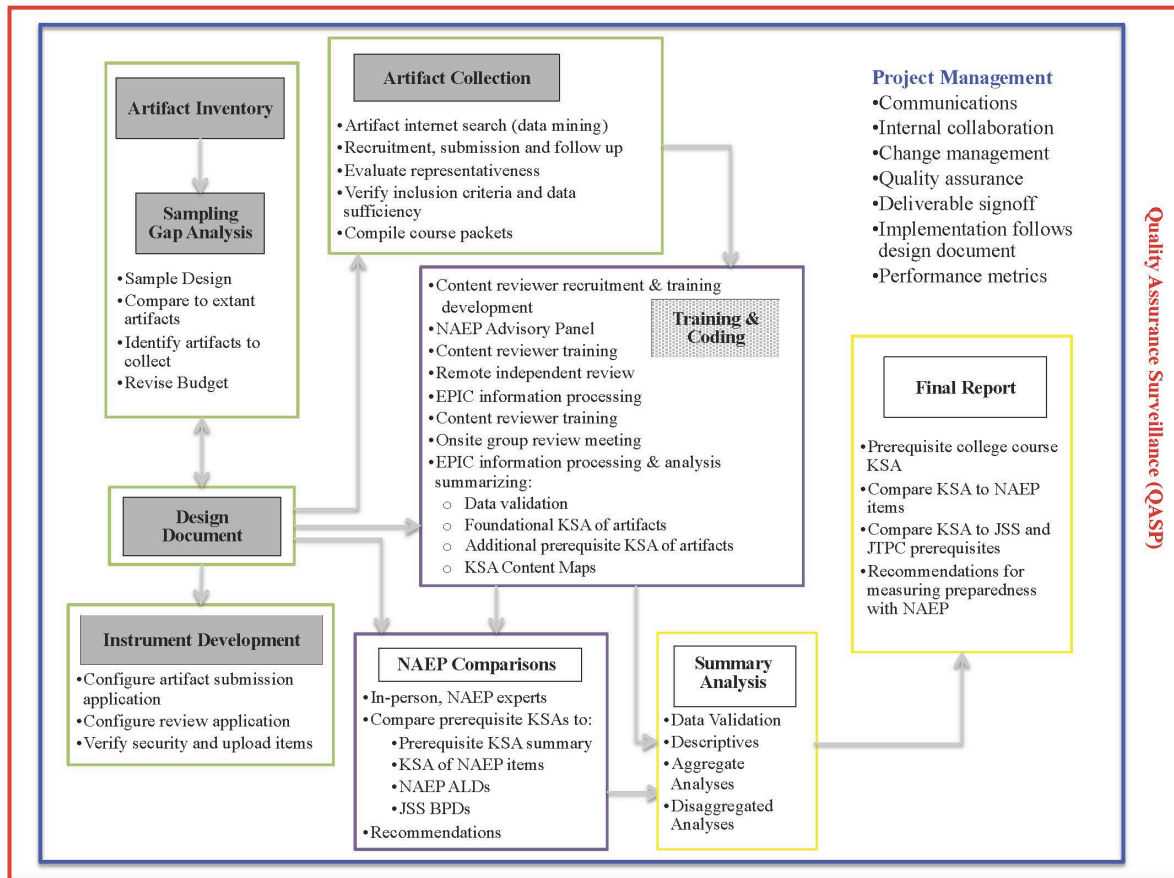
Illustration 2 displays an updated schedule of the CCCA study. As a result of feedback from the NAEP Advisory Panels, the schedule for content reviewer training has been changed to accommodate two sessions of training: an orientation session focusing on a holistic review of the packets, and a second training session after the reviewers are familiar with the packets. That second training will address the coding scheme, decision rules and use of the NAEP frameworks. Completed events are shaded black. Upcoming events and changes to the schedule are shaded.

