FINAL REPORT

Alignment Between the 2013 NAEP Grade 8 Mathematics Assessment and ACT EXPLORE Mathematics Assessment

NOVEMBER 6, 2015

PRESENTED TO:
National Assessment Governing Board
Munira Mwalimu, Contract Officer
800 North Capitol Street NW,
Suite 825
Washington, DC 20002

Project Officer:
Michelle Blair,
National Assessment Governing Board

PRESENTED BY:
NORC at the University of Chicago
Dr. Rolf Blank, Project Director
55 East Monroe Street, 30th Floor
Chicago, IL 60603

Consultant:
Dr. Norman L. Webb,
Wisconsin Center for Education
Products and Services

NORC at the UNIVERSITY of CHICAGO

Table of Contents

List of Tables

Appendices

Executive Summary

For the past decade the National Assessment Governing Board has been exploring the potential use of 12th grade NAEP reading and mathematics assessments as indicators of how well students are academically prepared for college and for job training opportunities after high school. In 2014, new research studies were initiated by the Governing Board to examine the content alignment of 8th grade NAEP and other student assessments, providing an opportunity to improve understanding of 8th grade achievement and to study the extent to which 8th grade students are on track for being academically prepared for college by the end of high school. The Governing Board contracted with the NORC at the University of Chicago, along with its subcontractor, the Wisconsin Center for Education Products and Services (WCEPS), to analyze and report on the degree of content alignment of the 2013 National Assessment of Educational Progress (NAEP) Grade 8 Reading and Mathematics assessments and the ACT EXPLORE assessments in the same subjects. For each subject area, the studies compared the two assessments—NAEP and EXPLORE—to the 2013 NAEP Framework and to the ACT College Readiness Standards (CRS). The project was conducted by a team led by NORC and used the content alignment methodology designed by Dr. Norman Webb for the Preparedness Research Program commissioned by the Governing Board.

Three research questions guided the design of the content alignment process and the analysis of data. The research questions were:

1. What is the correspondence between the mathematics content domains assessed by NAEP and EXPLORE?
2. To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on EXPLORE?
3. Are there systematic differences in content and complexity between NAEP and EXPLORE assessments in their alignment to the NAEP framework and between NAEP and EXPLORE assessments in their alignment to the ACT College Readiness Standards (CRS)? Are these differences such that entire mathematics subdomains are missing or not aligned?

Study Design

Most frequently alignment analysis is conducted between curriculum standards and student assessments rather than between two assessments. However, the methodology designed by Webb can also be used to analyze the content overlap and agreement between two assessments. The Design Document outlines how this process applies for comparing two assessments. This document was extensively reviewed and approved by the Governing Board to be used for the study. As noted in the study Design Document, two or more documents have content alignment if they support and serve student attainment of the same ends or learning outcomes. More specifically, two assessments are aligned to the degree that they are judged to target the same content knowledge at a similar level of complexity. The study design included two major steps—a framework analysis and a Content Alignment Institute (CAI).

The NAEP-EXPLORE alignment study was conducted with a bi-directional analysis process that used the NAEP Mathematics Framework as one representation of the assessment content and the CRS for the EXPLORE assessments as a second representation of the content. The CRS are separate from the content specifications of the EXPLORE assessment, and represent performance level descriptors of what students typically know and are able to do in the different score ranges derived from actual student performance on EXPLORE. The CRS were used in this alignment study because of their availability and their high level of detail on the content of the EXPLORE assessment. EXPLORE is a domain-sampled test. Forms are created by sampling the larger pool of items and do not cover exactly the same content. Equivalence of forms is achieved both by meeting multiple constraints on the number of items in each content area, the cognitive scope of the items, and match to a difficulty distribution in addition, as well as through fine-tuning using equivalent-population equating. The CRS are performance level descriptors that are articulated using the categories of ACT's content framework. The standards are derived from actual performance of students within score ranges. A standard at a score range represents something that 80 percent of students who scored in that range demonstrated that they knew or were able to do. In this study, each content representation, the CRS and the 2013 NAEP Mathematics Framework, is referred to as a framework for convenience. Before the CAI was held, the similarities and differences between the two frameworks were identified through a review conducted by an external mathematics education consultant.

The NAEP assessments are designed to monitor educational progress in the nation. Each 8th grade student tested took two 25-minute blocks of items. Matrix sampling then is used to report on the performance of the national population and subpopulations of 8th graders. The NAEP 2013 Mathematics Framework was used as the description of the content on the 2013 NAEP Grade 8 Mathematics assessment. The content is divided into five content areas:

Number Properties and Operations
Measurement
Geometry
Data Analysis, Statistics, and Probability
Algebra

These content areas are furthered delineated by subtopics and objectives. All of the 101 grade 8 mathematics objectives to be assessed were included in this analysis.

The EXPLORE assessments are designed to assess a specific student's academic progress at the 8th and 9th grade levels, especially with respect to being on track for college and career readiness. ACT College Readiness Standards (CRS) were used as the best available description of the content on the EXPLORE assessments. It should be noted that the CRS as performance-level descriptors do not include all of the content assessed by EXPLORE and that each standard represents the performance of most students at a score range, but not all. The CRS for EXPLORE Mathematics used in this study were released in 2005, and the standards relevant to EXPLORE have seven strands:

BOA: Basic Operations and Applications
PSD: Probability, Statistics, and Data
NCP: Numbers: Concepts and Properties
XE1: Expressions, Equations, and Inequalities
GRE: Graphical Representations
PPF: Properties of Plane Figures
MEA: Measurement

An eighth strand, Function (FUN), was added for the purposes of this study. Even though the CRS do not list any standards under the Function strand for grade 8, because the NAEP assessment may have items that assess, for example, function notation, the Function strand without any underlying standards was included. The CRS for EXPLORE Mathematics consist of 55 standards across the seven strands, divided into four score ranges—13-15, 16-19, 20-23, 24-25.

A mathematics education content expert conducted an analysis of the CRS mathematics framework and the NAEP Mathematics Framework. The analysis report indicated that differences were found in several key areas: what is considered the number domain in the NAEP framework appears in a single NAEP content area but is distributed among different CRS content strands; applications are specifically mentioned as a topic by the NAEP framework and are a part of some of the CRS statements but also cross most of the other topics of the CRS through a cognitive dimension; and, similarly, reasoning is mentioned explicitly in the NAEP Mathematics objectives, and specific types of reasoning is included in CRS statements but other reasoning is included through the cognitive dimension. Overall, the NAEP Mathematics Framework addresses a broader range of topics than the CRS, particularly in the content areas of geometry; data analysis, statistics, and probability; and algebra. Because not all topics assessed by EXPLORE are represented in the CRS, this comparison must be interpreted responsibly. This framework analysis report was used to prepare for the CAI.

The study design takes into consideration the different purposes for the two assessments and different structure of the two frameworks by mapping each assessment to each of the frameworks. The data from the four parts of the study can be used to determine the correspondence between the mathematics content domains assessed by each assessment; the emphasis of mathematics content given by each assessment; and any systematic differences between the content of the two assessments.

Content Alignment Institute

At the Institute convened by NORC the week of February 9-13, 2015, two panels of eight content experts compared the NAEP assessment and EXPLORE to the NAEP framework and to the CRS. The panelists were selected through a national search process. From over 100 nominees, 16 panelists were selected for each content area. These panelists represented all regions of the country, a range in years in service, and a range in ethnicity, race, and gender. They included current grade 8 classroom teachers, high school teachers, special education specialists, curriculum leaders, assessment coordinators, and graduate students and professors from higher education. At the Institute, panelists received training on the alignment process and the definition of the depth-of-knowledge levels used to describe the content complexity of standards and items. Each group was led by an experienced facilitator who had served in this capacity for a number of alignment studies and who had over 30 years of experience as classroom teachers and curriculum leaders.

The 2013 NAEP Mathematics Framework and NAEP assessment were analyzed in the Institute by the panelists. The NAEP Grade 8 Mathematics is comprised of an item pool of 153 items, which were used in the content alignment analysis. The NAEP Mathematics assessment took under 60 minutes for a student to complete, and was administered in “blocks” of items that differed from booklet to booklet and included 14-17 items per block. The 2013 NAEP Grade 8 Mathematics assessment had items written in three different formats: multiple choice (with 5 response options), short constructed response, and extended constructed response. The test was divided evenly in testing time between multiple choice and constructed response items. The maximum point value for a correct response on a NAEP item went from one to five points. The Institute analysis weighted the NAEP items by point value giving a total of 227 points for the NAEP Mathematics assessment. Weighting the NAEP items by point value provided a means for accounting for the effort required by the students in answering an item. Items with point values of two or more often had multiple parts. Thus, the point value of an item represented better what a student was likely required to do.

Two forms of the EXPLORE Mathematics assessment were analyzed in this study. Each form of EXPLORE has 30 multiple-choice items, each worth 1 point, for a total of 30 points per form. EXPLORE is a domain-sampled test with forms created by sampling from the mathematics

domain. Equivalence of forms is achieved both by meeting multiple constraints on the number of items in each content area, the cognitive scope of the items, and match to a difficulty distribution in addition, as well as through fine-tuning using equivalent-population equating. The Mathematics test is administered along with the English, Reading, and Science tests. The complete set of tests takes 2.5 hours and is usually administered in a single session.

The NAEP assessment uses matrix sampling to support group-level inferences for the nation and various jurisdictions, and so each student who takes the NAEP Mathematics assessment only encounters a subset of the full NAEP item pool of 153 items. EXPLORE is designed to report at the individual-student-level, and so each student who takes EXPLORE encounters a set of items representative of the entire EXPLORE assessment. Even with these differences, this study is focused on the content of each test, rather than what a particular student would see by taking either NAEP or EXPLORE.

The first step in the analysis process was for each of the two mathematics panels to assign depth-of-knowledge (DOK) levels to the NAEP Framework objectives. Then, the adjudication of inconsistencies was conducted to reach consensus between the two panels on the assigned DOK levels for the NAEP Framework objectives. Next, working in two panels, each educator individually assigned a DOK level to each NAEP item and then mapped the item to one to three NAEP Framework objectives. Each panel conducted within-group adjudications of the individual codes, and this was followed by adjudications between the two panels. The same procedures were then used to analyze and code the two testing forms of the EXPLORE assessment in relation to the NAEP Framework. This was followed by analysis of the NAEP assessment items to the CRS, and then the analysis of the EXPLORE assessment forms to the CRS.

Four alignment criteria developed by Webb were used to indicate the degree of alignment between the NAEP and EXPLORE assessments:

- Categorical Concurrence—the same or consistent categories of content appear in both assessments.
- Depth-of-Knowledge Consistency—the same depth of content knowledge is elicited from students by both assessments.
- Range-of-Knowledge Correspondence—there is a comparable span of knowledge within topics and categories that are targeted by both assessments.
- Balance of Representation—a similar emphasis is given to different content topics and subtopics on each assessment, as indicated by the number and weighting of assessment items.

Findings on Mathematics Content Alignment

Results from each of the four content analyses were used to describe the alignment between the NAEP and EXPLORE assessments.

The alignment between the two 30-item EXPLORE Mathematics assessments and the NAEP Mathematics Framework was found to have low values on the four alignment criteria. The EXPLORE items were found to have DOK Consistency with the assigned objectives on the NAEP framework. However, EXPLORE items were found to target from four to 30 percent of the underlying objectives for each of the five NAEP content areas. Coverage of less than one-half of the objectives for each of the content areas and only two to four corresponding items for some of the content areas are considered as weak alignment between the assessment and framework. This comparison should be interpreted in light of the fact that since the NAEP assessment uses matrix sampling, no single student ever takes all of the 153 items.

The NAEP Grade 8 Mathematics assessment aligned reasonably well with the CRS on the four alignment criteria, but also assessed a broader range of content. Both panels found about 15 percent of the NAEP items that did not correspond to any standard under the CRS. However, the other 85 percent of the items (about 130 items) aligned well to the seven strands of the CRS (Basic Operations and Applications (BOA), Probability, Statistics, and Data (PSD), Numbers: Concepts and Properties (NCP), Expressions, Equations, and Inequalities (XEI),

Graphical Representations (GRE), Properties of Plane Figures (PPF), and Measurement (MEA). Each of the strands used in the analysis were found to have from nine to 50 item point values corresponding to each strand. The content complexity of the matched items was very consistent with the content complexity of the standards, as indicated by 70 to 100 percent DOK Consistency ratings. Range-of-Knowledge was also high with 50 percent or more of the standards under each strand having at least one corresponding item. The Balance of Representation was low for four of the seven CRS strands (BOA, NCP, XEI, and PPF).

The alignment between NAEP Grade 8 Mathematics assessment and the NAEP Mathematics Framework was found to have high values for each of the four alignment criteria. The NAEP Mathematics assessment had over 25 point values for each of the five content areas (Number Properties and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra). (The point value for a correct response on a NAEP item ranged from one to five. The "point values" for a content area is the number of items times the point value per item). The two panels agreed that over 60 percent of the NAEP assessment items had a DOK level that was the same or greater than the DOK level of the corresponding objective. Items on the assessment targeted nearly 60 percent or more of the underlying objectives for each of the five content areas, and were evenly distributed among the objectives within four of the five content areas. Proportional reasoning to model and solve problems was emphasized slightly more than other objectives under the content area of Number Properties and Operations. The two panels had strong agreement in the summary results for each of the four alignment criteria.

The alignment between EXPLORE and the CRS was found to have mixed values on the four alignment criteria. The two panels had strong agreement on all four criteria. Each EXPLORE form had about eight items that targeted the Basic Operations and Applications strand (BOA), one of the seven strands, and from one to five corresponding items for each of the other strands. All seven of the strands had high DOK Consistency with the content complexity of the corresponding items. When the composite of the two EXPLORE forms was considered, four of the seven strands had an "acceptable Range-of-Knowledge (BOA, GRE, PPF, and MEA), two strands (PSD and XEI) varied by panel as to an acceptable level, and one strand had an unacceptable Range-of-Knowledge (NCP) as determined by data from both panels. Balance of

Representation was good for all seven strands indicating that the items were evenly distributed among the underlying standards for each strand.

Summary of Findings by Alignment Criteria

The findings below are based on the aggregation of the mappings by the panelists of the assessment items to the frameworks and the assignment of DOK levels. The analyses of these data from the panelists' coding results were conducted using criteria detailed by the Webb methodology and described in the Design Document for this study.

Categorical Concurrence. The NAEP Mathematics assessment covered more content than the EXPLORE forms and more content than described in the CRS. About 15 percent of the items on the NAEP assessment did not fit with the CRS standards. The NAEP Mathematics assessment had a sufficient number of point values of six or more when mapped to the five content areas of the NAEP framework or the eight strands (including Functions) of the CRS. The EXPLORE forms had a sufficient number of items for number and operations related topics, but had too few items to make reliable judgments of students' performance on the other content areas or strands, but the test is not designed to make judgments at the strand level. Both the NAEP assessment and EXPLORE had items that panelists mapped to all of the main topics on either framework. The NAEP assessment, in part because of the number of items, matched the strands with a greater number of items and addressed areas not targeted by the two EXPLORE forms.

Depth-of-Knowledge Consistency. The average DOK level of items on both assessments was nearly identical at 1.44 (with DOK levels from 1 to 4), based on the panelists' item mappings. In light of prior studies which have shown multiple choice items usually classified as a DOK level 1 or 2, the identical average DOK across the two assessments is noteworthy because EXPLORE Mathematics had multiple-choice items only, while the NAEP assessment had multiple-choice items as well as short and extended constructed response items. Both assessments had nearly the same and acceptable DOK Consistency for all of the content areas and strands.

Range-of-Knowledge Correspondence. Content coverage as represented by the Range-of-Knowledge criterion was a major difference between the two assessments. The NAEP assessment had a higher level of coverage of the underlying objectives or standards under each

of the content areas for both frameworks (from 56 to over 90 percent). EXPLORE had less than 50 percent Range-of-Knowledge for all five of the NAEP content areas. The NAEP Mathematics assessment covered more content than EXPLORE. This result may reflect the large difference in the number of items between the two assessments.

Balance of Representation. Items on the EXPLORE forms were distributed fairly evenly when mapped to either framework. Balance of Representation was high in the mapping of NAEP assessment to the NAEP framework, but not when the NAEP assessment was mapped to the CRS. The NAEP assessment over-emphasized some standards under one of the CRS strands—Numbers: Concepts and Properties.

Assessment to Assessment Alignment

Research Question 1 (Correspondence between content). Nearly all of the major topics targeted by the EXPLORE forms were also addressed by the NAEP assessment items. However, the NAEP assessments attended more to particular topics in measurement; geometry; and data analysis, statistics and probability. Both assessments targeted similar topics in number and algebra. Even though the NAEP assessment included constructed-response items, the average DOK level of items on both the NAEP assessment and EXPLORE was essentially the same.

Research Question 2 (Proportionality between content). The two assessments had similar proportions of items that corresponded to the five content areas on the NAEP Framework and the seven strands on the CRS. However, there were several differences. EXPLORE had a higher proportion of items that targeted number and operations while the NAEP had a higher proportion of items that targeted algebra and data analysis, statistics, and probability. The NAEP Mathematics assessment covered about three times the content of EXPLORE when compared to the NAEP Framework. When compared to the CRS, the NAEP covered about twice the amount of content of EXPLORE. This result may also reflect the large difference in the number of items between the two assessments.

Research Question 3 (Systematic differences between content). No large differences represented by major holes within a content domain were found between the NAEP assessment and the EXPLORE forms other than the sizeable difference in the number of items. The two mathematics assessments differed in degree rather than substance.

Conclusions

In summary, considering all four alignment criteria, there is a moderate degree of alignment between the NAEP Mathematics assessment and EXPLORE for mathematics. The NAEP assessment covered more content and was more consistent in matching expectations with regard to content complexity. In particular, the analysis showed that the NAEP Mathematics Framework and assessment had more content coverage of topics in measurement, geometry, and data analysis, statistics and probability. Both assessments targeted similar topics in number and algebra. Nearly all of the topics targeted by EXPLORE were also addressed by the NAEP assessment items. Even though the NAEP assessment included constructed-response items, the average DOK level of items on both the NAEP assessment and EXPLORE was essentially the same.

Study Limitations and Clarifications

The study was implemented very closely to the design that was planned. The process of content analysis at the Content Alignment Institute was carried out by mathematics teachers that were highly qualified and experienced, and as a group were representative of the population of mathematics teachers in the U.S. Even though there were some time pressures that resulted in not having as much time as desired for adjudication, all adjudication as specified by the methodology was completed. Some of the discussion among panelists was shortened because of this pressure. The proportion of time allocated to analyzing the NAEP assessment and to analyzing EXPLORE were similar to the proportion of items on each assessment.