Illustration 2: CCCA Study Gantt chart

OR EVENT	Start Date	End Date	Duration	QUARTER 2 APR MAY JUN	QUARTER 3 JUL AUG SEP	QUARTER 4 OCT NOV DEC	QUARTER 1 JAN FEB MAR	FINAL APR
NAEP Technical Panel Meeting	21-Jun	23-Jun	COMPLETE					
NAEP Technical Panel Meeting	5-Jun	9-Jun	COMPLETE					
Facilitator Training	8-Jul	12-Jul	3 days					
Math Content Holistic Training	8-Jul	12-Jul	3 days					
Reading Content Holistic Training	8-Jul	12-Jul	12 days					
Math Content NAEP Training	22-Jul	24-Jul	ADDED					
Reading Content NAEP Training	22-Jul	24-Jul	ADDED					
Independent Content Reviews	9-Jul	16-Aug	6+ weeks					
EPIC Data Analysis Period 1	19-Aug	6-Sep	1.5 week					
Math Content Review Meeting		22-Sep	1+ week					
Read Content Review Meeting		22-Sep	1+ week					
EPIC Data Analysis Period 2	22-Sep	18-Oct	4+ weeks					
NAEP Expert Math Review Meeting	21-Oct	25-Oct	4 days					
NAEP Expert Read Review Meeting	11-Nov	15-Nov	4 days					
EPIC Data Analysis Period 3	18-Nov	31-Jan	10.5 weeks					
Final Report Writing and Review	3-Feb	28-Mar	7.5 weeks					
Board Review and Presentation	1-Apr	30-Apr	4+ weeks					
Final Report Deliverable Due	30-Apr	30-Apr	Final Deliver					

PROGRESS UPDATE

Identification and Collection of Course Artifacts (Phase 1)

Table 1 contains the finalized list of entry-level courses to be included in the CCCA study.

Table 1: Course Titles Included in the CCCA Study

Mathematics	Reading
College algebra	English literature
Finite math	Introduction to psychology
Introduction to calculus/precalculus	U.S. government/Introduction to political science
Statistics	U.S. history

The criteria for the course titles is as specified in the Design Document are:

- Entry-level
- Credit-bearing
- Frequently taken
- No college level prerequisites
- Not honors level
- Not remedial
- Not for majors

For each course in the study, a course packet will be considered complete if it includes the following:
- Syllabus
- Textbook excerpt
- Textbook table of content
- Non-textbook based assignment or exam from the first third of the course

Collecting mathematics course artifacts has been challenging, particularly identifying calculus courses that do not require prerequisites. In order to meet the requirement of collecting artifacts for 24 course packets per course type, we have slightly relaxed some of our representativeness criteria. One end result is that sufficient mathematics packets will be collected but larger institutions will slightly overrepresented.

Another factor identified in the artifact collection effort is that mathematics courses often use assignments from the textbook and do not create assignments outside of the textbook. To address this issue, the criteria for a complete course packet has been relaxed to allow an assignment to be textbook-based when the assignment is specifically identified in the syllabus. This change was also supported by the NAEP content advisors and has resulted in an improvement in the overall math packet quality.

Collecting artifacts for English Literature packets has also been challenging than expected due to the common college requirement that student take a writing composition class prior to taking English classes.

Tables 2 and 3 are summaries of the characteristics of representative institutions where courses have been submitted as candidates for packet creation. All courses met the criteria for inclusion in the study and the packets are sufficiently data-rich. These percentages should be considered preliminary, as the set of packets to be used in the study has not been completely finalized. More packets than are needed for the study have been collected in order to have sufficient overage and be able to make substitutions, if necessary. Note the number of completed packets in the "*N*-count" in the headers of the tables.

Table 2: Updated Institutional Characteristics of Sample for Mathematics

Characteristic	College algebra (*N* = 20)	Finite math (*N* = 16)	Introduction to calculus (*N* = 20)	Statistics (*N* = 20)	Mathematics Overall (*N* = 76)
Program type					
2-Year	40%	25%	15%	25%	26%
4-Year	60%	75%	85%	75%	74%
Size					
Small	75%	44%	40%	20%	54%
Medium	10%	13%	20%	25%	17%
Large	15%	44%	40%	55%	29%
Control					
Public	60%	56%	65%	50%	58%
Private not-for-profit	40%	44%	35%	50%	42%
Geographic Region					
West	10%	6%	15%	15%	12%
Midwest	30%	13%	20%	30%	24%
East	10%	19%	10%	25%	16%
Southeast	30%	25%	30%	20%	26%
Southwest	20%	38%	25%	10%	22%