The NAEP Mathematics Framework and the CRS performance descriptors for mathematics were used in this study. Both included statement of performances of 8th grade students. The difference between the two documents used in this study is the purpose—i.e., why and how they were developed. The NAEP Framework was developed to guide the item writing and construction of a

comprehensive test to be used to make inferences about the performance of a national population of students. The CRS were developed as a result of ACT's analysis of empirical evidence that represents the typical performance of students who scored within a given score range.

The Webb methodology used in this study was first developed to analyze the alignment between curriculum standards and assessments used to determine students' attainment of these standards. The alignment process was slightly modified to analyze the alignment between two assessments. As described in the Design Document, the four alignment criteria (Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation) are as applicable to judging the degree of alignment between two assessments as they are in judging the degree of alignment between an assessment and curriculum standards. What is different for the assessment-to-assessment comparison are the decision rules used to describe what acceptable alignment is. Since EXPLORE is a domain-sampled test, it may be reasonable for any one form to have only one or two items for any one CRS strand or to cover a low percentage of standards under a strand. Another difference in this study is the large difference in the number of items of the NAEP assessment, which uses matrix sampling (153 items), and the number of items on each EXPLORE form (30 items). It cannot be expected that the two assessments would cover the same content range in all of the content domains of knowledge. The methodology examines the similarities and differences in content assessed by each test by considering the relationship of each to two different descriptions of performance, the NAEP framework and the CRS, enabling the findings to be grounded in more than one perspective of the content domain.

Introduction

Over the past decade, increasing attention in the United States has been given to academic readiness and preparedness for college, career, and the military. The National Assessment Governing Board (NAGB) has worked towards expanding the use of the National Assessment of Educational Progress (NAEP) and how it can be applied as an indicator for academic preparedness of students after they leave high school (Fields, 2014). In 2004, a blue-ribbon commission recommended that NAEP be re-tuned to report on the academic preparedness of 12th graders for college, job training, and the military. To this end, the Governing Board engaged in a series of actions to guide revisions of NAEP to improve reporting on the academic preparedness of 12th graders. In 2006, the Governing Board approved changes in the 12th grade NAEP Frameworks for reading and mathematics. In 2008, an expert panel appointed by the Governing Board recommended conducting a series of academic preparedness studies. Since this time, more than 30 studies have been conducted mainly directed toward the NAEP grade 12 assessments in reading and mathematics (NAGB, no date). The expert panel identified, as one area of investigation, comparison of the content and alignment of NAEP to the widely used examinations for college admissions. Other areas of investigation included statistical analyses of the relationships between NAEP and other assessment instruments, as well as judgmental standard setting. The results from these studies have been used as validity evidence to support NAEP reporting on student academic preparedness for grade 12 NAEP reading and mathematics.

Starting in 2010, a series of studies were conducted comparing the content and alignment of the NAEP grade 12 reading and mathematics assessments to examinations used for providing information on the academic preparedness of students for college admission and course placement. The content of the grade 12 NAEP assessments in Reading and Mathematics was compared to content in the SAT,[1] WorkKeys,[2] ACT,[3] and ACCUPLACER.[4] Most of these studies used the NAEP 2009 assessments in Reading and Mathematics. Subsequently, additional studies were planned to use the 2013 NAEP grade 8 and grade 12 assessments in these content

[1] SAT is the property of the College Board.
[2] WorkKeys is the property of the ACT, Inc.
[3] ACT is the property of the ACT, Inc.
[4] ACCUPLACER is the property of the College Board.

areas. The grade 8 studies were intended to explore whether students were on track to be academically prepared for college by the end of high school. Additional grade 8 studies included statistical linking studies of the grade 8 NAEP and ACT EXPLORE in reading and mathematics.

In 2014, the Governing Board contracted NORC at the University of Chicago, along with its subcontractor, the Wisconsin Center for Education Products and Services (WCEPS), to analyze and report on the degree of content alignment of the 2013 National Assessment of Educational Progress (NAEP) Grade 8 Reading and Mathematics assessments and the EXPLORE assessments in reading and mathematics. The purpose of this contract from the National Assessment Governing Board is to evaluate the extent to which the 2013 NAEP Grade 8 Reading and Mathematics assessments are aligned in content and complexity with EXPLORE assessments. For each subject area, the studies compared the two assessments (NAEP and EXPLORE) to the NAEP Framework, and also to the ACT College Readiness Standards (CRS). The project was conducted by a team led by NORC and used the content alignment methodology designed by Dr. Norman Webb for the Preparedness Research Program commissioned by the Governing Board (NAGB, 2009; Appendix A).

Three research questions guided the design of the content alignment process and the analysis of data. The research questions were:

1. What is the correspondence between the mathematics content domains assessed by NAEP and EXPLORE?
2. To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on EXPLORE?
3. Are there systematic differences in content and complexity between NAEP and EXPLORE assessments in their alignment to the NAEP framework and between NAEP and EXPLORE assessments in their alignment to the ACT College Readiness Standards (CRS)? Are these differences such that entire mathematics subdomains are missing or not aligned?

The alignment studies reported here are the first to be conducted with the 8th grade NAEP under the academic preparedness research. As a key step in this innovative study, NORC convened a Content Alignment Institute (CAI) at the NORC facility in Bethesda, Maryland, just outside Washington, D.C. in February 2015. The results from these studies of NAEP and EXPLORE

Mathematics and Reading assessments are to have several important applications, including improving understanding of test scores from NAEP. Additionally, the project results, as outlined in this report (and that of its reading counterpart), provide a number of products including framework analyses comparing the NAEP 2013 Mathematics and Reading Frameworks with the ACT College Readiness Standards (CRS) for EXPLORE assessments for reading and mathematics, and a Content Alignment Institute involving 8th grade educators from across the U.S. in review and analysis of assessment items.

Alignment

The main goal of this study is to determine the alignment between two assessments, those of the 2013 NAEP Grade 8 Mathematics and Reading and EXPLORE Mathematics and Reading. In general, alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is an aspect of the relationship between expectations and assessments and not an attribute of any one of these two system components. Most frequently alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). Alignment is intimately related to test "validity," most closely with content validity and consequential validity (Messick, 1994; Moss, 1992). Whereas validity refers to the appropriateness of inferences made from information produced by an assessment (Kane, 2006; Cronbach, 1971), content alignment refers to the degree to which content coverage is the same between an assessment and other curriculum documents (NAGB, 2009). This study differs from most alignment studies in that the alignment between two assessments is being analyzed. One purpose for doing this study is to determine if similar inferences about student academic preparedness can be made from the NAEP Grade 8 assessments as can be done by EXPLORE.

In 2008, the Governing Board contracted the services of Dr. Norman L. Webb, Senior Research Scientist Emeritus, Wisconsin Center for Education Research, to develop a Design Document for use in a series of content alignment studies focused on comparing two assessments. This document underwent extensive review and several responsive modifications until it was approved at the March 2009 meeting of the Governing Board. The goal of the current study is to ascertain the extent to which the EXPLORE frameworks and assessments in reading and mathematics are aligned with the NAEP Grade 8 frameworks and assessments in those subjects by implementing the design of the study as described in the Design Document prepared by Dr.

Webb (NAGB, 2009; Appendix A). In this study, each content representation, the CRS and the NAEP 2013 Mathematics Framework, is referred to as a "framework" for convenience. When a specific mathematics framework is considered, it will be noted as NAEP or as CRS for the ACT College Readiness Standards.

As noted in the Design Document, two assessments are aligned to the degree that the two assessments are judged to target the same content knowledge at a similar level of complexity. Both of the assessments for this study are composed of a sample of items from the content domains of reading and mathematics. For two or more assessments to have content alignment and similar content coverage, the assessments should sample content knowledge from the same content domains. Because the number of possible assessment items that could be used to assess students' knowledge of a domain is large, it is unlikely that any two assessments targeting the same domain will have precisely the same items. An item-by-item comparison between two assessments could result in a minimal match between the assessments. The likelihood of an item-by-item match between two assessments would decrease as differences in the purposes of the two assessments increase. The NAEP assessment uses matrix sampling to support group-level inferences for the nation and various jurisdictions, and so each student who takes the NAEP Mathematics assessment only encounters a subset of the full NAEP item pool of 153 items. EXPLORE is designed to report at the individual-student-level, and so each student who takes EXPLORE encounters a set of items that represents a carefully constructed balance of topics sampled from the entire EXPLORE domain. Even with these differences, this study is focused on the content of each test, rather than what a particular student would see by taking either NAEP or EXPLORE. The method of analyzing the alignment of the 2013 NAEP Grade 8 Mathematics assessment to EXPLORE for mathematics is designed to compare the assessments by how the items on each represent similar content domains. The alignment between these two assessments will be gauged by the extent of overlapping content knowledge targeted by the two assessments and by the extent of content knowledge that is targeted and unique for each assessment. A bi-directional process was employed that included using the NAEP Framework as a representation of the assessment content and using the CRS for the EXPLORE assessments as a second representation of the content. Four alignment criteria developed by Webb (1997) were used to indicate the degree of alignment between the NAEP and EXPLORE assessments:

- Categorical Concurrence—the same or consistent categories of content appear in both assessments.
- Depth-of-Knowledge Consistency—the same depth in content knowledge is elicited from students by both assessments.
- Range-of-Knowledge Correspondence—there is a comparable span of knowledge within topics and categories that are targeted by both assessments.
- Balance of Representation—a similar emphasis is given to different content topics and subtopics on each assessment, as indicated by the number and weighting of assessment items.

The Categorical-Concurrence criterion provides a general indication of alignment if both documents incorporate the same content. *The criterion of Categorical Concurrence between assessments is met if the same or consistent categories of content appear in both assessments.* This criterion is judged by determining the number of items each assessment includes for each content area and subtopic. Two assessments agree in Categorical Concurrence if the proportion of items from each assessment assigned to each content category is similar.

Two assessments can be aligned not only on the basis of the content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-Knowledge Consistency between two assessments indicates alignment if the cognitive demand of the two assessments is approximately equal.* For consistency to exist between two assessments, as judged in this analysis, the proportion of items at each level of complexity should be similar for the main content categories and subcategories. The DOK definitions for mathematics are included as Appendices H.6.

For two assessments to be aligned, the breadth of knowledge required on the two assessments should be the same, or very nearly so. *The Range-of-Knowledge criterion is used to judge whether a span of knowledge expected of students on one assessment is the same as, or very nearly the same as, the span of knowledge expected of students on the other assessment.* The range criterion considers the proportion of subcategories (e.g., subtopics or objectives) under a content category (e.g., content area or standard) with at least one corresponding assessment item. The Range-of-Knowledge Correspondence is comparable between two assessments if the proportion of subtopics assessed is the same or similar.

In addition to comparable depth and breadth of knowledge, aligned assessments require that knowledge be distributed equally in both. The Range-of-Knowledge Correspondence criterion only considers the number of subcategories within a content category match (a subtopic with a corresponding item); it does not take into consideration how the matched assessment items/activities are distributed among the subcategories (e.g., subtopics or objectives). *The Balance-of-Representation criterion is used to indicate the degree to which one content subcategory is given more emphasis on one assessment than the other assessment.* An index is used to judge the distribution of assessment items among subcategories underlying a content category. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a content category are equally distributed among the course-level expectations for the category. Index values that approach 0 signify that a large proportion of the items only correspond to one or two of all of the subcategories with at least one assigned item. Two assessments have comparable Balance of Representation if the distribution of items among subcategories is the same as determined by a comparable index value.

To provide some means to interpret the degree of alignment between standards and assessments specific acceptable levels have been used for the four alignment criteria (Webb, 2002, 2006). The acceptable levels for each of the four alignment criteria have been used most extensively when conducting studies of the alignment of state assessments and standards. These acceptable levels are considered the lowest desirable level for an assessment and standards to be aligned such that results from the assessment can be used to make inferences on a student's performance on the curriculum standards. Six items measuring a student's content knowledge of a standard or reporting category is considered the minimum number for an acceptable value for the Categorical-Concurrence criterion. Six items was determined using a procedure developed by Subkoviak (1988) to produce an agreement coefficient of about 0.63. This indicates that about 63 percent of the tested group, more than half in a group, would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The acceptable level for the Depth-of-Knowledge Consistency criterion is to have at least 50 percent of the items corresponding to a standard or reporting category having the same or higher DOK level than the corresponding objective underlying the standard. This acceptable level is based on the assumption that a minimal passing score for any one standard of 50 percent or higher would require the student to successfully answer at least some items at or above the Depth-of-

Knowledge level of the corresponding standards. The acceptable level for the Range-of-Knowledge Correspondence criterion is to have a least 50 percent of the objectives underlying a standard to have at least one corresponding item. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. The Balance-of-Representation criterion is determined by computing an index with values from 0 to 1.0. An index value of 0.7 or higher indicates that assessment items are distributed among all of the expectations to some degree (e.g., nearly every expectation assessed has at least the same number of corresponding items) and is used as the acceptable level on this criterion. In this study, these acceptable levels for the four alignment criteria were used to interpret the degree of alignment between the two assessments and the two frameworks as well as the comparable content addressed by the two assessments.

The NAEP assessments are designed to monitor educational progress in the nation. EXPLORE assessments are designed to assess a specific student's academic progress at the 8^{th} and 9^{th} grade levels, especially with respect to being on track for college and career readiness. The study design takes into consideration the different purposes for the two assessments and different structure of the two frameworks by mapping each assessment to each of the frameworks. The data from the four parts of the study can be used to determine the correspondence between the mathematics content domains assessed by each assessment; the emphasis of mathematics content given by each assessment; and any systematic differences between the content of the two assessments.

Study Design

The following section describes the NORC and WCEPS major activities in accordance with the proposed study design for this work. The two main components of the study design entailed a Framework Analysis prior to the Content Alignment Institute (CAI), as well as the implementation of the alignment work itself at the February 2015 Institute. The Framework Analysis was the first analysis performed in the study. The main purpose of this analysis was to identify the similarities and differences between the two frameworks, the NAEP 2013 Mathematics Framework and the ACT College Readiness Standards (CRS) for EXPLORE Mathematics. Information gained from this analysis was then used to prepare for the CAI and to develop instruments that were used by the panelists. The main research questions for this project were:

- What is the correspondence between the mathematics content domains assessed by NAEP and EXPLORE?
- To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on EXPLORE?
- Are there systematic differences in content and complexity between NAEP and EXPLORE assessments in their alignment to the NAEP framework and between NAEP and EXPLORE assessments in their alignment to the ACT College Readiness Standards? Are these differences such that entire mathematics subdomains are missing or not aligned?

Framework Analysis

As outlined in the 2009 study design (NAGB, 2009; Appendix A), one main feature of the specified design for analyzing the alignment between the NAEP and another assessment was conducting a framework analysis comparing the two frameworks for the assessments. The purpose of this framework analysis was to determine how the documents developed to specify the domain of knowledge to be assessed are the same or different. The main process for conducting the framework analysis was to develop a side-by-side chart listing the content standards and objectives in one framework and then filling in the comparable content expectations for the other assessment. The Framework Analysis for mathematics considered the

similarities and differences in the included mathematical topics as well as the expectations expressed for content complexity. The NAEP framework listed content topics under five general content areas and included specifications for the percentage of items by content complexity. The CRS organized standards representing four score ranges under seven strands. The guide to the assessment (ACT, Inc., 2011, p. 8) listed the percent and number of items for each form of EXPLORE under four content areas (Pre-Algebra, Elementary Algebra, Geometry, and Statistics and Probability). Another analysis considered the overlap in the verbs used in each framework. For example, "calculate," "evaluate," and "solve" were used in both frameworks, but "substitute," "locate," and "explore" were only used in the CRS and "analyze," "apply," and "justify" were only used in the NAEP Framework. Thus, the framework analysis noted similarities and differences in content coverage between the two frameworks and the content complexities represented by each.

A second feature of the specified design was to conduct a content alignment institute that is structured around panels of content experts, including teachers, who map the items from each assessment to each of the content frameworks. Then the alignment between the two assessments was to be determined by comparing the mapping of both assessments to each of the two frameworks.

The actual design of the current alignment study followed the specifications as described in the Design Document (NAGB, 2009; Appendix A), and some minor changes were made to adjust this methodology to grade 8 assessments, as the original focus of the Design Document was grade 12. A framework analysis for mathematics was conducted by Dr. Raven McCrory (Michigan State University). This report was reviewed by the Governing Board and ACT Inc. staff. Necessary changes and additions were made on further drafts of this report, and the finalized document was provided to the Content Alignment Institute facilitators in preparation of the February 2015 Content Alignment Institute meeting. The final version of the Mathematics Framework Analysis Report can be found in Appendix B of this report.

Content Alignment Institute (CAI)

The Institute was conducted February 9 through February 13, 2015, at NORC's Bethesda facility. It included a total of four facilitators (two for mathematics, two for reading) and 32 selected panelists (16 mathematics experts, and 16 reading experts). The design for this Institute was structured to conform to the specifications provided in the Design Document (NAGB, 2009; Appendix A). The goal of the CAI was to generate data that can be used to ascertain the degree of alignment between the 2013 NAEP Mathematics and Reading assessments and the 2013 EXPLORE for mathematics and reading. The Design Document specified that an institute be conducted over a span of five days.

The plan for this study included recruitment and selection of panelists who were experienced teachers or were curriculum or assessment specialists in the subject and target grade level. A panel of eight constitutes a sufficient number to ensure high reliability of the assigned depth-of-knowledge level to a standard or assessment item and the reliability of the assigned assessment item to a content standard. Two panels were included in the design to identify and analyze potential variations in coding results that may reflect legitimate differences. Some frameworks can have overlapping standards or objectives, which may result in an item measuring content in more than one objective. For example, the NAEP 2013 Mathematics Objective I.1.e (recognize, translate or apply multiple representations of rational numbers [fractions, decimals, and percentages] in meaningful contexts) and Objective V.4.c (analyze situations or solve problems using linear equations and inequalities with rational coefficients symbolically or graphically) both target the use of rational numbers. One panelist may appropriately decide that the main content knowledge to answer an item correctly requires solving an equation with rational numbers (V4.c) while another panelist may just as appropriately decide that the assessment item requires students to apply rational numbers in a real-world context (I.1.e). If two panels working in parallel arrive at the exact same coding result, then it can be assumed there is strong confirmation that the assessment item maps to the specified standard or objective. However, variations among the panels can point to true differences that may be caused by the way the standards are written, how the assessment items are written, or the way assessment items fit within the standards. Just the sheer number of standards that panelists have to consider can result in differences in assigning items to standards. The NAEP 2013 Mathematics Framework

included 101 objectives and the CRS for EXPLORE Mathematics included 55 standards. Variations can also be caused by insufficient training of the panelists and lack of experience among panelists in analyzing assessment items. Having two panels operate in parallel makes analysis and interpretation of variations in the coding process possible and makes the variations more visible.

Another feature incorporated into the design for collecting these data is the adjudication of coding results. In adjudication, panelists discuss their differences in their initial coding results to determine if any error in coding was committed. Possible errors include:

- a panelist coding an item to an adjacent standard on the menu listing the standards rather than the intended standard (e.g., a simple clerical mistake);
- a panelist not considering the full range of the standards and coding an item to a standard with some fit, but not the best possible fit;
- a panelist not fully understanding what student thinking and work is necessary to answer an item correctly.

The study design incorporated adjudication within panels that were conducted after mapping items on an assessment to standards and objectives in the framework. For within-panel adjudication, the technical coordinator (Dr. Webb) worked with the facilitator to identify instances of large variation in the coding data among the eight panelists. Then the panelists within the group discussed and explained the reasons they had for the code they assigned. The facilitators were trained to guide the discussion to move the panelists towards agreement. In some cases, complete consensus or agreement among panelists was not possible. The amount of time allowed for discussion of any one instance of variation in coding results was restricted because of the amount of work needed to be completed over the five-day Institute. After the discussion, each panelist was asked to decide if he/she wanted to change his/her initial coding. Panelists were not required to change a coding if they felt their original mapping was the most appropriate. As a result, adjudication served to improve the agreement among panelists, but the process did not necessarily lead to complete agreement within a panel on a given code or resolve the issue of overlap or redundancies in the content standards or objectives in the framework.

Between-panel adjudication was conducted when the results generated by the two panels for a content area varied greatly. This process was carried out after three steps were complete: a) panelists assign depth-of-knowledge (DOK) codes to standards and objectives, b) panelists map the assessment items to standards and objectives, and c) within-panel adjudication. The technical coordinator made the decision on the degree of variation between results for two panels, and then reported to each of the facilitators as to which standards, objectives, or assessment items needed to be discussed by all members of both panels, a total of 16 panelists, for the subject. In these discussions, one or two members of a panel presented their argument for their panel's coding results, followed by a presentation from one or two members from the second panel. As was the practice for within-panel adjudication, the individual panelists were not required to change their code if the arguments and discussion between panels were not persuasive.

Methodology

This section describes the specific methodology and decisions that were made to perform this study. The methodology includes information about the following critical tasks: selecting panelists, identifying the content frameworks, identifying the assessments, training the facilitators, training the panelists, and implementation of the Content Alignment Institute (CAI).

Panelist Selection

Alignment was judged by content experts who participated in a content analysis of the frameworks, assessments, and their relationship. A total of 32 content experts were needed to have two panels of eight members each for mathematics and to have two panels of eight members each for reading. Because both the NAEP and EXPLORE assessments were administered to students across the nation, a decision was made to recruit qualified panelists from four regions of the country—the West, the Midwest, the South, and the East. The intent was to have the selected panelists be representative of grade 8 teachers across the characteristics of gender, race/ethnicity, and region of the U.S. A joint letter (Appendix E.1) was sent by NAGB and NORC to educational leaders in each state and to leaders of national professional organizations representing teachers and content specialists in reading/English language arts and mathematics (See Appendix D for a list of those organizations). Leaders were asked to nominate qualified educators to serve as panelists. Qualifications included significant teaching or assessment experience in mathematics or reading at the 8th grade level. The nominees could include classroom teachers, curriculum coordinators, instructional coaches, content area assessment specialists, or district- or state-level specialists. Letters were sent to the state mathematics and reading specialists and to the NAEP state coordinators to seek their support for the project and to support the effort to identify qualified panelists who would be able to participate in the planned five-day CAI. The projected composition of each panel of eight members was to include:

- one high school teacher or educator (so each panel had a representative with experience on how learning in grade 8 is used in high school courses),
- at least four grade 8 practicing or recently retired teachers,
- teachers of other middle school grades, and
- school, district, or state subject specialists.

Nomination letters were received by NORC for a total of 105 nominees (45 for reading and 60 for mathematics) by December 1, 2014. Nominees were from 25 states, representing each of the four designated regions. From this list, eight panelists and two alternates were identified by the technical coordinator and project director for each of the four panels. Selection was based on the following criteria: desired qualifications and representation by role and experience, as well as in relation to the population of grade 8 teachers and students by gender, race/ethnicity, and region. A summary of the characteristics of the selected panelists for mathematics and reading are given in Table 1. The composition of the panels are representative of the nation's teachers by race/ethnicity and gender (i.e., 76 percent of all middle grades teachers are female, and 82 percent of all teachers and 89 percent of mathematics teachers are White, non-Hispanic (NCES, 2012; Horizon Research, 2012)). The selection purposely sought inclusion of teachers who are Black, Hispanic, Asian, or American Indian. The selection also sought a balance of teachers from all four regions of the country. Three of the initially selected panelists were unable to participate due to family illness and asked to withdraw prior to the February 2015 Institute. Alternate panelists were contacted by the project director, and they agreed to participate so that each content area had a total of 16 panelists. Panelist information indicates that one mathematics panelists was a Ph.D. candidate.

Table 1 Frequency of panelists selected by subject, region, gender, race/ethnicity, and experience

Characteristic	Mathematics (N=16)	Reading (N=16)
Region		
East	4	4
South	3	5
Midwest	3	4
West	6	3
Gender		
Female	12	13
Male	4	3
Ethnicity/Race		
Black	1	4
White	13	9
Hispanic	1	1
American Indian/Alaska Native	0	1
Asian	0	1
Two or more races Asian and White	1	0
Employment level		
Classroom	4	3
District	6	6
State	5	5
Higher education	1	2
Years in Education		
<10	2	3
10 to 15	6	8
>15	8	5

Content Frameworks

National Assessment of Educational Progress (NAEP)

The NAEP 2013 Mathematics Framework (NAGB, 2012) was used to create the 2013 NAEP Grade 8 Mathematics assessment. The content is divided into five content areas:

- Number Properties and Operations
- Measurement
- Geometry
- Data Analysis, Statistics, and Probability
- Algebra

The same five content areas are used to structure the assessments for all three grades—4, 8, and 12. The content areas are further delineated by subtopics and objectives. Not all subtopics or objectives describe content to be assessed at all grades so the numbering system or the labeling of objectives has some breaks for some of the subtopics or objectives. These breaks are generally the case when objectives have been collapsed from grade 4. The NAEP Mathematics Framework was written to guide the development of the main NAEP assessments at the national, state, and district levels. It was also designed to specify what mathematics skills should be assessed at grades 4, 8, and 12. The framework is not intended to be a curriculum framework for guiding instruction. All of the 101 grade 8 mathematics objectives to be assessed were included in this analysis (Appendix H.4). Panelists mapped the assessment items to the objectives.

EXPLORE

The CRS are separate from the content specifications of EXPLORE, and represent performance level descriptors of what students typically know and are able to do, derived from actual student performance on EXPLORE exams. The CRS were used in this alignment study because of their availability and their high level of detail on the content of the EXPLORE assessment. It should be noted that the CRS as performance-level descriptors do not include all of the content assessed by EXPLORE and that each standard represents the performance of most students at a score range, but not all. The CRS are performance level descriptors that are articulated using descriptive categories. The standards are derived from actual performance of students within score ranges using normative data from EXPLORE (the test for 8th and 9th graders) along with PLAN (the test for 10th graders) and the ACT (the college admission test for 11th and 12th graders). ACT analyzed these normative data along with college admission criteria and information about actual college course placement to describe the skills and knowledge needed to achieve each score range. A standard at a score range represents something that 80 percent of students who scored in that range demonstrated that they knew or were able to do (NORC, 2014). The CRS for EXPLORE Mathematics assessments used in this study were released in 2005, and the Standards relevant to EXPLORE are organized into seven strands:

BOA: Basic Operations and Applications

PSD: Probability, Statistics, and Data

NCP: Numbers: Concepts and Properties

XE1: Expressions, Equations, and Inequalities

GRE: Graphical Representations

PPF: Properties of Plane Figures

MEA: Measurement

An eighth strand, Function (FUN), was added for the purposes of this study. Even though the CRS do not list any standards under the Function strand for grade 8, because the NAEP and EXPLORE assessments may have items that assess knowledge of functions, the Function strand without any underlying standards was included. The CRS for EXPLORE Mathematics consist of 55 standards across the seven strands, divided into four score ranges—13-15, 16-19, 20-23, 24-25. Although the strands overlap, each standard has been assigned to a primary strand. All of the CRS strands and standards used in the analysis are listed in Appendix H.3. Relative to EXPLORE, the CRS description states that "lack of a CRS statement in a score range indicates that there was insufficient evidence with which to determine a descriptor" (ACT, Inc., 2009, p. 7).

Comparison of the Two Mathematics Frameworks

A mathematics education content expert conducted an analysis of the CRS mathematics framework and the NAEP Mathematics Framework. The analysis report indicated that the NAEP 2013 Mathematics Framework for grade 8 and the CRS include the same general content topics, but differ on how more specific topics are addressed. Both frameworks addressed the general topics of number, measurement, geometry, data analysis/statistics/probability, and algebra. However, the CRS listed number ideas in several strands whereas the NAEP Framework constrained number concepts to one content area (Number Properties and Operations). The NAEP content statements explicitly referred to applications as "meaningful contexts," while the CRS included application as one of the three cognitive levels to which 27 percent of the items were to be allocated. The NAEP Framework also provides for application problems derived from targeting a high level of complexity for a given objective. The different treatment of applications may result in a variation of how application items and problems are treated on each assessment.

It is possible that more than 27 percent of items on EXPLORE could target applications. Reasoning was referenced explicitly by a few NAEP objectives but was included in the CRS as one aspect of one of the three cognitive levels (understanding concepts and integrating conceptual understanding). Items in this cognitive level may also require application skills. In multiple-choice items, students are prompted to select a response from a set of answer options, where as in constructed-response items, students can possibly express their reasoning. The NAEP Framework required the percent of testing time to be evenly divided between multiple-choice items (50 percent) and constructed-response items (50 percent). EXPLORE was limited to multiple-choice items.

The NAEP 2013 Mathematics Framework listed a higher number of objectives as well as more detailed objectives compared to the CRS under several content areas, and consequently, more items were required on the NAEP assessment to adequately represent these content areas. Overall, the NAEP Mathematics Framework addresses a broader range of topics than the CRS, particularly in the content areas of geometry; data analysis, statistics, and probability; and algebra. For example, the CRS did not mention symmetry, transformations, or solid geometry, and included locating points only in the coordinate plane (GRE 201 and 401). Because not all topics assessed by EXPLORE are represented in the CRS, this comparison must be interpreted responsibly. This framework analysis report was used to prepare for the CAI.

Assessments

NAEP

The 2013 NAEP Grade 8 Mathematics assessment of 153 items was used in this analysis. Other studies explore statistical relationships between the 2013 NAEP data and data from the 2013 administration of EXPLORE. For this study to have comparable data to these statistical studies, the 2013 assessments and frameworks were analyzed for alignment. The NAEP Mathematics test took under 60 minutes for a student to complete, and was administered in “blocks” of items that differed from booklet to booklet. The test was administered to a student in two 25-minute sections, with 14 to 17 items on each section. Any one student only took two blocks. The number of items and the administration time (90 minutes including time to answer background questions) was designed so that all students would be able to complete the work in the allocated time.

Samples from subgroups assessed are assigned sampling weightings before the scores are analyzed. A scale score is then produced for each of the five mathematics content areas for the total population and subpopulations. From this application of matrix sampling, inferences are made to the full population. No information is provided on individual students.

The 2013 NAEP Grade 8 Mathematics assessment allowed a calculator for one-third of the blocks of items. For students who did not bring one, NAEP provided a scientific calculator for test takers as part of the protocol. Students were allowed to bring whatever calculator, graphing or otherwise, they were accustomed to using in the classroom with some restrictions for test security purposes. Some items required manipulatives (e.g., number tiles, geometric shapes, rulers, protractors), which were provided. The alignment analysis took into consideration if a calculator was available for use and if any other materials were required such as manipulatives.

The 2013 NAEP Grade 8 Mathematics assessment had items written in three different formats: multiple choice (with 5 response options), short constructed response, and extended constructed response. The test was divided evenly in testing time between multiple-choice and constructed-response items. Short constructed-response items were scored at two levels (correct or incorrect) or three levels (correct, incorrect or partially correct). Extended constructed-response items had multiple parts and required more than a short answer. They were scored at either four or five levels. Scoring rubrics were used for all constructed-response items (NAGB, 2010, Chapter 4, p. 51).

The NAEP 2013 Mathematics Framework for grade 8 (NAGB, 2012) specified that the number of items on the assessment should be distributed among the five content areas as follows:

Number Properties and Operations	20 percent
Measurement	15 percent
Geometry	20 percent
Data Analysis, Statistics, and Probability	15 percent
Algebra	30 percent

EXPLORE

Two forms of the EXPLORE mathematics assessments were analyzed in this study. Each form of EXPLORE had 30 multiple-choice items, each with five answer choices. EXPLORE is a domain-sampled test. Forms are created by sampling the larger domain, strategically, to obtain representative student scores but do not include exactly the same topics on each form. Equivalence of forms is achieved by meeting multiple constraints on the number of items in each of EXPLORE' s four content areas (Pre-algebra, Elementary Algebra, Geometry, Statistics/Probability), the cognitive distribution of the items, and the match to a difficulty distribution, as well as through fine-tuning using equivalent-population equating. The mathematics test was 30 minutes long and was administered along with the English, reading, and science tests. The complete set of tests took 2.5 hours to administer and was usually administered in one block of time, including a short break. The test was given year round, at the discretion of the district or school. There was no penalty for incorrect answers and the test was not speeded. That is, most students finished in the allocated time.

Calculators were recommended to be used, wisely, when taking EXPLORE Mathematics and made available to students but students were not required to use them. Students were advised to use the calculators they were most comfortable with, including four-function, scientific, or graphing calculators except for those explicitly prohibited such as calculators with a built-in or downloaded algebra computer system functionality. The EXPLORE assessment included items where a calculator was recommended for use, items where a calculator was not recommended, and items where a calculator and a non-calculator approach were both appropriate.

The EXPLORE Mathematics assessments were designed to include items in four general content areas:

Pre-algebra	10 Items (33 percent)
Elementary Algebra	9 Items (30 percent)
Geometry	7 Items (24 percent)
Statistics and Probability	4 Items (13 percent)

Possible Impact Due to Different Nature of Frameworks and Assessments

The two assessments, NAEP and EXPLORE, have different purposes and are constructed differently to fulfill their intended purposes. The main NAEP assessment was designed to provide periodic information on student achievement of the national population of students at grades 4, 8, and 12. The results are intended to inform citizens about the nature of students' comprehension of the subject, curriculum specialists about the level and nature of student achievement, and policymakers about factors related to schooling and its relationship to student proficiency (NAGB, 2012, p. 1). For NAEP grade 8 mathematics, a large battery of items (N=153) were administered in 2013 to a national sample stratified by state and select urban districts using a matrix sampling technique. The NAEP assessment uses matrix sampling to support group-level inferences for the nation and various jurisdictions, and so each student who takes the NAEP Mathematics assessment only encounters a subset of the full NAEP item pool of 153 items. With this design a grade 8 student in the chosen sample would only take two blocks of assessment items, about one hour of testing time. Items were multiple choice, short constructed response, and extended constructed response. The point value for a correct response on a NAEP item ranged from one to five. Results from this testing were reported by select large urban districts, state, and the nation, but not by individual student. The findings are disaggregated and reported by gender, race/ethnicity, disability status, socioeconomic status, and geographic region.

The EXPLORE assessments are designed to assess a specific student's academic progress at the 8th and 9th grade levels, especially with respect to being on track for college and career readiness. EXPLORE results provide information useful to begin exploring career options, and assist in developing a plan for high school courses to prepare students to achieve their post-high school goals (ACT, Inc., 2009, p. 1). EXPLORE Mathematics was designed to emphasize quantitative reasoning rather than memorization of formulas or computational skills. Items were selected that assess knowledge and skills that are prerequisite for success in high school. Several forms were created, each with 30 items. Each form was designed to have similar psychometric properties and the same general content coverage, but varied some on specific content topics within the general content areas. For EXPLORE Mathematics, a student took one form of 30 multiple-choice items, with each item assigned a score of one point. The results of the EXPLORE Mathematics, along

with assessments in Reading, English, and Science, were given for the individual student with each student receiving information reported in four sections: Your Scores, Your Plans, Your Career Possibilities, and Your Skills.

Reflecting the different purposes of the NAEP and EXPLORE assessments, the content represented in the framework for each assessment varied. The NAEP 2013 Mathematics Framework attempted to specify a wide range of content that 8th grade students, as a group, should know. This includes content that more advanced 8th graders would be exposed to and content typically covered by the lower performing students in the grade. The NAEP 2013 Mathematics Framework was organized by five content areas that were further subdivided into subtopics and objectives with the intent for the objectives to delineate content that should be assessed under each content area. In contrast, the CRS were created to represent typical performance of 8th and 9th grade students at particular score ranges, representing students performing advanced work at the highest score range and other students in lower score ranges. Similar to the NAEP 2013 Mathematics Framework, the CRS for EXPLORE Mathematics was organized by seven strands representing four content categories. An eighth CRS strand (Function) was added to accommodate some items from the two assessments. (This is not to be confused with the Functions strand that is part of the CRS for PLAN or the CRS for the ACT, which encompasses function notation and advanced functions.) Content standards were specified under each of the CRS strands to represent the performance of 8th and 9th graders at specific score ranges on EXPLORE (score ranges 13-15, 16-19, 20-23, and 24-25). As a consequence, CRS standards progress from what students in lower score ranges can do up to what students in higher score ranges can do. Therefore, the CRS present empirically derived descriptions of knowledge and skills that students are likely to demonstrate, based on their test score.

The study design takes into consideration the different purposes for the two assessments and different structure of the two frameworks by mapping each assessment to each of the two frameworks. The study design is also intended to determine how the two assessments vary by content complexity.