Table 3: Updated Institutional Characteristics of Sample for Reading

Characteristic	English literature (N = 17)	Introduction to psychology (N = 20)	U.S. government (N = 20)	U.S. history (N = 20)	Reading Overall (N = 80)
Program type					
2-Year	25%	35%	15%	15%	23%
4-Year	75%	65%	85%	85%	78%
Size					
Small	55%	60%	50%	55%	55%
Medium	15%	20%	25%	25%	21%
Large	30%	20%	25%	20%	24%
Control					
Public	60%	60%	50%	50%	55%
Private not-for-profit	40%	40%	50%	50%	45%
Geographic Region					
West	0%	20%	20%	15%	21%
Midwest	40%	25%	30%	25%	30%
East	15%	20%	20%	25%	20%
Southeast	15%	30%	20%	25%	23%
Southwest	30%	5%	10%	10%	6%

NAEP Advisory Panel Review of Course Artifacts (Phase 1)

A preliminary review of packets was completed at two NAEP Advisory Panels, one in reading and one in mathematics, both conducted in June of 2013. The table below provides an overview of the type and number of packets to be reviewed at each CCCA review session.

Table 2: Allocation of Packets Across CCCA Events

Total number of course packets and purpose	NAEP Advisory Panel Pre-coded by NAEP experts	Training Coded by content reviewers and alternate reviewers	Independent Review Coded by content reviewers and alternate reviewers	Group Content Review Reviewed in content teams
4 Training packets	4 Total, 2 for math and 2 for reading	4 Total, 2 for math and 2 for reading	NA	NA
4 Qualifying packets	4 Total, 2 for math and 2 for reading	4 Total, 2 for math and 2 for reading	NA	NA
16 Validation packets	16 Total, 8 for math and 8 for reading	Not coded in training	16 Total, 8 for math and 8 for reading	Depends on the number of packets that need to be reviewed during the group review process
160 Operational packets	Not pre-coded by NAEP experts	Not coded in training	160 Operational packets	Depends on the number of packets that need to be reviewed during the group review process
TOTAL PACKETS				
184 Total 92 for math 92 for reading	24 Total 12 for math 12 for reading	8 Total 4 for math 4 for reading	176 Total 88 for math 88 for reading	

The primary goals of the NAEP Advisory Panel meetings were: (1) to obtain comprehensive guidance regarding the process of content review training; and (2) to establish benchmarks for coding a subset of the packets. Twelve (12) packets were reviewed in the NAEP Advisory meeting – 8 validation packets, 2 training packets, and 2 qualifying packets for both reading and mathematics. The NAEP experts were briefed on the proposed training process and provided with an initial set of training materials that included a coding schema, decision rules and a reference sheet. The request was that they begin coding the 12 packets using the existing guidance. As they coded, they provided feedback and suggested new or different approaches regarding the process and decision rules.

The outcomes from each of the Advisory Panel meetings were:
1. Guidance for making improvements to the training process
   a. Training in two separate sessions – the first for a holistic review of the entire packet without reference to the NAEP frameworks and then a second session to train on how to review using the NAEP frameworks
   b. Training in small groups of four instead of a large group
   c. Deeper understanding of the complexity of the task and advice on how to make it easier for the reviewers
2. Updated, and simplified, coding schema
3. Updated, and simplified, decision rules
4. Updated, and simplified, reference sheets
5. Guidance for establishing the criteria for sufficiency of a packet
6. Guidance for establishing a review procedure for course packets
7. Guidance for identifying the criteria for qualities of good training examples
8. Guidance for estimating time to complete review tasks independently and in group
9. Benchmarks for 12 packets and recommended usage for one of three possible uses – validation, training and/or qualifying.

Most of the guidance from the NAEP experts was integrated into the development of the training materials and the planned implementation of the CCCA study. In order for guidance to be accepted, it had to be feasible, not be in conflict with other design factors in the Design Document and support comparability with the JSS and JTPCS study findings.

Coding Schema and Review of the Course Artifacts (Phase 2)

Using the Design Document as the roadmap, phase 2 activities are well underway.