The two assessments, NAEP and EXPLORE, varied greatly in the unit of analysis. EXPLORE had 30 items on each form whereas the NAEP had a total of 153 mathematics items. The NAEP Mathematics Framework had 101 objectives and the CRS had 56 standards (including the added category for a Functions strand). The differences in the unit of analysis had implications for the alignment analysis results. When an EXPLORE form of 30 items was mapped to the NAEP framework with 101 objectives clustered under the five NAEP content areas, less than one-third of the NAEP objectives would have a corresponding EXPLORE item if, for example, each item targeted only one objective. The alignment between EXPLORE assessment and the NAEP Framework would have a value that would be lower than what is generally considered acceptable on the Categorical Concurrence criterion of at least six items per content area. The alignment analysis would also result in a value far below the threshold level on Range-of-Knowledge Correspondence due to the wide difference in the number of items on one EXPLORE form and the number of objectives in the NAEP framework. The Range-of-Knowledge Correspondence value would likely improve if the number of EXPLORE items is increased by including in the analysis items from two of the EXPLORE forms. When the NAEP Mathematics assessment was mapped to the NAEP framework there was potential for high values on these alignment criteria because the number of items exceeded the number of objectives.

When the assessment items from EXPLORE were mapped to the 56 mathematics standards clustered under eight strands, 30 items on one EXPLORE form were unlikely to meet the generally acceptable level for the Categorical Concurrence criterion of at least six items for each strand. Some of the items would have to be sufficiently robust to be mapped to standards under more than one strand. The NAEP Mathematics assessment mapped to the CRS could have an acceptable level on the Categorical Correspondence criterion with the 153 NAEP items being mapped to the eight CRS strands. Both assessments had the potential for an acceptable level on the Range-of-Knowledge Correspondence criterion, i.e., 50 percent of the standards with at least one corresponding item. However, because of the relative difference in the number of items on each assessment, the maximum Range-of-Knowledge for an EXPLORE 30-item form mapped to the CRS is 54 percent whereas the maximum Range-of-Knowledge for the NAEP 153-item mathematics assessment is 100 percent.

Overall, the wide differences in both the number of items on the two assessments and the number of objectives/standards in the two frameworks had implications for the likelihood that the assessments and frameworks could be considered to be aligned. Using the study methodology and considering two of the alignment criteria, Categorical Concurrence and Range-of-Knowledge Correspondence, the degree of content alignment between the EXPLORE assessment and the NAEP Framework and the CRS is likely to be lower than the degree of alignment for the NAEP assessment with these two frameworks. The other two alignment criteria—Depth-of-Knowledge Consistency and Balance of Representation—are not affected by differences in the number of items or the number of objectives.

The data from the four studies can be used to determine the correspondence between the mathematics content domain assessed by each assessment; the emphasis of mathematics content given by each assessment; and any systematic differences between the two assessments. If two assessments have exact agreement on the mathematics content domain, then there will be a strong correspondence when both assessments are mapped to one framework and a similar correspondence when both assessments are mapped to the second framework. Also, for exact agreement, the degree of emphasis among the different content topics will be the same when both assessments are mapped to the same framework. Because of the different purposes between the two assessments and the large differences in the number of items and point-values assigned to the items, it is very unlikely the two assessments will be found to have exact agreement or one-to-one correspondence.

As noted in the Design Document (NAGB, 2009, p. 7; Appendix A), the degree of alignment between two assessments will be determined by the amount of overlapping content knowledge and skills targeted by both assessments and the content knowledge and skills unique to each assessment. Because of the big differences in the number of items on each assessment relative to the number of CRS standards, the NAEP assessment has a wider opportunity to cover the content domain assessed by EXPLORE. If the NAEP assessment covers more of the CRS than EXPLORE, the content knowledge assessed by EXPLORE would be a subset of the content knowledge assessed by the NAEP. If this is the case, then inferences from parts of the NAEP assessment can be made that are similar to those made from EXPLORE. Consequently, additional inferences would be possible from the NAEP assessment that would not reasonably be

made from EXPLORE because of the additional content knowledge assessed by the larger assessment. One conclusion in this case is that the NAEP assessment is aligned with EXPLORE, but not vice-a-versa. Alternatively, it could be possible that the content knowledge and skills covered by EXPLORE partially overlaps with the content knowledge covered by the NAEP assessment. In this case there is a common set of content knowledge assessed by both assessments. In addition, each assessment could target unique content knowledge. If the unique coverage by each assessment is large, then the degree of alignment between the two assessments will be low, and it is unlikely that similar inferences can be made from both assessments.

By mapping the items of each assessment to the two frameworks, each framework provides a language system for analyzing each assessment. Even though the two mathematics frameworks address the same general mathematics topics, there is some difference in the structure between the two frameworks. The NAEP Framework partitions the content and was developed so that the underlying objectives are in general distinct. The CRS clusters the content standards (the most specific content statements) by categories and score ranges, with the higher score ranges containing the content knowledge expected for higher scoring students as compared to the standards for the lower score ranges.

Content coverage is the determining factor for alignment and not item format, population sampling, variation in scoring, or other administration differences that may exist between the two assessments. Characteristics of the two assessments, NAEP and EXPLORE, may have some impact on the results of the study mainly because of the large difference in the number of items on each assessment.

Organization for Content Analysis

The content analysis and coding process was organized according to the NAEP Mathematics Framework and the ACT College Readiness Standards (CRS). The NAEP Framework included all 101 objectives listed in the NAEP 2013 Mathematics Framework for grade 8. The labels for each objective corresponded with those included in the framework preceded by a roman numeral indicating the content area:

I. Number Properties and Operations
II. Measurement
III. Geometry
IV. Data Analysis, Statistics, and Probability
V. Algebra

For example, I.3.a represents the first objective, "perform computations with rational numbers," found under the content area of Number Properties and Operations, within its third subtopic (Number Operations).

The analysis of the CRS included all of the standards in that document designated for EXPLORE. Numbers following the strand abbreviation —200s, 300s, 400s, and 500s—were used to identify a specific standard. All of the CRS Mathematics strands and standards used in this analysis are listed in Appendix I. Those standards at the 200 level represented content knowledge and skills students who scored at the 13-15 level are likely to demonstrate (i.e., 80 percent or more of the students who scored in this range demonstrated the knowledge and skills described by the standard, and less than 80 percent of students who scored in the next-lower range demonstrated these knowledge and skills); those at the 300 level represent content knowledge and skills students who scored at the 16-19 level are likely to demonstrate; those at the 400 level represent content knowledge and skills students who scored at the 20-23 level are likely to demonstrate, and those at the 500 level represent content knowledge and skills students who scored at the 24-25 level are likely to demonstrate. Note that some content knowledge and skills assessed by EXPLORE are not demonstrated by 80 percent of the students in the 24-25 level, and so would not be included in the CRS. Some of these knowledge and skills are included in the full CRS because the PLAN and ACT assessments also address them. As included in the ACT documents, the expectations included in the CRS will be referred to as standards.

Pre-Institute Preparation

Four facilitators were identified to lead the panelists during the CAI, two for mathematics and two for reading. Each of the facilitators had over 30 years of experience in education, including serving as classroom teachers and in leadership positions such as district mathematics specialist, reading K-12 specialist, and state assistant superintendent of public instruction. All four

facilitators had served in this capacity for other alignment studies for over 10 studies across more than 5 years.

Two weeks prior to the Institute, the facilitators were given written instructions (Appendix H.8) describing their responsibilities as a facilitator. They were also sent the Framework Analysis for their subject (Appendix B), other supporting materials such as the NAEP 2013 Mathematics Framework and ACT documents, and the coding forms to be used for mapping assessment items to standards for the NAEP Framework and for the CRS. In preparation for the CAI, the facilitators reviewed the Framework Analysis and developed rules to be given to the panelists to help them make coding decisions when some ambiguity may exist. For example, guidance for determining which depth-of-knowledge level should be assigned to a standard that includes more than one level of performance or how to differentiate among some of the CRS that are cumulative such as the CRS XEI 202 and XEI 403, both of which involve solving linear equations.

Five days prior to the Institute, the technical coordinator conducted a conference call with all four facilitators to explain the design of the study, to explain how the study was similar and different from previous studies, to review the decision rules for each content area, and to respond to any questions. The technical coordinator met again with all four facilitators in person the evening prior to the Institute to review the plans and procedures for the Institute. At this time, the facilitators reviewed the Content Alignment Institute Meeting Agenda (Appendix C) and discussed possible contingencies in case more time was needed to complete a task than what was allocated.

Panelist Training

Prior to the CAI, the project director sent a letter to all of the panelists indicating they would be trained on the alignment process; expected to assign Depth-of-Knowledge levels to expectations of the NAEP 2013 Framework for grade 8 and the CRS; and code the items for the NAEP and EXPLORE assessments to each of the sets of expectations. The letter described the general structure of the Institute, together with an information packet to assist them in their arrival and accommodations during the Institute. They received the Institute agenda, a description of the work during the five-day schedule, and travel and contact information. Panelists were not sent

any of the training materials, however, because from previous experience, it has been found that panelists perform better when all receive consistent training and information concurrently. As information resources, the panelists were provided a list of references for prior content alignment studies and reports from NAEP and ACT. All materials provided to panelists prior to and during the Institute can be found in Appendix H.

At the CAI, each panelist was instructed to "assign a Depth-of-Knowledge level to objectives and items based on your knowledge of a typical grade 8 student." In so doing they were told to think about the central challenge of the item for the typical grade 8 student. In assigning an assessment item to an objective or standard, each panelist was told to "find the objective or standard that best corresponded to the central challenge that was necessary to perform in order to answer the item correctly." If the assessment item required knowledge from more than one distinct objective or standard, then they were instructed to map the item to each of these objectives or standards, up to three at most.

Panelists were trained on the Depth-of-Knowledge (DOK) language system to distinguish among four levels of content complexity: 1) recalling information and verbatim text 2) applying skills, concepts, simple inference, and comprehension; 3) conducting significant reasoning and drawing complex inferences about implied information; and 4) extended thinking over a sustained period of time.

Additionally, the facilitators ensured that the panelists had a common understanding of the DOK levels by discussing with the panelists the definitions and then having them use the DOK levels to code sample content objectives and assessment items. When coding a DOK level to an objective or other expectations, they were instructed to consider the central performance required by the objective and then code this based on the DOK definitions they were given. Panelists also were cautioned to assign a DOK level to an expectation by what was explicitly stated in the expectation and not on what could be inferred as included in the statement. Panelists were to think about the cognitive demand and processes that would be required by a typical 8th grade student to successfully perform what was stated in the expectation. When mapping an assessment item to an objective from the NAEP framework or CRS, panelists were instructed to find the expectation that most closely matched the central performance required by the item. If an

assessment item had some relationship to the central performance required by two or more expectations addressing a common performance, then the panelists were instructed to choose the most specific expectation. If the knowledge in more than one distinct objective was necessary to correctly answer the item, then the panelists were instructed to code the item to no more than three expectations. However, panelists were cautioned to assign one item to multiple objectives sparingly and only if the knowledge expressed by two or more distinct expectations were required to answer the item correctly. Panelists were told to consider all of the distractors of multiple-choice items in assigning a DOK level to the item and in mapping the item to an expectation. For open-ended items, the panelists were instructed to consider the scoring rubric to determine the DOK level and to match the item to an expectation. It was anticipated that the panelists would find some items on either of the assessments that did not correspond to any of the expectations given in a framework. In these cases, panelists were instructed to code the item to the next higher level (e.g., topic) of expectation that corresponded to the central performance of the item. If an assessment item did not require all of the topics or only a small part of an expectation, they were instructed to write a note on what content was not addressed by the assessment item. After finishing coding all of the items for an assessment, the panelists were asked to answer three Debriefing Questions:

A. What major topics or subtopics were only partially covered by assessment items or did not have any corresponding items?

B. In what ways did the performance (DOK levels) required by the assessment items meet or did not meet the full performance as expected by the standards?

C. Compared to other assessments being analyzed, how does this assessment align to the set of standards or expectations?

In preparation for the CAI, the facilitators for each content area reviewed the framework analysis report to develop any decision rules that should be used to help panelists reduce ambiguity during the coding process. The mathematics facilitators provided a few rules for assigning items to specific standards under the College Readiness Standards. For example, mathematics panelists were instructed to assign an item to PSD 201 if whole numbers were used and to PSD 301 if decimals were used.

Observers

Governing Board staff attended the CAI. Two NAEP representatives, one for mathematics and one for reading, were present for the first two days of the Institute and on call for the remaining days. ACT, Inc. provided observers in mathematics and reading for the complete schedule of the CAI. The observers were there to answer questions about assessments or the frameworks on an as-needed basis via consultations outside of panelist sessions, but were not to participate in activities with the panelists.

Logistics

The Institute was held at the NORC facility in Bethesda, Maryland. This facility had the necessary conference room space including a large room for all participants and observers to meet on Monday and Tuesday (Days 1 and 2, respectively) and four breakout rooms, one for each panel. A personal laptop with a wireless connection was provided for each panelist and facilitator. The laptops were connected to the NORC wireless system. Prior to the CAI, the assessments were acquired by the project director from the National Center for Education Statistics (NCES) and from the ACT, Inc. An answer key for each assessment was provided and made available to the panelists. One room at the NORC offices was designated as the repository where the assessments were kept in a secured location. A procedure was established for each panelist and facilitator to check out and return the assigned assessment copies each day, as well as her/his assigned laptop. Each participant was required to sign and record the time on a form in order to receive an assessment. When the assessment was returned, the participants initialed and recorded the time. Lunch was arranged to be served at the NORC facility each day so that security could be maintained during the day and to reduce the time needed for panelists to be away from the facility.

Content Alignment Institute

Introductory Session

The Content Alignment Institute began with presentations by the project director, the Governing Board contracting officer representative (COR) as the project officer, and representatives from ACT. These presentations provided information on the structure of the project, the objectives and organization of each of the assessments, and the context for administering the assessments to students. The project technical coordinator gave more details on the design of the study, the alignment methodology, and the process to be used during the week for the content analysis. The presentation provided: a) essential training in the Depth-of-Knowledge definitions for ascertaining content complexity of standards and assessment items, b) the process for assigning DOK levels to the NAEP objectives, c) the process for mapping the NAEP and EXPLORE assessment items to the NAEP Framework, and d) how the process would repeat in relation to the ACT College Readiness Standards.

On Day 1, panelists were asked to arrive at 8:00 a.m. for registration. The focus of the first day was training and developing knowledge of the content analysis process with both large and small groups; work continued until 5:30 p.m. The work schedule for the second through fourth days of the Institute was 8:30 a.m. to 5:00 or 5:30 p.m. On the last day, the work concluded at 2:00 p.m. Each day, two 15-minute breaks were scheduled, as well as a one-hour lunch break. As part of the post-CAI analysis, Dr. Webb conducted a side-by-side comparison of the planned agenda (schedule) and the actual activities as unfolded for the week. This is discussed further in this report.

Panelist In-Person Training

On the first day of the CAI, the technical coordinator provided a general overview of the content alignment process, and the panelists were then divided into two 16-person groups by subject. In these groups, the two facilitators conducted training on the DOK definitions. In this training, panelists read the DOK definitions, discussed the differences among the four levels, and applied the definitions by assigning DOK levels to a sample of objectives and assessment items. The purpose of this part of the training was to ensure the panelists had a common understanding of the DOK definitions, and that all 16 panelists for a subject had the same basic training. However,

it was anticipated that the panelists would continue to deepen their understanding of the DOK levels as they worked in their group to develop consensus on the DOK codes for the standards and objectives and assessment items during group adjudication.

One decision rule for analysis and coding addressed the use of general vs. specific standards. If no particular specific standard or objective could be identified for a given assessment item, reviewers were instructed to code the item to the next more general level. For the NAEP 2013 Mathematics Framework, the next level was either the subtopic or the content area. For coding to the CRS, the next general level was the strand. This coding to a generic standard, or objective, sometimes indicates that the item targets an objective at a grade level other than the target grade level for the assessment. However, if the item is grade-appropriate, then this situation may instead indicate that there is a part of the content not precisely described in the set of standards or objectives. In this study, assessment items were mapped to frameworks other than the one created by the developers, and it was expected that generic standards or objectives would need to be used, for example, in mapping the NAEP assessment items to the CRS and with mapping EXPLORE items to the NAEP Framework.

Data Collection

The Version 2 of the Web Alignment Tool (WATv2) was used to enter all of the content analysis codes during the CAI. The WATv2 is a web-based tool connected to the server at the Wisconsin Center for Education Research (WCER) at the University of Wisconsin-Madison. It was designed to be used with the Webb process for analyzing the alignment between assessments and standards. Prior to the Institute, a group number was set up on the WATv2 for each of the four panels. Each panel was assigned an identification number and the facilitator was assigned as the group leader. Then the standards and objectives were entered into the WATv2 along with the information for each assessment, including the number of items, the weight (point value) given to each item, and additional comments such as the identification number for the item to help panelists find the correct item.

Timeframe for Completing Agenda

The CAI agenda (Appendix C) was developed to describe the intended or ideal timeline for the Institute. The agenda included time for training, reaching consensus on the DOK levels for the

standards and objectives of each framework, mapping each of the assessments to both frameworks, and within-group and between-group adjudication. However, the project staff anticipated that adjustments to the agenda would have to be made as the Institute proceeded due to several factors. Many of the participating educators serving as panelists had never participated in a content alignment study and were not familiar with the process. Before the beginning of the Institute, it was difficult to anticipate how much training and ongoing support would be required by these panelists and difficult to predict the speed with which the panelists would review, analyze and code the items. In retrospect, staff observed that the small number of experienced panelists were always among the first to finish the coding.

At the end of each day (Monday through Thursday), generally from 5:30 to 6:30 p.m., a debriefing meeting was held for the project director, the technical coordinator, the contract officer representative, and the four facilitators. These debriefing meetings were held to discuss any issues that may have arisen that day, review the progress made in completing the assigned work, and make any needed adjustments to the agenda. For example, on Monday, Day 1, the mathematics panelists took longer than planned to assign DOKs to the objectives in the NAEP Framework and to reach consensus. As a result, a few panelists took additional time on Tuesday morning. A chart comparing the actual time for completing the agenda vs. the planned agenda is displayed in Appendix C.1. Even though the official start time each morning was 8:30 a.m., most panelists had already begun coding for the day before this time. Two exceptions to full attendance can be noted. One panelist had to be absent on Wednesday morning, but was able to complete all of her assigned content analysis by working through breaks and the lunch period. Two panelists had to leave the Friday session at 1:00 p.m. to catch early flights home. These panelists had completed all of their assignments and had participated in all of the adjudications. All of the other panelists were at the Institute until 2:00 p.m. on Friday.

On Tuesday morning, from 8:30 to 9:00 a.m., the technical coordinator spoke to all of the panelists as a group on a few issues that the facilitators raised at the prior evening's debriefing meeting, he reminded the panelists that the NAEP frameworks were developed to design an assessment and should not be considered as a curriculum framework. He also noted that the DOK levels should not be considered a scoring rubric requiring calibration among reviewers. The process was designed for the panelists to become more knowledgeable of the DOK

definitions and the frameworks through assigning DOK levels to a framework and the adjudication process. Panelists were reminded to listen to the others when trying to decide what student knowledge a standard required and the level of content complexity that was demanded. The technical coordinator emphasized that it was important for a panel to come to a common understanding of items and the frameworks. It was suggested that the panelists should not think about the process in terms of a correct or incorrect DOK or standard. Finally, the group was reminded in coding a standard or objective, to think of not only how the standard or objective had been assessed, but to think about how it *should* be assessed. The panelists were thus encouraged to think more broadly about the standards and objectives. The intent was for the panelists not to be restrained by their knowledge of more traditional multiple-choice items, but to think about constructed-response or computer-enhanced assessment items that may be more suitable for assessing knowledge at a deeper level.

On Friday afternoon, at 1:45 p.m., the Institute concluded with a general meeting of all participants. The project director went over some of the logistics such as reference letters that would be sent to supervisors if requested. The COR congratulated the group on their hard work and thanked them. The technical coordinator reminded the group on the importance of keeping all items confidential and not to discuss any of the content with others outside the Institute. He finished by thanking everyone for their participation and by wishing all a safe trip home.

Coding Process

The following section describes the day-by-day activities of the coding process of the Institute, including steps taken during adjudication and issue resolution.

Days 1-2

After the panelists had gained some understanding of the DOK definitions, the panelists in mathematics were organized into two panels to proceed with the analysis process. In their groups of eight, the panelists registered and logged onto the WATv2. Then the panelists assigned DOK levels to the NAEP objectives, with each panelist assigning a code for a DOK level for each of the 101 objectives in the NAEP Mathematics Framework and then entering the value into the WATv2. Using a chart from the WATv2 that showed the results of all eight panelists, the facilitator engaged the reviewers in a consensus development process until all reviewers reached

agreement on the same DOK level for each objective. The technical coordinator then reviewed the DOK values for both of the mathematics panels and identified 28 objectives (28 percent) where the two panels differed, mainly between assigning a DOK 1 or 2 to the objective—that is, a level 1 demanding student recall of information or a level 2 demanding student conceptual understanding. The two mathematics panels then met together and resolved their differences on the 28 objectives. The adjudication process also was intended to increase the understanding of the DOK definitions by all panelists. For four of the 28 objectives, the 16 panelists decided on the lower DOK level (a DOK 1) and on 24 of the objectives the panelists decided on the higher value (a DOK 2 or 3).

Next, after analyzing a small sample of five assessment items from the NAEP assessment, each mathematics panel separately mapped the 153 NAEP Mathematics assessment items to the NAEP Framework. For this step, each panelist coded individually each assessment item by first assigning it a DOK level and then finding the NAEP objective that best represented the knowledge that was necessary for a student to answer the item correctly. Panelists could assign an item up to three objectives if the knowledge expressed in each objective was necessary to answer the item correctly. However, panelists were cautioned to approach the use of multiple objectives sparingly. When all of the members in one panel and seven in the other panel completed mapping all of the items, each facilitator led the panelists through an adjudication process to resolve large variation in the coding of items to objectives and the assignment of DOK levels to specific items. Items that were adjudicated within each panel included those without a majority of the panelists agreeing on the corresponding objective and those for which three or more different DOK levels were assigned. During the analysis of NAEP items, one panelist for Panel 2 was slower in coding than members of her panel, in part due to illness. This panelist continued individual coding during the adjudication session.

The technical coordinator reviewed the analysis results of both mathematics panels after they had completed within-group adjudication. He found 10 items that needed to be adjudicated between the panels. For these items, the assigned objective by the majority of one panel did not coincide with the objective assigned by the majority of the second panel; one or two panelists in each group assigned the item to the objective assigned by the majority in the other panel. The two panels met as a group of 16 with the two facilitators and adjudicated the 10 items. After the

adjudication discussion, panelists could change their code for an item if they felt there was a compelling reason.

On Day 2, the two mathematics panels coded the two forms of EXPLORE Mathematics to the NAEP Framework objectives. Two EXPLORE forms were analyzed in order to have a basis for comparison of the variation among the forms. After the panels completed the within-group adjudication, the technical coordinator identified four of the 30 items from Form 1 and six items from Form 2 that needed to be adjudicated between groups. The DOK coding results were very similar; the two panels differed significantly in assignment of DOK levels on only three items on both forms. The two mathematics panels met together to adjudicate the 10 items across the two forms. Because of time pressures, the six items across the two EXPLORE forms that varied in the assigned DOK between the two groups were not adjudicated as specified in the Design Document. For these six items the majority within a group only differed from the majority within the other group by one DOK level. As in the Design Document the DOK of items should be discussed further if three or more DOK levels were assigned to an item or two non-contiguous levels were assigned to an item).

Days 3-5

The next step was for the two panels to analyze the NAEP and EXPLORE assessments in relation to the CRS, which included 56 standards (55 standards plus one category for the Functions strand). First, each panel assigned a DOK level to each standard and then reached a consensus within their group. The Function category (FUN) was assigned a DOK 2. The two mathematics panels differed in the DOK they assigned on only four of the 56 standards. The between-group adjudication resulted in all four of the standards being assigned a DOK level 2, the higher of the values between the two panels.

Following the DOK analysis for the standards, the two panels mapped the 153 NAEP items and then the 30 items from each form of EXPLORE Mathematics assessment to the standards in the CRS. Eleven of the NAEP items assigned to standards in the CRS had to be adjudicated between the two mathematics panels because of disagreement between the two panels on the corresponding standards. The two mathematics panels varied in the assignment of items to a CRS strand only on two items on Form 1 of EXPLORE and on one item on Form 2 of EXPLORE. Because of the high agreement between panels, no between-group adjudication was conducted for the results from coding the two EXPLORE forms to the CRS. Of the 60 items, the

two groups differed significantly in assigning items to standards on only three items. This is within the five percent margin specified in the Design Document for deciding on agreement between the two groups. As discussed above for Day 2, the DOK levels of items did not vary beyond the criterion for needed discussion as indicated in the Design Document. As a validity check, the data show that not discussing the differences between the two groups for the EXPLORE-CRS alignment study had no significant impact on the overall agreement between panels or the results from the analysis.

Variations in the Process

In general, the content analysis process in the mathematics panels was performed in the intended sequence as outlined in the agenda. There were a few variations to the intended agenda. On Day 1, most members of mathematics Panel 2 completed assigning DOK levels to the NAEP standards and were waiting before engaging in the consensus process. They were instructed to go ahead and assign DOK levels to the standards in the CRS. Also, the amount of time required to assign a DOK level to each of the 101 NAEP Mathematics standards and to reach consensus within the group took longer than planned. As a consequence, the between-group adjudication was done on the morning of Day 2. This required adjusting the agenda, and the between-group adjudication of the NAEP assessment items to the NAEP Framework was done early on Day 3. One panelist from the mathematics Panel 2 was much slower in coding than the other panelists. Even with an intervention, this panelist coded items at a rate that required twice the time needed by other panelists. Rather than lose more time, the panelist continued individual coding in a separate room, while that panel was adjudicating the results and completed the coding process. The panel missed the potential contributions of this member in the adjudication process, but all the panelist's codes were completed and included in the results. The degree of agreement among panelists in assigning DOKs to items and assigning items to mathematics standards was slightly lower for mathematics Panel 2 than for Panel 1 likely in part due to this panelist not participating in the adjudication process.

The study was implemented very closely to the planned design. The process of content analysis at the Content Alignment Institute was carried out by mathematics teachers that were highly qualified and experienced, and as a group were representative of the population of mathematics teachers in the U.S. Even though there were some time pressures that resulted in not having as

much time as desired for adjudication, all within group adjudication as specified by the methodology was completed. Between group adjudications were also conducted as determined appropriate by the two group facilitators and the discussion among panelists was shortened because of this pressure. The proportion of time allocated to analyzing the NAEP assessment and to analyzing EXPLORE were similar to the proportion of items on each assessment.

Feedback Survey

Panelists were requested to complete a brief online survey at the end of the second day and the fourth day. With these two surveys, NORC requested feedback from the panelists on their views of the training for content analysis and how they thought the process was going for them. The surveys can be reviewed in Appendices J.1 and J.2. The information from the Day 2 survey was reviewed by the project director and the project technical coordinator, and adjustments were made accordingly. A description of the survey results and data tables can be found in Appendices J.3 and J.4.

Regarding their *training*, the survey results were positive. Over 90 percent of panelists responded that the training materials were easy to understand, and 94 percent responded that they understood the criteria used in coding. While the data do show overall positive responses to the training process, the data also show room for improvement. From panelist surveys, 72 percent of respondents indicated there were a sufficient number of examples to practice, and while 63 percent of panelists said they had adequate time to practice coding, 20 percent indicated a need for more practice. The questions for panelists regarding the *process* of content analysis showed very positive responses, and the panelists' views of the process did improve as the Institute proceeded. In the survey on Day 2, 88 percent of panelists indicated they were adequately prepared for the coding, and 84 percent reported that the facilitator was effective in assisting panelists and the coding process. Change in responses on several evaluation items indicate improved attitudes by the end of the week. On Day 4 of the Institute, a total of 97 percent of panelists thought their facilitator was effective, 97 percent felt at ease in applying the analysis criteria (improving from 84 percent on Day 2), and 78 percent felt they had adequate time to do the coding work (improving from 63 percent on Day 2).

Findings

Assessments and Content Complexity of Frameworks

The two assessments varied on a few attributes. Each EXPLORE form had 30 multiple-choice items, each scored as one point, whereas the 2013 NAEP Grade 8 Mathematics assessment was composed of 153 items of which three-quarters were multiple-choice items (50 percent of the assessment time). The other one-fourth of the NAEP items were constructed response (short or extended). Table 2 lists the items by the number of points given for a correct answer. Twenty-nine (29) of the NAEP assessment items were given a maximum of three to five points. The items were weighted by point value and thus 227 points was the total possible for NAEP and 30 points was the total for an EXPLORE form. Weighting the NAEP items by point value provides a means for accounting for students applying more cognitive strategies in answering an item and essentially the elimination of any possibility for correct guessing. Items with point values of two or more often have multiple parts. Thus, the point value assignment for an item gauges the likely effort required from students to complete compound items and originate ideas rather than recognize them, as would be required in a multiple-choice item. As noted above under the discussion of assessments, the large difference between the two assessments in the total number of items and point values is likely to have an impact on the content coverage of the NAEP assessment compared to EXPLORE, but weighting by point value does not have any impact on the Range-of-Knowledge Correspondence criterion. The assessments varied also in the type of items by content complexity. EXPLORE materials indicated that about one-fourth of the items targeted knowledge and skills, one-fourth of the items targeted direct applications, and about one-half targeted understanding concepts or integrating conceptual understanding. The NAEP Framework used three categories—low, moderate, and high—to describe differences in the demands that an item may make on a student. The NAEP specifications indicated that about one-fourth of the total testing time should be low in complexity, one-half moderate, and one fourth high.

Table 2 Number of items with multiple point values for the 2013 NAEP Grade 8 Mathematics Assessment and EXPLORE Mathematics Forms 1 and 2

Point Value	Number of Items	Total Point Value
NAEP Assessment		
1	115	115
2	9	18
3	25	75
4	1	4
5	3	15
Total	153	227
EXPLORE Form 1		
1	30	30
Total	30	30
EXPLORE Form 2		
1	30	30
Total	30	30

The two frameworks varied in the complexity of the NAEP objectives and the ACT College Readiness Standards (Tables 3 and 4). The panelists judged that 78 percent of the 56 CRS had a DOK level 1 (recall of information, straightforward procedures) and 22 percent had a DOK level 2 (concepts and skills), whereas 40 percent of the 101 objectives in the NAEP 2013 Mathematics Framework was judged as a DOK 1, 56 percent of the NAEP objectives were judged as having a DOK 2, and 5 percent of the NAEP objectives were judged to have a DOK 3 (strategic thinking). For the measurement content area, the NAEP and CRS were about evenly distributed between expectations calling for recall of information and straightforward procedures or conceptual understanding and skills (DOK 1 or 2). Considering the six other strands under the CRS (all except the Function category), more than two-thirds of the standards under each strand were judged to have a DOK level 1. In contrast, the NAEP Framework had from 20 percent to 52 percent of the objectives for the four content areas other than measurement with a DOK 1. Each of these four content areas also had one or two of the objectives that were judged to require strategic thinking (DOK 3). It is important to note that the CRS standards were derived from many EXPLORE, PLAN, and ACT items and represented the kind of mathematics knowledge and skills that a great majority of 8th and 9th grade students in an EXPLORE score range will demonstrate. This implies that the level of complexity of the CRS should reflect heavily the level of complexity of the items that students in a score range can solve reliably. The purpose of the

NAEP Framework was to specify the type of items to be on the assessment so the distribution of the objectives would represent the intended DOK levels by testing time and not necessarily the actual distribution of mathematics items by number in terms of content complexity.

Table 3 Percent of objectives under content areas by Depth-of-Knowledge (DOK) Levels for the NAEP 2013 Mathematics Framework for grade 8

NAEP Content Areas	Total Number of Objectives	DOK Level	Number of Objectives by DOK Level	Percent within Content Areas by DOK Level
Number Properties and Operations	27	1 2 3	14 12 1	52 44 4
Measurement	13	1 2	6 7	46 54
Geometry	21	1 2 3	8 12 1	38 57 5
Data Analysis, Statistics, and Probability	22	1 2 3	5 15 2	23 68 9
Algebra	18	1 2 3	7 10 1	39 56 5
Total	101	1 2 3	40 55 5	40 55 5

Table 4 Percent of standards under content areas by Depth-of-Knowledge (DOK) Levels for the ACT College Readiness Standards for EXPLORE Mathematics

CRS Strands	Total Number of Standards	DOK Level	Number of Standards by DOK Level	Percent within Content Areas by DOK Level
Basic Operations and Applications (BOA)	7	1 2	6 1	86 14
Probability, Statistics, and Data (PSD)	13	1 2	9 4	69 31
Numbers: Concepts and Properties (NCP)	9	1	9	100
Expressions, Equations, and Inequalities (XEI)	12	1 2	10 2	83 17
Graphical Representations (GRE)	3	1	3	100
Properties of Plane Figures (PPF)	4	1 2	3 1	75 25
Measurement (MEA)	7	1 2	4 3	57 43
Functions (FUN)	1	2	1	100
Total	56	1 2	44 12	78 22

Alignment of Assessments to the Frameworks

An important step in judging the alignment between the NAEP and EXPLORE assessments was mapping, or coding, each assessment to both frameworks. The mapping results were then used to draw conclusions on the degree of alignment between the two assessments. Each mathematics panelist individually mapped the items for each assessment to a content framework (as described in the *Coding Process* section of this report). The final results for the reporting categories for a panel were determined by averaging the results among the eight panelists within a panel.

One indicator of the alignment between an assessment and a framework is whether each panelist found in the framework an appropriate objective or standard that corresponded to the content the item was measuring. If a panelist could not identify a corresponding objective or standard, then he/she was asked to select a generic standard that provided a match between the assessment item and the more general content categories such as a subtopic, content area, or strand. Items are reported in relation to generic standards only if two or more panelists coded an item to the generic standard. If a panelist could not find even a generic standard that corresponded to an item, then she/he was to enter “uncodeable” for the item.

None of the panelists from either mathematics panel coded any of the NAEP assessment items to a generic standard when the items were mapped to the NAEP Framework (Table 5). For the coding of EXPLORE assessment forms to the NAEP Framework, one item per form was mapped to a generic standard by two or more panelists within a panel (Tables 6-8). When the two EXPLORE forms were mapped to the NAEP Framework, eight panelists from Panel 1 coded an item of EXPLORE Form 1 assessment to the generic standard V.3 (variables, expressions, and operations) (Table 6). This item required evaluating a non-linear expression whereas the NAEP Framework only had objectives that required students to develop an expression for a context or evaluate a linear expression. Five panelists in Panel 2 coded another item of EXPLORE Form 2 to the generic standard IV.3 (experiments and samples). These panelists did not find any NAEP objective that related to categorical versus numerical data. Another panelist coded the item under Number Sense and coded the item to IV.1.a (reading and interpreting data).

Two or more panelists in both panels mapped items to generic standards when they compared the 2013 NAEP Grade 8 Mathematics assessment items to the CRS (Table 7). Two or more panelists from Panel 1 coded 37 items and two or more panelists from Panel 2 coded 50 items to a generic standard. Twenty-four of the items were in common between the panels. Twenty-three items in both panels were coded by five or more panelists to generic standards. A large number of NAEP items for which panelists could not find a good match in the CRS related to geometry, i.e., Graphical Representations (GRE) or properties of plane figures (PPF). The expert framework analysis report prior to the Institute also indicated that a number of NAEP objectives on geometry did not have a corresponding standard in the CRS. Panelists in both panels also found NAEP items that were related to functions, but did not find a good fit for these items other than mapping the items to the functions (FUN) category.

When the two EXPLORE forms were mapped to the CRS, two or more panelists from both panels found two to four items on each form that did not precisely match any of the standards in the CRS (Table 8). Four or more panelists from both groups agreed that two of the items on EXPLORE Form 1 mapped to a generic standard, one item mapped to PPF and another to Probability, Statistics, and Data (PSD). For Item 5, some panelists did not find any standard that addressed identifying attributes of a polygon or identifying shapes. For Item 21, some panelists did not find any standard related to sample space. When mapping EXPLORE Form 2 to the CRS,

a majority of panelists from both groups agreed that one item mapped to PSD did not correspond to any CRS standard because there were no standards related to evaluating a survey question or experimental design. Both panels had members that assigned an item on EXPLORE Form 2 to Functions, while three Panel 1 members assigned it to Graphical Representations as did all of the Panel 2 members. This item required knowledge of slope and the *y*-intercept.

Summary of coding to generic standard. The analysis and coding of NAEP Mathematics assessment items to CRS resulted in a large number of items (37 in one panel, 50 in the second) being coded to a generic standard, with a majority of these items relating to the geometry standards or the functions standard. Only one mathematics item per EXPLORE form was coded to a generic standard in the NAEP Framework. The panelists did not code any NAEP mathematics items to a generic standard in the NAEP Framework. Two EXPLORE items were coded to a generic CRS standard.

Summary of coding items as "uncodeable." Panelists mark an item as "uncodeable" when the review indicates that the item did not correspond, even generally, to any of the topics in the framework. The data in Table 9 show that the panelists identified 17 items on the NAEP assessment that did not fit well with any of the standards under the CRS. Seven NAEP items were marked as uncodeable by more than one panelist reviewer, including two NAEP items each identified as uncodeable by four reviewers.