Development of the content reviewer coding instruments for both mathematics and reading, based on the coding schema and decision rules in the design document specifications, was completed. Both the coding schema and decision rules were thoroughly reviewed by the NAEP Advisory Panel experts, in conjunction with the review of the sample course title packets. As a result of that work, the recommendation was to simplify the coding schema from six levels to three levels. The coding instrument has been updated to reflect the three level coding schemas. The change from six levels of coding to three levels will not impair comparability with JSS or JTPCS findings.

Table 4: Simplified Coding Schema

	Design Document Coding Schema	**Post NAEP Advisory Panel Coding Schema**
Applicability and Importance	1—KSA is **NOT** applicable to this course	1—KSA is **NOT PREREQUISITE** to this course.
	2—KSA is **NEW** content taught in this course	
	3—KSA is **PREREQUISITE** for this course and is **NOT IMPORTANT.** Although a prerequisite, possessing this KSA will make little difference on course outcomes.	2—KSA is **PREREQUISITE** for this course.
	4—KSA is **PREREQUISITE** for this course and is **MINIMALLY IMPORTANT.** This KSA is a prerequisite, which if possessed, is likely to result in better course outcomes.	
	5—KSA is **PREREQUISITE** for this course and is **IMPORTANT.** Without this KSA, students will struggle with the course.	3—KSA is **PREREQUISITE** for this course and is **IMPORTANT.**
	6—KSA is **PREREQUISITE** for this course and is **VERY IMPORTANT.** Without this KSA students are not prepared for and will be unlikely to complete this course.	

While the task of reviewing mathematics packets is very different from the reading task, it was clear to both panels that the task of content review needed to be simplified to assure manageability of the task and a reasonable time commitment for review activities. Changes are not expected to have any negative effect on the content maps that will eventually be created for review by the NAEP experts at the NAEP comparison meetings in Fall 2013. Both panels endorsed the practice of reviewing the entire packet in a holistic manner to get familiar with the packet, the task, and the overall variation among the packets with the goal of identifying the most relevant prerequisite KSAs in a summary manner. This approach ensures that the content reviewers have the big picture in mind and that the potential risk of bias toward only identifying NAEP KSAs is mitigated.

The overarching guidance from both of the NAEP Advisory Panels was that the content reviewer task of reviewing 28 packets was challenging because of complexity and time-consuming because of the number of packets to be reviewed. The advice was to review the entire training and review process with the goal of simplifying wherever possible.

Several additional changes, and improvements are being considered. The estimated time needed to review each individual packet has increased (informed by the NAEP content advisors, and so an increase in stipend and an increase in the length of the group review session by half a day are under discussion.

Design Document and Analysis and Reporting (Phase 3)

Preparation for analysis and the final reporting have begun with the majority of the effort in data management. Staff are working with sample data and testing to ensure that accurate data collection protocols and routines of effective quality control, data cleaning procedures and data storage/security protocols are in place and use.

The final report is also underway. The table of contents has been established and preliminary table shells have been drafted. The final report will be written in sections and reviewed throughout the rest of the study.

Attachment E-2
Job Training Program Content Analysis Final Report

In October 2011, the Governing Board began work with WestEd and its subcontractor, the Education Policy Improvement Center (EPIC), to conduct follow-up research relative to the NAEP preparedness judgmental standard setting (JSS) research, wherein panelists reviewed NAEP questions and made judgments about the content knowledge needed by minimally prepared students. The research results from this project are intended to supplement the JSS research findings by providing a clearer understanding of the knowledge and skills required for entry- and exit-level coursework in designated occupational programs. By reviewing course artifacts such as syllabi, text books, and assignments, this study will help to determine if the knowledge, skills, and abilities (KSAs) required of students in the training programs are appropriately represented by the borderline preparedness descriptions (developed in the JSS research), by all the items on the 2009 NAEP, and by the 2009 NAEP items in the scale score ranges identified by panelists in the JSS research project.

The executive summary for the final report was included with the May 2013 COSDAM materials. The full report is now available online at:

http://www.nagb.org/what-we-do/preparedness-research/types-of-research/jss.html

Attachment E-3
Phase 2 Academic Preparedness Research Plans

Continued research plans call for NAEP-SAT, NAEP-ACT, and NAEP-EXPLORE statistical linking studies, more research partnerships with states, analysis of course content prerequisites for job training programs and freshman college courses, and efforts to partner with experts in military occupational training. A summary of each proposed research study follows. At the November 2012 Board meeting, COSDAM began discussion on these research plans.

National and State Statistical Linking Studies with the SAT and with the ACT

In 2013, the Governing Board will partner again with the College Board, as it did in 2009, to conduct a statistical linking study at the national level between NAEP and the SAT in reading and mathematics. Through a procedure that protects student confidentiality, the SAT records of 12th grade NAEP test takers in 2013 will be matched, and through this match, the linking will be performed. A similar study at the national level is planned in partnership with ACT, Inc.

In addition, the state-level studies, begun in 2009 with Florida, will be expanded in 2013. Again using a procedure that protects student confidentiality, the postsecondary activities of NAEP 12th grade test takers in the state samples in partner states will be followed for up to five years using the state longitudinal data bases. Five states will be partners in these studies: Florida, Illinois, Massachusetts, Michigan, and Tennessee. These studies will examine the relationship between 12th grade NAEP scores and GPA, placement into remedial versus credit-bearing courses, and scores on admissions and placement tests. Data sharing agreements are in development for each state partner.

August 2013 Update: No updates at this time.

In 2013, linking studies between 8th grade NAEP in reading and mathematics and 8th grade EXPLORE, a test developed by ACT, Inc. that is linked to performance on the ACT, are planned with partners in two states, KY and TN. The objective is to determine the feasibility of identifying the point on the NAEP scales that indicate students are "on track" for being academically prepared for college and job training by 12th grade. As a foundation for the linking study, content alignment studies between 8th grade NAEP reading and mathematics and 8th grade EXPLORE would also be conducted as a part of the planned partnership with Act, Inc.

August 2013 Update: No updates at this time.

The Governing Board is conducting a procurement (1) to design a comprehensive and multi-method evaluation of the grade 12 NAEP frameworks and item pools in both reading and mathematics as measures of academic preparedness for college and job training; and (2)

based on the evaluation, to produce specific recommendations for changes that may be needed to further refine 12th grade NAEP in reading and mathematics as a measure of academic preparedness for college and to determine the extent to which changes would be needed to make 12th grade NAEP in reading and mathematics a valid measure of academic preparedness for entry into job training programs that require at least three months of post-secondary training, but not a bachelor's degree in college.

The review of the 12th grade reading and mathematics frameworks by Achieve, Inc. in 2005 and 2006 led to changes in the frameworks for the 2009 assessments intended to measure 12th grade academic preparedness for college and job training. The content alignment studies between 12th grade NAEP reading and mathematics and the SAT and ACT college admissions tests in reading and mathematics tests found a high degree of overlap in content widely recognized as representing academic preparedness for college. The content alignment study with WorkKeys, as well as the Judgmental Standard Setting studies for job training, surfaced questions about the capacity of the current 12th grade NAEP to measure academic preparedness for job training. The planned evaluation is part of the continuing program of preparedness validity research.

In this procurement, the Board seeks innovative, practicable design proposals for evaluations that will provide the foundation needed to make valid statements about academic preparedness.

August 2013 Update: The contract has been awarded to HumRRO. A kickoff meeting has been conducted. The project is now just getting underway.

███

Reporting on academic preparedness for college and job training is a challenging and important new direction for NAEP. Hence, the Governing Board is also conducting a procurement to seek proposals for research designs and studies that are feasible. The objective of the research is to advance the Governing Board's efforts to identify locations on the 12th grade NAEP reading and mathematics scales that represent the knowledge and skills to qualify for training in various occupations.

August 2013 Update: The procurement process did not result in a contract award.

Attachment E-4
Overview of the Types of NAEP Preparedness Research

As part of the ongoing updates to COSDAM, the following is a summary of each research study category from phase 1 of the Board's program of research for reporting academic preparedness.