Table 5 NAEP items assigned to generic content expectations on NAEP Framework by two or more panelist reviewers

Assessment Panel	Generic Content Expectation on NAEP Framework
Panel 1	No generic standards coded by two or more panelists
Panel 2	No generic standards coded by two or more panelists

Table 6 EXPLORE items assigned to generic content expectations on NAEP Mathematics Framework by two or more panelists

Assessment Panel	Generic Content Expectation NAEP Framework	Number of EXPLORE Items (N Panelists)
EXPLORE 1 Panel 1	V.3	one item (8)
EXPLORE 1 Panel 2	None	No generic standards coded by more than one panelist
EXPLORE 2 Panel 1	None	No generic standards coded by more than one panelist
EXPLORE 2 Panel 2	IV.3	one item (5)

Table 7 Frequency of NAEP items assigned to generic content on ACT College Readiness Standards for EXPLORE Mathematics

Assessment Panel	Generic Content Expectation CRS Strands	Frequency of Items by Number of Panelists
Panel 1	PSD	four items by more than 6 reviewers
	NCP	one item by 2 reviewers
	XEI	one item by 4 reviewers
	GRE	four items by 2-3 reviewers; five items by 5-8 reviewers,
	PPF	eight items by 5-7 reviewers
	MEA	two items by 6 reviewers
	FUN	six items by 5-7 reviewers
Panel 2	PSD	four items by 6-8 reviewers
	NCP	one item by 2 reviewers
	XEI	two items by 4-5 reviewers
	GRE	one item by 8 reviewers; four items by 2-3 reviewers
	PPF	six items by 2-3 reviewers; eight items by 4-7 reviewers
	MEA	five items by 5-8 reviewers; three items by 2-4 reviewers
	FUN	eight items by 5-7 reviewers; five items by 2-4 reviewers

Table 8 EXPLORE items assigned to generic content expectations on ACT College Readiness Standards for EXPLORE by two or more panelists

Assessment Panel	Generic Content Expectation CRS Strands	EXPLORE Items by Number of Panelists
EXPLORE 1 Panel 1	PSD	one item by 7 reviewers
	PPF	one item by 4 reviewers
EXPLORE 1 Panel 2	PSD	two items by 7-8 reviewers
	GRE	one item by 4 reviewers
	PPF	one item by 7 reviewers; one item by 4 reviewers
EXPLORE 2 Panel 1	PSD	one item by 7 reviewers
	FUN	one item by 3 reviewers
EXPLORE 2 Panel 2	BOA	one item by 8 reviewers
	PSD	one item by 6 reviewers
	GRE	one item by 8 reviewers
	MEA	one item by 4 reviewers

Table 9 NAEP items marked as uncodeable on the ACT College Readiness Standards by number of panelists

Number of NAEP Items marked as uncodeable	Number of Panelists
10	1
3	2
2	3
2	4
17	Total items

Study 1: Alignment of the 2013 NAEP Assessment and the NAEP Mathematics Framework

The degree of alignment between an assessment and a set of standards or framework depends on how each document relates to the four alignment criteria—Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation. The alignment between the NAEP 2013 Mathematics Framework and the 2013 NAEP Grade 8 Mathematics assessment is high when considering the threshold levels. One contributing factor is the large number of items and point values on the NAEP Mathematics. In this analysis, each item is weighted by its maximum point value, which ranges from one to five points.

Categorical Concurrence. The tabulation for the Categorical-Concurrence criterion represents the total possible points for each of the content areas and not the total number of items. As the

results show in Table 10, the point values for each of the five content areas range from 26 (Measurement) to 70 (Algebra). This is well above the six items or point values used as the threshold level (see page 19) to make a reliable judgment on a student's performance on a content area. Note that assessments may not report on performance by individual content areas, and so Categorical Concurrence should not be interpreted as a reflection on the assessment itself. Even if the items were not weighted by point value, the number of items by content areas for each of the two panels is well above the threshold level. Panel 1 is 38, 20, 29, 25, and 44 items for the five respective content areas in the order as seen in Table 10 and Panel 2 is 38, 23, 27, 26, and 42 items, respectively. Among the five content areas, the greatest emphasis is given to Algebra (28 percent) and Number Properties and Operations (24 percent). The percentage of point values as determined by the panelists corresponds closely to the percentages of the number of items specified in the framework (NAGB, 2012). The two panels agreed very strongly on the average number of point values mapped to Number Properties and Operations (55 by Panel 1 and 54 by Panel 2) and to Data Analysis, Statistics and Probability (43 by Panel 1 and 46 by Panel 2). The two panels differed some in mapping items to Measurement and Geometry. The total point values across these two content areas for both panels were the same: 67 points. However, Panel 1 coded more point values to Geometry than to Measurement, whereas Panel 1 coded about the same point value to each these two content areas. Even panelists within a group did not agree if some items should be mapped to Geometry or to Measurement. The total number of items mapped to Algebra differed between panels by only two to three items, a difference of less than 5 percent of the total number of items mapped to that content area.

Depth-of-Knowledge Consistency. The Depth-of-Knowledge Consistency was relatively high for all five content areas for both panels. Panel 1 found that 83 percent to 92 percent of the weighted point values had a DOK level that was the same as or higher than the DOK level of the assigned objective. Panel 2 found that from 62 percent to 88 percent did so. All five content areas met the threshold level of 50 percent of the items with a DOK level that is at least the same as the DOK level of the corresponding objective. Panel 2 members coded more items at a DOK level 1 than did Panel 1. One or two members of Panel 2 seemed to convince fellow panelists that students should be able to do the mathematics routinely without significant mental processing. For example, items that required applying an equation in some context were coded as DOK 1 (recall of information) by Panel 2 and as DOK 2 (concept and skills) by Panel 1. This

was also the case for some geometry items. In the content areas of Number Properties and Operations, Measurement, and Data Analysis, Statistics, and Probability, the two panels agreed strongly on the level of complexity of items and also concurred that the complexity of the items matched the complexity as expected by the NAEP objectives. Even with the lower DOK values of Panel 2, the DOK Consistency exceeds the threshold level of 50 percent by more than 12 percent for Algebra and by more than 30 percent for the other four content areas. The DOK value assigned to an item correlated with the point value of the item (Table 11). Most of the multiple-choice items were assigned a DOK 1. Two- and three-point items were generally assigned DOK 1 or 2. Four- and five-point items were generally assigned DOK 2 or 3.

Table 10 Item numbers and percentages on four alignment criteria by panels for the 2013 NAEP Grade 8 Mathematics Assessment mapped to the NAEP 2013 Mathematics Framework for grade 8

NAEP Content Areas	NAEP Assessment Items Mapped to NAEP Framework Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Number Properties and Operations	55	54	83	83	62	60	0.64	0.64
Measurement	27	32	92	88	65	61	0.72	0.73
Geometry	41	36	86	73	66	64	0.75	0.78
Data Analysis, Statistics, and Probability	43	46	87	83	60	59	0.76	0.74
Algebra	70	63	86	62	80	83	0.73	0.74
Total Point Value	236	231						

Table 11 Average depth-of-knowledge level of 2013 NAEP grade 8 mathematics items by item type, panel and across panels

Item Type	Number of Items	Average DOK Panel 1	Average DOK Panel 2	Average DOK Across Panels
Multiple Choice (1 Point)	115	1.5	1.3	1.4
Constructed Response (2 Points)	9	1.6	1.4	1.5
Constructed Response (3 Points)	25	1.7	1.6	1.6
Constructed Response (4 Points)	1	2.1	2.1	2.1
Constructed Response (5 Points)	3	2.5	2.2	2.4
Total	153	1.53	1.36	1.44

Range-of-Knowledge Correspondence. The Range-of-Knowledge Correspondence measure indicates that at least one item on the NAEP assessment corresponded to 60 percent to 80 percent of the objectives under each of the five content areas. Across content areas, each panel found at least one item that corresponded to 65 or 66 of the objectives. This is high considering the large number of objectives in the grade 8 mathematics framework. At the subtopic level with 23 subtopics, the majority of Panel 1 found at least one corresponding item per objective for 20 of the subtopics (87 percent) and the majority of Panel 2 found at least one item per objective for 21 of the subtopics (91 percent). For one item, the members in Panel 2 decided that this item corresponded to Objective V.5.1 (make, validate, and justify conclusions and generalizations about linear relationships) whereas those in Panel 2 coded this item to two objectives—one under Geometry (III.3.f—describe or analyze simple properties of, or relationships between, triangles, quadrilaterals, and other polygonal plane figures) and one under Objective V.1.b (generalize a pattern appearing in a numerical sequence, table, or graph using words or symbols). The two panels strongly agreed on the Range-of-Knowledge. Each panel found at least one item that mapped to more than 50 percent of the objectives under a content area. The two panels only differed by 5 percent for any of the content areas. Both panels found Algebra to have the highest proportion of objectives with corresponding items.

Balance of Representation. Balance of Representation was also very similar between the two panels. The Balance Index[5] represents the degree of emphasis given to objectives under a content area. If the same number of items corresponds to each objective under a content area, then the Balance Index will be 1. The greater the distribution of items represents a monomial or a binomial distribution, the lower the index. An index value of 0.70 is used in this analysis as the acceptable threshold for balance as indicated in the Design Document (NAGB, 2009; Appendix A). For the NAEP alignment analysis, only the Number Properties and Operations content area did not have a Balance Index that reached this minimum level. The majority of panelists in both mathematics groups coded a total of 9 or 10 items to Objective I.4.c (use proportional reasoning to model and solve problems). Except for Objective I.3.f, panelists only found one to three items that mapped to any other objective under this content area. The overemphasis on proportional reasoning lowered the Balance Index for the Number Properties and Operations content area. However, because the other alignment criteria were met for this content area, the additional emphasis on proportional reasoning is not considered a deficit in alignment and can be explained by noting that proportional reasoning is an important topic for the middle grades, which appropriately receives additional emphasis on a grade 8 assessment. The Balance Index went from a low of 0.64 for both panels for the Number Properties and Operations content area to a high of 0.78 by Panel 2 for Geometry. The index values across the two groups differed by only 0.03 for any content area.

Panelists Responses to Debriefing Questions for Study 1. Panelists' comments provided additional information related to the Study 1 findings on NAEP assessment alignment to the NAEP Framework, using their expertise as educators. Most panelists thought the DOK levels were distributed appropriately, however, two panelists from Panel 2 noted that too many of the items had a DOK 1 ("you either know the answer or you do not know the answer"). Two panelists from each group noted that mathematical reasoning was less represented on the NAEP Mathematics assessment than they had expected based on their prior experience. No NAEP assessment item was assigned a DOK 3 by eight or more panelists

[5] Balance Index:

$1-(\sum_{k=1}^{n} | 1/(0) - I(k)/(H)|)/2$

Where O = Total number of objectives hit under a standard
I(k) = Number of items hit corresponding to objective k
H = Total number of items hit for the standard
N = Total number of objectives

(Note: Objectives are considered as underlying a standard.)

across the two panels, however four or more panelists from Panel 1 agreed that four NAEP items were a DOK 3, and three of the panelists from Panel 2 rated these items as a DOK 3. The four items were among the items assigned the highest average DOK ratings by Panel 2.

Summary of Study 1 Findings:

- **Categorical Concurrence**. Among the five NAEP content areas, the greatest emphasis is given to Algebra (28 percent) and Number Properties and Operations (24 percent). The percentage of items by topic as determined by the panelists corresponds closely to the percentages of items specified in the framework.
- **Depth-of-Knowledge Consistency** was relatively high for all five content areas for both panels. Panel 1 found that 83 percent to 92 percent of the weighted item point values had a DOK level that was the same as or higher than the DOK level of the assigned objective. Panel 2 found that from 62 percent to 88 percent of item point values had a DOK level as high as or higher than the objective.
- **The Range-of-Knowledge Correspondence** measure indicates that at least one item on the NAEP Mathematics assessment corresponded to 60 to 80 percent of the objectives under each of the five content areas.
- **Balance of Representation** index represents the degree of emphasis given to objectives under a content area, with 0.70 used as the acceptable threshold. The level was reached for four of the five content areas. The Number Properties and Operations content area did not have a Balance Index that reached this minimum level, with the majority of panelists in both mathematics groups coding a total of 9 or 10 items to Objective I.4.c (use proportional reasoning to model and solve problems).
- **Overall,** the alignment between NAEP Grade 8 Mathematics assessment and the NAEP Mathematics Framework was found to have high values for each of the four alignment criteria. The NAEP Mathematics assessment had over 25 point values mapped to each of the five content areas (Number Properties and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra). The two panels agreed that over 60 percent of the NAEP assessment items had a DOK level that was the same or greater than the DOK level of the corresponding objective. Items on the assessment targeted nearly 60 percent or more of the underlying objectives for each of the five content areas, and were

evenly distributed among the objectives within four of the five content areas. Proportional reasoning to model and solve problems was emphasized slightly more than other objectives under the content area of Number Properties and Operations. The two panels had strong agreement in the summary results for each of the four alignment criteria. Panel 2 did assign slightly lower DOK levels to NAEP items than did Panel 1. Even with this difference between the two panels, the DOK Consistency as assigned by Panel 2 was above the acceptable threshold.

Study 2: Alignment for EXPLORE Forms 1 and 2 and the NAEP 2013 Mathematics Framework

The alignment between EXPLORE Forms 1 and 2 and the NAEP 2013 Mathematics Framework was generally weak overall, considering all four alignment criteria and threshold levels. The alignment of the NAEP framework with EXPLORE Form 1 also varied some from the alignment of the framework with EXPLORE Form 2, indicating some variation between the two different assessment forms. The alignment criteria for Categorical Concurrence and Depth-of-Knowledge Consistency were found to be higher for Form 2 than for Form 1.

Categorical Concurrence. Both of the EXPLORE forms had only 30 items, making it very difficult to have a sufficient number of items (at least six) for each of the five NAEP content areas. The two panels agreed that about half of the items on each of the forms corresponded to expectations under the Number Properties and Operations content area (Tables 12 and 13). Both panels also agreed that about four items on both forms targeted content related to the Data Analysis, Statistics, and Probability content area. But the panels differed some on how many items from EXPLORE Form 2 mapped to Algebra (5 or 7 items) but did agree that EXPLORE Form 1 had four algebra items. Panel 2 mapped two items to Algebra whereas Panel 1 mapped one of those items to Geometry and one of the items to Number Properties and Operations. Both panels agreed that EXPLORE Form 1 had eight items that targeted Measurement and Geometry content and that EXPLORE Form 2 had five (Panel 2) or six (Panel 1) items that targeted these NAEP content areas. However, of the six or eight items, both panels found more items that mapped to Geometry on Form 1 and more items that mapped to Measurement on Form 2. EXPLORE forms have a sufficient number of items to assess a student on Number Properties and Operations and the Geometry NAEP content areas reliably using six items as the threshold.

For three NAEP content areas, the two to four EXPLORE items mapped are fewer than needed to provide reliable information on the content areas. If the content areas of Measurement and Geometry were combined, then EXPLORE forms would have a sufficient number of items to make a reliable judgment on the combined content area. Note that EXPLORE is not designed to report performance on these individual content areas.

Depth-of-Knowledge Consistency. The Depth-of-Knowledge Consistency was generally acceptable for the alignment threshold, considering the two EXPLORE forms relative to the NAEP Mathematics Framework. Over 50 percent of the items had a DOK level that was at least as high as the DOK level of the corresponding objective on both forms for Number Properties and Operations; Geometry; and Algebra. For Data Analysis, Statistics, and Probability, the four items on Form 1 had comparable DOK levels at least as high as the DOK levels of the assigned objectives, but on Form 2, two of the four items had DOK levels at least as high as the DOK level of the corresponding objectives, which just meets the 50 percent level for the DOK Consistency criterion. The results between the two panels for the Data Analysis, Statistics, and Probability content area are not consistent between the two forms, making it difficult to judge if the data items would always have a matching DOK level for any one EXPLORE form. Items mapped to the NAEP Measurement content area had a similar finding. On EXPLORE Form 2, both panels agreed that at least two of the four items had a DOK level comparable to the level of the matching objectives. However, on Form 1 both panels agreed that two Measurement items had DOK levels that were below the DOK levels of the corresponding objectives. Overall, the DOK Consistency was acceptable for alignment between the EXPLORE forms and the NAEP Framework for three of the content areas, but weaker for Measurement and Data Analysis, Statistics, and Probability. When compared to just the 115 multiple-choice items on the NAEP assessment, the 60 multiple-choice EXPLORE items had a slightly higher average DOK level as coded by both panels, an average of 1.40 for NAEP and 1.44 for EXPLORE (Tables 11 and 14).

Range-of-Knowledge Correspondence. The analysis of EXPLORE forms relative to the NAEP Framework did not show acceptable levels for Range-of-Knowledge Correspondence. On both forms, panelists found items that targeted from 4 percent (Geometry on Form 2) to 31 percent (Measurement on Form 2) of the objectives under a content area. The Range-of-Knowledge

Correspondence was increased when the composite[6] of the two EXPLORE forms was considered (Table 15). The increase in the Range-of-Knowledge Correspondence for the composite forms indicates that the two EXPLORE forms did not target exactly the same NAEP objectives. However, even when considering the composite form of 60 items mapped to the 110 objectives, less than 50 percent of the underlying objectives for each of the five NAEP content areas were targeted. An assessment with items that corresponded to less than one-third of the objectives for four of the five content areas is not considered to have sufficient coverage of content from the specified content domain to be considered aligned under the study methodology.

Balance of Representation. The EXPLORE forms had items that were distributed fairly evenly among the objectives of the NAEP Mathematics Framework. Additionally, four items mapped to Objective I.3.f on both forms and three items mapped to one or two other objectives under the Number Properties and Operations content area. The rest of the objectives under this content area with assigned items only had one or two items.

Panelists Responses to Debriefing Questions for Study 2. Panelists' comments to the debriefing questions supported the Study 2 findings, including the high number of items that corresponded to the Number Properties and Operations content area by the EXPLORE forms while other NAEP topics were not covered as well. One panelist noted that EXPLORE items were more skills-based.

Summary of Study 2 Findings:

- **Categorical Concurrence**. The EXPLORE forms had a sufficient number of items to assess a student on Number Properties and Operations reliably using six items as the threshold, as well as for the Algebra content area, but not for the other three NAEP content areas.
- **Depth-of-Knowledge Consistency** was acceptable for alignment between EXPLORE forms and the NAEP Framework for three of the content areas, but weaker for Measurement and Data Analysis, Statistics, and Probability. Compared to the 115 multiple-choice items on the NAEP assessment, the 60 multiple-choice EXPLORE items had a higher average DOK level.

[6] The composite EXPLORE form refers to the aggregation of the two EXPLORE forms, 1 and 2, into one form of 60 items. The composite EXPLORE form is discussed in this report to consider the possibility that more content is covered over two forms (the composite form) than any one form.