Content Alignment Studies

Content alignment studies are a foundation for the trail of evidence needed for establishing the validity of preparedness reporting, and are, therefore, considered a high priority in the Governing Board's Program of Preparedness Research. The alignment studies will inform the interpretations of preparedness research findings from statistical relationship studies and help to shape the statements that can be made about preparedness. Content alignment studies were recommended to evaluate the extent to which NAEP content overlaps with that of the other assessments to be used as indicators of preparedness in the research.

A design document was developed by Dr. Norman Webb for the NAEP preparedness research alignment studies, and this design was implemented for the studies of the 2009 NAEP with the SAT and ACUPLACER in reading and mathematics. This design, with minor modifications, has also been used for the alignment of the 2009 NAEP with WorkKeys tests in these subject areas.

Content alignment studies for the first phase of the Board's Program of Preparedness Research have been completed for NAEP in reading and in mathematics with WorkKeys, the SAT, and ACCUPLACER. In addition, a content alignment study was designed and conducted by ACT for the ACT and NAEP in reading and mathematics before the content alignment design document was developed.

Studies to Establish Statistical Relationships

Highest priority has generally been placed on these studies. Currently, two main sets of studies have been conducted under this heading. One set addresses *statistical linking* of NAEP with other assessments, and the other set examines *longitudinal data* for NAEP examinees.

For statistical linking, there has been a study to relate SAT scores in reading and in mathematics to the national sample of NAEP scores for grade 12. The objective was to provide a statistical linking of SAT and NAEP scores for all students in the 2009 grade 12 NAEP who had taken the SAT by June 2009. ETS staff reported that the match rate of approximately 33% of NAEP scores to SAT scores compares favorably to the national SAT participation rate of approximately 36% of public school students. The final sample used for linking the NAEP reading and SAT critical reading included approximately 16,200 students. For NAEP and SAT mathematics, the linking sample included approximately 15,300 students.

For longitudinal data, a series of analyses were conducted to examine statistical relationships for Florida's NAEP examinees. NAEP's 2009 state-representative sample of Florida 12th graders was used to match NAEP scores for reading and mathematics to student scores on several tests

collected by the Florida Department of Education (FLDOE). The data sharing agreement with FLDOE provides access to scores for the SAT, ACCUPLACER, and WorkKeys. Additionally, ACT, Inc. has given permission to the Florida Department of Education to share ACT scores with the Governing Board for purposes of conducting the grade 12 preparedness research. A plan to allow for electronic transfer of data was developed to keep secure the identity of students, consistent with the NAEP legislation, FLDOE requirements, and requirements of each assessment program.

Records for roughly half of the Florida grade 12 NAEP examinees in 2009 could be matched to an ACT score and half to an SAT score. This match rate is consistent with other data for Florida students. The match of WorkKeys scores to the total 2009 state NAEP sample of 12th graders was only about 6%. FLDOE reported that around 89,300 Florida 12th graders were enrolled in vocational-technical programs in school year 2008-09. The match of WorkKeys examinees to NAEP examinees was not sufficient to warrant additional analyses for the 2009 cycle. The state of Florida has only recently implemented the testing of high school students in vocational programs with the WorkKeys exam, and we anticipate that the number of examinees will increase in subsequent years.

Judgmental Standard Setting Studies

A series of judgmental standard setting studies was planned to produce preparedness reference points on the NAEP scale for entry into job training programs and for placement in college credit-bearing courses. Within this category of studies, the Technical Panel for 12th Grade Preparedness Research placed highest priority on the judgmental studies related to preparedness for job training programs in 5-7 exemplar jobs. This priority is largely related to the paucity of national data available for statistical studies in these areas. The Governing Board has not assumed that academic preparedness for college and for job training are the same. Rather, our studies are aimed at determining the level of performance on NAEP that represents the reading and mathematics knowledge and skills needed to qualify for job training programs for each of the occupations included in our research studies and for placement in credit-bearing college courses that fulfill general education requirements for a bachelor's degree.

In order to maximize the standardization of judgmental standard setting (JSS) studies within and across post-secondary areas, a design document was developed to specify the number of panelists, the eligibility criteria for panelists, the procedures for drafting and finalizing borderline performance descriptions, the methodology to be implemented, feedback to be provided, key aspects to be evaluated, and reports to be produced. The methodology and basic procedures specified for the design of these studies were those implemented for the achievement levels-setting process for the 2006 grade 12 economics NAEP and for the 2009 science NAEP for grades 4, 8, and 12.

The five exemplar jobs approved by COSDAM for inclusion in these studies are as follows:
1. automotive master technicians
2. computer support specialists
3. heating, ventilation, and air conditioning technicians
4. licensed practical nurses

5. pharmacy technicians

A pair of replicate panels with 10 panelists each was convened for each subject and post-secondary area for a total of 24 operational panels.

Higher Education Survey

A survey of two-year and four-year post-secondary institutions was conducted in Fall 2011 to gather information regarding (1) the placement tests used and (2) the cut scores on those tests in reading and mathematics below which need was indicated for remedial/developmental courses in reading and mathematics, and at or above which placement in credit-bearing entry level courses was indicated. The sample of accredited postsecondary education institutions was nationally representative. A weighted response rate of 81% was achieved.

Benchmarking Studies

Benchmarking studies in the preparedness research context are studies in which NAEP is administered to groups of interest, e.g., college freshmen enrolled in credit-bearing college level courses that fulfill general education requirements for a four-year degree without the need for remediation. Determining the average NAEP performance of this group would then provide a "benchmark" score that can be considered as one of the reference points on the NAEP scale. A benchmarking study in combination with reference points from other studies in the Program of Preparedness Research can assist the Board in determining the areas of the NAEP scale that indicate preparedness. A benchmarking study of Texas college freshmen was planned, and it had the support of the Texas Commissioner of Higher Education and the cooperation of nine Texas higher education institutions. A small scale pilot study to evaluate the feasibility of the study design was implemented.

The Governing Board and the National Center for Education Statistics (NCES) collaborated on the implementation of this small scale pilot study, which was carried out by Westat, the NAEP sampling and administration contractor to NCES. The data collection phase for the pilot ended on October 15, 2010. Of the eligible sample of 1,234 students, 255 actually attended a NAEP session, for an overall response rate of 20.7 percent. As announced at the November 2010 meeting of COSDAM, NCES, Westat, and Governing Board staff met to discuss alternatives. Board staff decided that we will not proceed to the operational phase of this study due to low participation rates and the lack of feasible alternatives to increase participation.

No additional benchmarking studies are planned for the 2009 NAEP preparedness research.

OVERVIEW OF REFERENCED ASSESSMENTS

For additional background information, the following list presents a brief description of the assessments that the Technical Panel on 12th Grade Preparedness Research recommended for analysis in NAEP preparedness research. Many of these assessments are the primary focus of the proposed content alignment studies and statistical relationship studies. In each case, only the

mathematics and reading portions of the assessments are the targets for analysis, although analyses with the composite scores may be conducted.

- ACCUPLACER – ACCUPLACER is a computer adaptive test used for college course placement decisions in two-year and four-year institutions. It is produced by the College Board and includes assessments of sentence skills, reading comprehension, arithmetic, elementary algebra, college level math, and written essays.

- ACT – The ACT assessment is a college admissions test used by colleges and universities to determine the level of knowledge and skills in applicant pools, including reading, English, and mathematics tests. ACT has *College Readiness Standards* that connect reading or mathematics knowledge and skills and probabilities of a college course grade of "C" or higher (75%) or "B" or higher (50%) with particular score ranges on the ACT assessment.

- ACT WorkKeys –WorkKeys is a workplace focused set of tests that assess knowledge and skills in communication (business writing, listening, reading for information, writing) as well as problem solving (applied technology, applied mathematics, locating information, observation). There is also an interpersonal skills section of WorkKeys.

- COMPASS – ACT Compass is a computer-adaptive college placement test. It is produced by ACT and includes assessments of Reading, Writing Skills, Writing Essay, Math, and English as a Second Language.

- SAT – The SAT reasoning test is a college admissions test produced by the College Board. It is used by colleges and universities to evaluate the knowledge and skills of applicant pools in critical reading, mathematics, and writing. The College Board has provided SAT score data to be used in research studies to establish a statistical relationship between the SAT and NAEP.