- **The Range-of-Knowledge Correspondence** analysis of EXPLORE forms relative to the NAEP Framework did not reach threshold levels for Range-of-Knowledge Correspondence. On both forms, panelists found items that targeted from 4 percent of the objectives (Geometry) to 31 percent (Measurement) of the objectives under a content area, lower than the threshold of 50 percent.
- **Balance of Representation**. The EXPLORE Mathematics forms had an acceptable Balance of Representation in relation to the NAEP Mathematics Framework, with all NAEP content areas showing Balance Indices above the 0.70 criterion level.
- **Overall**, the alignment between each of the two 30-item EXPLORE Mathematics assessments and the NAEP Mathematics Framework was found to have low values on some of the four alignment criteria. Items on EXPLORE forms mapped to the objectives from the NAEP Framework, but there were too few EXPLORE items to have adequate alignment. Both EXPLORE forms placed a relatively large emphasis on the Number Properties and Operations content area with nearly half of EXPLORE items on each form mapped to this NAEP content area. The DOK levels of EXPLORE items for the NAEP areas of Measurement and Data Analysis, Statistics, and Probability varied and the results produced inconsistent findings for these content areas. EXPLORE items were found to have DOK Consistency with the assigned objectives on the NAEP framework. However, EXPLORE items were found to target from four to 30 percent of the underlying objectives for each of the five NAEP content areas. Coverage of less than one-half of the objectives for each of the content areas and only two to four corresponding items for some of the content areas is considered weak alignment between the EXPLORE assessment and NAEP framework, according to the study alignment criteria. This EXPLORE assessment to NAEP framework comparison should be interpreted while considering the 101 objectives on the NAEP framework, relative to the 30 items on each of the two EXPLORE assessment forms.

Table 12 Item numbers and percentages on four alignment criteria by panels for EXPLORE Mathematics Form 1 mapped to the NAEP 2013 Mathematics Framework for grade 8

NAEP Content Areas	EXPLORE Form 1 Mapped to NAEP Framework Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Number Properties and Operations	14	14	91	69	31	30	0.75	0.75
Measurement	2	2	44	19	9	11	1.00	0.96
Geometry	6	6	83	69	22	23	0.84	0.87
Data Analysis, Statistics, and Probability	4	4	90	79	13	12	0.83	0.83
Algebra	4	4	63	71	21	18	0.98	0.94
Total Point Value	30	30						

Table 13 Item numbers and percentages on four alignment criteria by panels for EXPLORE Mathematics Form 2 mapped to the NAEP 2013 Mathematics Framework for grade 8

NAEP Content Areas	EXPLORE Form 2 Mapped to NAEP Framework Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Number Properties and Operations	14.5	13.5	78	62	31	27	0.73	0.73
Measurement	4	4	94	56	28	32	0.94	0.98
Geometry	2	0.9	100	86	5	4	1.00	1.00
Data Analysis, Statistics, and Probability	4	4	51	46	19	18	0.98	0.96
Algebra	5	7	80	56	24	28	0.89	0.82
Total Point Value	30	30						

Table 14 Average depth-of-knowledge level of EXPLORE Mathematics items across two forms by item type, panel and across panel

Item Type	Number of Items	Average DOK Panel 1	Average DOK Panel 2	Average DOK Across Panels
Multiple Choice (1 Point)	60	1.53	1.36	1.44

Table 15 Percent of objectives under a content area with at least one corresponding item from the composite of two EXPLORE Mathematics forms (1 and 2) mapped to the NAEP 2013 Mathematics Framework for grade 8

NAEP Content Areas	Composite of EXPLORE Forms 1 and 2 Mapped to NAEP Framework Mathematics by Panels 1 and 2	
	Range-of-Knowledge (percent of standards with at least one hit)	
	Panel 1* %	Panel 2* %
Number Properties and Operations	44	44
Measurement	31	31
Geometry	24	24
Data Analysis, Statistics, and Probability	32	26
Algebra	37	33

*An objective was considered hit by a panel if four or more panelists coded the same item from either form to the objective.

Study 3: Alignment of 2013 NAEP Grade 8 Mathematics Assessment with ACT College Readiness Standards for EXPLORE

In general, the 2013 NAEP Grade 8 Mathematics assessment and the CRS were aligned according to three of the four alignment criteria (Table 16). Both panels found a sufficient number of item point values for each of the eight strands, including Functions (FUN). The DOK levels of the items were in general consistent with the DOK levels of the corresponding standards and a reasonable number of standards under each strand had at least one corresponding item to satisfy the threshold level for Range-of-Knowledge (50 percent or more of standards with at least one item). However, the Balance Index was lower than the minimum acceptable level (0.70) for four or five of the strands indicating that for these strands a high percentage of the items mapped to only one or two of the underlying standards. Also, there was a relatively large

number of items, at least 24 (16 percent), that did not map to any specific standard but mapped only at the strand level (Table 7).

Categorical Concurrence. The Categorical Concurrence between the NAEP assessment and the CRS was high with both panels coding items with more than eight point values to each of the eight strands (Table 16). For the seven strands, not including FUN, the NAEP point values were at least 12. The CRS strands with the highest number of NAEP point values by both panels were Probability, Statistics, and Data (PSD) (about 52 points) and Basic Operations and Applications (BOA) (about 44 points). Both panels mapped more items to the generic standard for the Properties of Plane Figures (PPF), eight items, compared to about six items mapped to underlying standards (Table 7). So even though 20 points were found to correspond to PPF, over half of these did not match any specific standard listed under the strand. The point values corresponding to each strand were very consistent between the panels except for the Numbers: Concepts and Properties (NCP) strand. Both panels agreed on the same 10 items that mapped to the NCP strand, but Panel 1 also mapped several multiple-point NAEP items to the strand, e.g., one three-point item and one five-point item. Panel 2 mapped these items to PPF and to PSD, respectively. Of interest is that about 50 percent of the NAEP items were found to target standards at the 400 or 500 levels, representing skills and understandings likely to be demonstrated by students scoring in the higher range, 20-25, on the EXPLORE scoring scale. At least 24 (16 percent) of the items did not map to any specific standard, but only at the strand level (Table 7).

Depth-of-Knowledge Consistency. Depth-of-Knowledge Consistency was high. Across all eight CRS strands, 70 percent or more of the items had a DOK level that was at least as high as the DOK level of the assigned standard. For six of the eight strands, the DOK Consistency was over 90 percent. The DOK Consistency with the CRS Measurement (MEA) strand was lower because five of the nine NAEP items that panelists coded to Standard MEA 201 (with DOK 2) were judged to have a DOK level 1. Standard MEA 201 (estimate or calculate the length of a line segment based on other lengths given on a geometric figure) was assigned a DOK 2 whereas the items were judged to involve only finding the length. One factor for such a high Depth-of-Knowledge Consistency is that 78 percent of the standards were judged to have a DOK Level 1,

and so any items mapped to these standards would support DOK Consistency since DOK Level 1 is the lowest level.

Range-of-Knowledge Correspondence. Range-of-Knowledge Correspondence was reasonable between the grade 8 NAEP Mathematics assessment and the CRS. The NAEP assessment had items that mapped to more than half of the standards underlying a CRS strand for all seven strands where this applied. For three of the strands (BOA, GRE, and MEA), the NAEP items mapped to more than 80 percent of the underlying standards. The results across panels had values that varied from 3 percent (MEA- 81 and 84 percent) to 19 percent (GRE- 81 and 100 percent). The GRE strand is comprised of three standards. Panel 1 coded Item 6 to GRE 201, but Panel 2 coded this item to GRE 301, and this was the only item coded to this standard by four or more panelists. The 10 percent difference between panels in Range-of-Knowledge for PPF was similar. PPF only had four underlying standards. About half of the panelists in Panel 1 coded a NAEP item to CRS standard PPF 501 whereas only one panelist from Panel 2 coded any item to this standard. The panels in general had high agreement on the Range-of-Knowledge, varying only by one or two items across the strands. The Range-of-Knowledge was lower for NCP because most items mapping to this strand corresponded to one standard (CRS standard NCP 401). Very few of the items corresponding to NCP were coded to the lower standards in the 200 to 300 range. Overall, the Range-of-Knowledge Correspondence between the NAEP assessment and the CRS indicated that the NAEP Mathematics assessment matches a high percentage of the standards, particularly those representing the higher score ranges.

Balance of Representation. The Balance of Representation between the NAEP assessment and the CRS met the threshold for four of the eight strands. Four of the strands (BOA, NCP, XEI, and PPF) had a Balance Index below the threshold of 0.70 as analyzed by both panels. Two other strands (PSD and MEA) were below this level as determined by Panel 2. Even though Range-of-Knowledge was fairly high for all of the strands, panelists in both groups found most of the items that corresponded to a strand mapped to only one standard. The most severe case of this was for Numbers: Concepts, and Properties (NCP). Panelists found 20 to 29 point values worth of NAEP items that mapped to this strand, but nearly all of the items mapped to one standard, CRS standard NCP 401. Panel 1 mapped 10 of 14 items targeting the NCP strand to the CRS Standard NCP 401. Panel 2 mapped eight of 12 items targeting the NCP strand to the CRS standard NCP

401. The lower Balance of Representation between the NAEP assessment and the CRS could be partly due to the cumulative structure of the standards. Panelists chose to map items to the standards representing the higher score intervals. Considering the Balance of Representation and Range-of-Knowledge together, the NAEP assessment was found to match a reasonably high number of the standards under the CRS, but many of the standards only had one or two mapped NAEP items.

Panelists Responses to Debriefing Questions for Study 3. Under four or five of the strands, a relatively large number of NAEP items mapped to one standard. Even though there was general alignment with most items, 16 percent of the NAEP items did not match with any of the CRS standards. Most of these items were in the NAEP content areas of Geometry and Measurement. Panelists in their responses to the debriefing questions explained the issue further. A panelist from Panel 1 observed, "The NAEP assessment had a wide variety of questions. However, the standards for the ACT did not align well at all. There were no items that addressed functions[7], minimal geometry, no slope/y-intercept or estimation." A panelist from Panel 2 observed, "Many [NAEP] items are not listed in the [CRS] but their building blocks are." The latter panelist's comment suggests some match between NAEP items and the CRS, but not at the full level of knowledge needed to answer the NAEP items correctly. Other comments by panelists supported the view that the range of DOK levels required by the NAEP items was broader than the range in the CRS.

Summary of Study 3 Findings:

- **The Categorical Concurrence** between the NAEP assessment and the CRS was high with both panels coding items with more than eight point values to each of the eight strands. The strands with the highest number of point values were Probability, Statistics, and Data (PSD) (about 52 points) and Basic Operations and Applications (BOA) (about 44 points).
- **Depth-of-Knowledge Consistency** was high. Across all eight CRS strands, 70 percent or more of the items had a DOK level that was at least as high as the DOK level of the assigned standard.

[7] The Function Strand was included in case some items on the NAEP assessment may target content related to this topic.

- **Range-of-Knowledge Correspondence** was reasonable between the grade 8 NAEP Mathematics assessment and the CRS. The NAEP assessment had items that mapped to more than half of the standards underlying a strand for all seven strands where this applied.
- **Balance of Representation** between the NAEP assessment and the CRS was at a threshold level for four of the eight strands. Four of the strands (BOA, NCP, XEI, and PPF) had a Balance Index below the acceptable level of 0.70 as analyzed by both panels.
- **Overall,** the NAEP Grade 8 Mathematics assessment aligned well with the CRS on most of the four alignment criteria, but also assessed a broader range of content than captured by the CRS. The panels found 16 percent of the NAEP items that did not correspond to any standard under the CRS. However, the other 84 percent of the items (about 130 items) aligned well to the eight strands of the CRS, with panelists mapping from nine to 50 NAEP item point values to each strand. The content complexity of the matched NAEP items was very consistent with the content complexity of the CRS standards, as indicated by 70 to 100 percent DOK Consistency ratings.

Table 16 Item numbers and percentages on four alignment criteria by panel for the 2013 NAEP Grade 8 Mathematics Assessment mapped to the ACT College Readiness Standards for EXPLORE Mathematics

CRS Strands	NAEP Grade 8 Mathematics Assessment Mapped to ACT College Readiness Standards for EXPLORE Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Basic Operations and Applications (BOA)	43.5	45	98	97	95	91	0.68	0.67
Probability, Statistics, and Data (PSD)	52	52	96	97	69	62.5	0.70	0.62
Numbers: Concepts and Properties (NCP)	29	20	100	100	59	53	0.43	0.47
Expressions, Equations, and Inequalities (XEI)	38	38	92	78	71	66	0.62	0.65
Graphical Representations (GRE)	14	12	100	100	81	100	0.78	0.76
Properties of Plane Figures (PPF)	20	20	93	98	85	75	0.67	0.68
Measurement (MEA)	31	31	73	74	81	84	0.73	0.68
Functions (FUN)	8.5	11	100	100	100	100	1.00	1.00
Total Point Value	236	229						

Study 4: Alignment of EXPLORE Mathematics Assessment with ACT College Readiness Standards for EXPLORE

EXPLORE was found to have at least some items that targeted each of the seven CRS strands, although only one strand had six or more items, which is considered the threshold needed to make a reliable judgment about a student's performance on the strand. The content complexity of the items matched the DOK levels of the corresponding standards. Range-of-Knowledge was low for four to six of the strands, but items were adequately distributed among standards with corresponding items without overemphasizing any one standard. Panelists found only slight differences between the EXPLORE Mathematics forms in the two CRS geometry areas of Graphical Representations and Properties of Plane Figures (Tables 17 and 18). For example, whereas most panelists agreed that Form 1 of the assessment had two or three items that mapped to these strands respectively, Form 2 only had one or two items that did so.

Categorical Concurrence. The Categorical Concurrence between the EXPLORE assessment and CRS did not meet the threshold of six items per strand for seven of the CRS strands. Panelists matched an average of four EXPLORE items per CRS strand. The BOA strand had the most corresponding items at eight. With seven strands (excluding FUN), it is very difficult to have six or more items for each strand with a 30-item assessment. The six-item level was attained for one strand—Basic Operations and Applications (BOA) (Tables 17 and 18). For BOA, Form 1 had seven or eight items and Form 2 had nine items that mapped to this strand. Panelists found four items that mapped to most strands. The only exceptions to this were for GRE with Form 1 and GRE and PPF with Form 2.

Depth-of-Knowledge Consistency. The Depth-of-Knowledge Consistency between the two EXPLORE forms and the CRS was rated at a high level with over 80 percent of the items assigned a DOK level that was the same as or higher than the DOK level of the corresponding standard. The Form 2 items that mapped to PPF and MEA are the exception. The DOK for these strands and the assessment items on Form 2 had a DOK Consistency ranging from 46 to 76 percent. The panels did not agree on the DOK levels of the items mapped to the MEA content area. Other than for these five or six items, the level of complexity on the assessment represented the level as expected by the standards.

Range-of-Knowledge Correspondence. Range-of-Knowledge Correspondence was low with very similar results for both forms and panels. At most, only three or four standards had one or more corresponding items under each of the seven strands. Each strand had from three to 13 underlying standards. The low number of items on each EXPLORE form, as compared to the larger number of NAEP items, contributed to a low percentage of the standards being well-matched with assessment items. The Range-of-Knowledge was improved if the composite of both forms was used (the aggregation of two test forms to create one test form of 60 items, see Table 19). Each form targeted standards not assessed by the other form thus increasing the Range-of-Knowledge. This was true for both panels. When just one form was considered, only one to three of the seven strands had a range of over 50 percent by either panel. However, when a composite of the two forms was considered, five of the seven strands had a range of at least 50 percent by either panel. From the results, it is evident that a higher number of standards under nearly all seven strands (excluding FUN) was assessed when both EXPLORE forms were considered.

Balance of Representation. Balance of Representation was rated as high for both assessment forms with each strand reporting a Balance Index above the 0.7 level. A key explanatory factor was that no standard was overemphasized because most standards were matched to only one or two EXPLORE items.

Panelists Responses to Debriefing Questions for Study 4. The panelists noted some topics that appeared to be missing from the assessment forms, such as topics related to properties of plane figures, sample space, and linear algebraic functions. Panelists felt the DOK levels of the items were generally comparable to those of the standards, but at least one panelist noted that some of the items had a higher DOK level than expected by the corresponding standard.

Summary of Study 4 Findings:

- **Categorical Concurrence** of EXPLORE and the CRS in mathematics was found to be adequate if a lower threshold is used, four items rather than six items, for five of seven strands. Both panels found fewer than four items that mapped to the two geometry strands, GRE and PPF. If these two strands were combined as a geometry strand, then all six of the strands would have at least four corresponding items.
- The **Depth-of-Knowledge Consistency** between the two EXPLORE forms and the CRS was rated at a high level with over 80 percent of the items for most strands assigned a DOK level that was the same as or higher than the DOK level of the corresponding standard. The Form 2 items that mapped to PPF and MEA were the exception. The DOK for these strands and the assessment items on Form 2 had a DOK Consistency ranging from 46 to 76 percent.
- **Range-of-Knowledge Correspondence** was rated as low compared to the acceptable level with the analysis showing similar results for both EXPLORE forms and across each mathematics panel. Only three or four standards had one or more corresponding items under each of the seven of the strands.
- **Balance of Representation** was high for both assessment forms and this result is partly explained by at most one or two EXPLORE items matching any standard.
- **Overall,** the alignment between EXPLORE and the CRS was found to have mixed values on the four alignment criteria. The two panels had strong agreement on all four criteria. Each EXPLORE form had about eight items that targeted the Basic Operations and Applications strand (BOA), one of the seven strands, and from one to five corresponding items for each of the other strands. All seven of the strands had high DOK Consistency with the content complexity of the corresponding items. When the composite of the two EXPLORE forms was considered, four of the seven strands had an acceptable Range-of-Knowledge (BOA, GRE, PPF, and MEA), two strands (PSD and XEI) varied by panel as to an acceptable level, and one strand had an unacceptable Range-of-Knowledge (NCP) as determined by data from both panels. Balance of Representation was good for all seven strands indicating that the items were evenly distributed among the underlying standards for each strand. Note that EXPLORE is not designed to report performance on individual strands.

Table 17 Item numbers and percentages on four alignment criteria by panels for EXPLORE Mathematics Form 1 mapped to the ACT College Readiness Standards for EXPLORE Mathematics

CRS Strands	EXPLORE Mathematics Form 1 Mapped to ACT College Readiness Standards for EXPLORE Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Basic Operations and Applications (BOA)	8.5	7	100	93	64	62.5	0.78	0.80
Probability, Statistics, and Data (PSD)	4	4	97	100	23	24	0.87	0.86
Numbers: Concepts and Properties (NCP)	4	5	100	100	29	36	0.78	0.83
Expressions, Equations, and Inequalities (XEI)	4	5	91	81	32	33	0.96	0.90
Graphical Representations (GRE)	2	2	100	100	62.5	55	1.00	0.97
Properties of Plane Figures (PPF)	3.5	3.5	96	84	62.5	49	0.93	0.90
Measurement (MEA)	4	3.5	87.5	80	37.5	39	0.84	0.90
Functions (FUN)	0	0	NA	NA	NA	NA	NA	NA
Total Point Value	30	30						

Table 18 Item numbers and percentages on four alignment criteria by panel for EXPLORE Mathematics Form 2 mapped to the ACT College Readiness Standards for EXPLORE Mathematics

CRS Strands	EXPLORE Mathematics Form 2 Mapped to ACT College Readiness Standards for EXPLORE Mathematics by Panels 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	Panel 1 No.	Panel 2 No.	Panel 1 %	Panel 2 %	Panel 1 %	Panel 2 %	Panel 1 0-1	Panel 2 0-1
Basic Operations and Applications (BOA)	9	9	100	100	62.5	59	0.71	0.8
Probability, Statistics, and Data (PSD)	5	5	89	90	29	34	0.89	0.96
Numbers: Concepts and Properties (NCP)	5	4.5	100	100	31	40	0.73	0.89
Expressions, Equations, and Inequalities (XEI)	5	5	84	89	35	30	0.94	0.82
Graphical Representations (GRE)	0.6	1.4	100	100	21	34	1.00	1.00
Properties of Plane Figures (PPF)	1.8	1.4	69	62.5	31	25	1.00	1.00
Measurement (MEA)	4	4	76	46	40	40	0.84	0.86
Functions (FUN)	0.4	0	NA	NA	NA	NA	NA	N/A
Total Point Value	30.8	30.3						

Table 19 Percent of standards under a content area with at least one corresponding item from the composite of two EXPLORE Mathematics forms (1 and 2) mapped to the ACT College Readiness Standards for EXPLORE Mathematics

CRS Strands	Composite of EXPLORE Forms 1 and 2 Mapped to the ACT College Readiness Standards by Panels 1 and 2	
	Range-of-Knowledge (percent of standards with at least one hit)	
	Panel 1* %	Panel 2* %
Basic Operations and Applications (BOA)	71	88
Probability, Statistics, and Data (PSD)	43	57
Numbers: Concepts and Properties (NCP)	44	44
Expressions, Equations, and Inequalities (XEI)	58	33
Graphical Representations (GRE)	67	50
Properties of Plane Figures (PPF)	80	60
Measurement (MEA)	71	62
Functions (FUN)	0	0

* An objective was considered hit if four or more panelists coded the same item from either form to the objective.

Alignment between the Two Assessments

To compare the content and coverage between the two assessments—the 2013 NAEP Grade 8 Mathematics assessment and the two forms of EXPLORE Mathematics assessment—the data were aggregated across panels and the two EXPLORE forms (Tables 20 and 21). For most categories the two mathematics panels had very similar results for both the NAEP and the EXPLORE assessments. The data in Tables 20 and 21 under the columns labeled NAEP are the averages for two panels for the given category. For the columns labeled EXP, the values are the averages across the two panels and the two forms of EXPLORE included in the analysis. It should be noted that the aggregated value for Range-of-Knowledge Correspondence was computed by first considering a composite of the two EXPLORE forms and then averaging across the two panels. The Range-of-Knowledge for the composite of the two test forms was determined by counting the number of objectives/standards across the two EXPLORE forms with at least one item assigned by four or more panelists. Then this number was divided by the total number of standards, including any generic standard, under the content area or strand.

The findings below are based on the aggregation of the mappings by the panelists of the assessment items to the frameworks and the assignment of DOK levels. The analyses of these

data from the panelists' coding results were conducted using criteria detailed by the Webb methodology and described in the Design Document for this study (NAGB, 2009; Appendix A).

Categorical Concurrence. The NAEP assessment targeted similar mathematics content as EXPLORE. Items were found on each assessment that corresponded to the objectives or standards under all of the content categories for both the NAEP Framework and the CRS, with the exception that no FUN items were found in EXPLORE. In addition, the NAEP assessment had 15 percent of its items that did not correspond to any of the CRS. The NAEP Mathematics assessment had a sufficient number of point values of six or more when mapped to the five content areas of the NAEP framework or the eight strands (including Functions) of the CRS. The EXPLORE forms had a sufficient number of items for number and operations related topics, but had too few items to make reliable judgments of students' performance on the other NAEP content areas or CRS strands. It should be noted that EXPLORE was not designed to make judgments at the strand level. There were some differences between the assessments in proportion of items corresponding to the different topics. The EXPLORE assessment had a higher proportion of items that targeted Number Properties and Operations. When compared to the NAEP Framework, 47 percent of the items on the two EXPLORE forms corresponded to the Number Properties and Operations content area (Table 20) compared to 23 percent of the point values of the NAEP. Also, 28 percent of the items on the two EXPLORE forms targeted standards under the Basic Operations and Application strand of the CRS, compared to 19 percent of the NAEP point values (Table 21). In addition, 16 percent of the items on the two EXPLORE forms mapped to the CRS targeted standards under the Numbers: Concepts and Properties (NCP) Strand compared to 11 percent of the NAEP items. The NAEP assessment also had a slightly higher percentage of items that targeted the NAEP areas of Algebra and Data Analysis, Statistics, and Probability. When compared to the NAEP Framework, 29 percent of the NAEP point values corresponded to the Algebra content area compared to 17 percent of the average between the two EXPLORE forms (Table 20). This difference is not as apparent when the two assessments were compared to the CRS because the NAEP Algebra content was distributed among the seven strands. The 4 percent of the NAEP point values that corresponded to the Function strand is another indication that a higher proportion of the NAEP items targeted the general area of algebra than of EXPLORE items. Comments by the panelists also noted that EXPLORE did not have very many items that required solving equations, slope, and *y*-intercept. The two

assessments also varied some on the NAEP content area of Data Analysis, Statistics, and Probability and the CRS strand of Probability, Statistics, and Data. The NAEP assessment had 5 percent more items that targeted this area when the assessments were mapped to the NAEP Framework (Table 20) and 7 percent more items when the assessments were mapped to the CRS (Table 21). The proportion of items in the area of geometry, including measurement and plane figures, were about the same on both assessments.

Depth-of-Knowledge Consistency. The two assessments were rated very closely on the Depth-of-Knowledge Consistency when mapped to each framework. The average DOK level of items on both assessments was nearly identical at 1.44 (with DOK levels from 1 to 4), based on the panelists' item mappings. Even though the NAEP Framework had a higher percentage of objectives with a DOK 2 or 3 (60 percent) than the percentage of standards in the CRS with a DOK 2 (22 percent), the two assessments were similar to each other in content complexity. However, a few differences were found. For example, the NAEP Measurement items were more comparable on content complexity with the NAEP objectives in this content area than the small number of EXPLORE items that mapped to this content area, 90 percent compared to 53 percent on DOK Consistency (Table 20). A similar difference was found for the Data Analysis, Statistics, and Probability content area. Panelists found only 67 percent of items on EXPLORE with a DOK at least as high as the corresponding objectives, compared to 85 percent of the items on the NAEP assessment (Table 21). When the assessments were compared to the CRS, 95 percent of the 20 NAEP point values that corresponded to standards under the Properties of Plane Figures (PPF) had DOK levels at least as high as the matching standards. The EXPLORE forms had a small number of items corresponding to PPF and the items had a lower DOK Consistency of 78 percent.

Range-of-Knowledge Correspondence. The big difference in the number of items on each assessment, 153 items on the NAEP and 30 items on each EXPLORE form, greatly influenced the results on the Range-of-Knowledge Correspondence criterion. The NAEP assessment had a higher level of coverage of the underlying objectives or standards under each of the content areas for both frameworks (from 56 to over 90 percent). EXPLORE had less than 50 percent Range-of-Knowledge for all five of the NAEP content areas. The NAEP assessment targeted at least 60 percent of the objectives under each of the given content areas on the NAEP Framework whereas

each of the EXPLORE forms targeted at most 44 percent of the objectives. The EXPLORE assessment fared better when compared to the CRS, with from 44 percent to 79 percent of the standards under each strand matching at least one item from the EXPLORE forms. Only two strands did not meet a threshold level of 50 percent for EXPLORE items—Numbers: Concepts and Properties (44 percent) and Expressions, Equations, and Inequalities (46 percent). The NAEP assessment Range-of-Knowledge in relation to the CRS varied from 56 percent to 93 percent of the underlying standards under each strand. The Range-of-Knowledge Correspondence for EXPLORE improved when the composite EXPLORE form was used.

Balance of Representation. Both the NAEP assessment and EXPLORE had adequate balance when compared to either the NAEP Framework or the CRS without overemphasizing any one objective or standard. Of course, the characteristics of each framework contributed to how items were distributed. The balance of the assessments when compared to the NAEP Framework were all reasonable except for a slightly lower value for the NAEP assessment and the Number Properties and Operations content area with an index value of 0.64 (Table 20). When the NAEP assessment was mapped to the CRS, the Balance Indices were much lower (.45 to .77) while the EXPLORE Balance Indices with respect to the CRS remained high (.77 to .96) (Table 21). This suggests that the NAEP items are discriminatory among fine-grained topics. This is evident when the NAEP items are mapped to the NAEP Framework with 101 objectives but not when the assessment is mapped to more general statements of content as in the CRS with 56 standards. The reasonable level of Range-of-Knowledge between the NAEP and the CRS indicates that the NAEP assessment did map to the content in general, but the low Balance is possibly due to the structure of the CRS framework with not as many distinctions as in the NAEP Framework (i.e., 101 NAEP objectives vs. 56 CRS objectives).

Summary of Alignment between the Two Assessments: Overall, the results from this analysis indicate that the two assessments are moderately aligned.

- Both assessments target similar content areas with only some variation in the relative emphasis—for example, the content area related to number and operations was emphasized more on EXPLORE, and the content areas related to algebra and data analysis/statistics/probability were emphasized more on the NAEP.
- The two assessments were similar on DOK Consistency. Even though NAEP included constructed response items, none of the items targeted the more complex objectives, Objectives I.6.b, III.5.a, and V.5.a, which relate to making a mathematical argument, generating conjectures, and justifying conclusions. Only two items were judged by the majority of panelists in one panel to have a DOK level 3. Nearly all of the items on NAEP were either a DOK 1 or 2. EXPLORE multiple-choice items had an average DOK level higher than the NAEP multiple-choice items.
- Both assessments primarily targeted recall of information, conceptual understanding, and skills. However, NAEP measured more breadth in content knowledge than EXPLORE. This is clearly related to just the number of items used in this research project, but NAEP also covered a reasonable range of content relative to both frameworks.
- NAEP items targeted more detailed objectives within the main content topics than EXPLORE.

Table 20 Mean number of items and percentages on four alignment criteria for content areas of the NAEP 2013 Mathematics Framework for grade 8 mapped to the NAEP Mathematics assessment and two forms of EXPLORE Mathematics test

NAEP Content Areas	Assessments mapped to NAEP 2013 Mathematics Framework for grade 8 Averaged Across Panels 1 and 2 and EXPLORE Forms 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	NAEP	EXP	NAEP	EXP	NAEP	Com EXP*	NAEP	EXP
Number Properties and Operations	54.8 (23%)	14.2 (47%)	83	75	61	44	0.64	0.74
Measurement	29.3 (13%)	3.1 (10%)	90	53	63	31	0.73	0.97
Geometry	38.1 (16%)	3.7 (12%)	79	85	65	24	0.77	0.93
Data Analysis, Statistics, and Probability	44.5 (19%)	4.1 (14%)	85	67	59	29	0.75	0.90
Algebra	66.4 (29%)	5.0 (17%)	74	68	82	35	0.74	0.91
Total Point Value	233.1	30.1						

* The Range-of-Knowledge for the composite of the two EXPLORE forms was computed for each content area and each panel by counting the number of standards with at least one item coded by four or more panelists and then dividing by the total possible standards under the content area including the generic standard if appropriate. Then the averages of the Range-of-Knowledge percent for each content area for the composite test of 60 items was computed across the two panels.

Table 21 Mean number of items and percentages on four alignment criteria for strands of the ACT College Readiness Standards mapped to the 2013 NAEP Grade 8 Mathematics Assessment and two forms of EXPLORE

CRS Strands	Assessments mapped to CRS Strands Averaged Across Panels 1 and 2 and EXPLORE Forms 1 and 2							
	Categorical Concurrence (mean hits weighted by point value)		Depth-of-Knowledge Consistency (percent of hits at or above DOK of standard)		Range-of-Knowledge (percent of standards with at least one hit)		Balance of Representation (Balance Index)	
	NAEP Avg.	EXP Avg.	NAEP %	EXP %	NAEP %	Com EXP* %	NAEP Index	EXP Index
Basic Operations and Applications (BOA)	44.3 (19%)	8.3 (28%)	97	98	93	79.5	0.68	0.77
Probability, Statistics, and Data (PSD)	51.9 (22%)	4.4 (15%)	97	94	66	50	0.66	0.90
Numbers: Concepts and Properties (NCP)	24.8 (11%)	4.7 (16%)	100	100	56	44	0.45	0.81
Expressions, Equations, and Inequalities (XEI)	37.4 (16%)	4.7 (16%)	85	86	69	45.5	0.64	0.91
Graphical Representations (GRE)	13.3 (6%)	1.5 (5%)	100	100	91	58.5	0.77	0.99
Properties of Plane Figures (PPF)	20.2 (9%)	2.5 (8%)	95	78	80	70	0.68	0.96
Measurement (MEA)	31 (13%)	3.8 (12%)	73	73	83	66.5	0.71	0.86
Functions (FUN)	9.8 (4%)	0.1 (0%)	100	100	100	N/A	1.00	N/A
Total Point Value	232.7	30.0						

* The Range-of-Knowledge for the composite of the two EXPLORE forms was computed for each content area and each panel by counting the number of standards with at least one item coded by four or more panelists and dividing by the total possible standards under the content area including the generic standard if appropriate. Then, the average of the Range-of-Knowledge percentages for each content area for the composite test of 60 items was computed across the two panels.

Reliability of Data

Based on a recommendation from an NORC internal group of experts, different statistics were considered to represent agreement among the panelists including the Cohen Kappa. Two statistics were chosen that were appropriate to use with multiple panelists assigning categorical levels to items and from which the average across panelists was computed to report findings. The Shrout-Fleiss (1979) intraclass correlation (ICC) for a mean rating reliability was used to determine the agreement among reviewers in assigning DOK levels to items. The Winer

Reliability was used as a second measure of agreement in assigning the DOK levels. Both of these were computed by a NORC education researcher who conducted the analyses independently. A pairwise comparison was used to determine the degree of agreement among reviewers coding items to objectives/standards and content areas/strands. The pairwise agreement was computed by comparing the content code assigned by each panelist with the content code assigned by each of the other panelists. The number of exact agreements were counted across the 28 comparisons (the number of possible pairwise comparisons among 8 panelists). Then the number of agreements was divided by 28. The average agreement for each item was then totaled and divided by the total number of items. This value was used as the pairwise agreement.

The overall intra-class correlation among the mathematics panelists' assignment of DOK levels to items was high for eight participants in each panel for all 12 analyses (Tables 22 and 23). This was true as indicated by both the Shrout-Fleiss ICC and the Winer Reliability. An intra-class correlation value greater than 0.8 generally indicates a high level of agreement among raters. The intra-class correlations and Winer Reliability for assigning DOK levels to items for all 12 analyses were 0.83 or higher.

A pairwise comparison was used to determine the degree of agreement among the panelists coding at the objective/standard level and at the content area/strand level. The pairwise content area/strand agreements for 10 of the 12 analyses were all above 0.80, which is reasonably high for most alignment studies. Both mathematics panels had a slightly lower agreement, 0.77 for Panel 1 and 0.76 for Panel 2, when assigning the NAEP assessment items to the CRS (Table 23). Panelists disagreed in assigning about 10 items to Basic Operations and Applications (BOA) and either to Number: Concepts and Properties (NCP) or Measurement (MEA). The pairwise agreement in assigning items to specific objectives or standards ranged from 0.24 to 0.31. These values are lower than for prior alignment studies. One explanation involves the number of objectives and standards included in each framework and the time pressures. The panelists had strong agreement on assigning items to the content area or strand for nearly all analyses, but within these levels the panelists differed on the precise objective or standard that the items matched. The reporting categories used in this study are the content area and strand. Low

agreement in assigning items to objectives or standards does not strongly impact the overall findings.

Panelists did engage in an adjudication of their data after all panelists finished their coding for an assessment. These discussions were used to identify any mistakes in coding or to improve on a panelist's coding. Panelists were not required to change their coding unless they found a compelling reason to change. A spot check was made on what changes were made from adjudication by comparing pre- and post-adjudication for the Mathematics Panel 1 for mapping the NAEP assessment to the NAEP Framework and for the Mathematics Panel 2 for mapping EXPLORE Form 1 to NAEP Framework. After adjudication, the mathematics panels made changes in the coding of objectives to an item on about one quarter of the items (27 percent Panel 1 and 22 percent Panel 2) and in assigning DOK levels to items on about one third of the items (31 percent Panel 1 and 33 percent Panel 2). The number of panelists who made changes in coding an objective to an item varied from one to all eight panelists for Panel 1 and from one to six of the eight panelists for Panel 2. When changes in assigned objectives were made, they were made most commonly by five panelists from Panel 1 and three panelists from Panel 2. All panelists from both panels made changes in their DOK codes on at least seven items. The percentage of items that any one panelist changed in DOK level varied form 7 percent to 46 percent for Panel 1 and 17 percent to 47 percent for Panel 2. Once a panel completed the within-group adjudication, the panelists were reluctant to make changes from the between-group adjudication. Thus, the between-group adjudications only resulted in few changes in either assigning items to an objective or assigning a DOK level to an item. There was increased confidence in the data after adjudication, particularly after the within-group adjudication. The process provided the panelists the opportunity to be sure they considered the full range in possibilities and understood better what a student had to do in order to correctly answer an item. Over the course of the CAI, panelists increased in their agreement and thus had to adjudicate the codes for fewer items.

Table 22 Intra-class correlations, Winer reliability, and pairwise comparisons for the alignment analysis of the 2013 NAEP Grade 8 Mathematics Assessment and EXPLORE Mathematics Forms 1 and 2 mapped to NAEP 2013 Mathematics Framework for grade 8 (Panel 1 N=8 and Panel 2 N=8)

Study and Panel	Shrout-Fleiss Intra-class Correlation for DOK	Winer Reliability: Mean of 8 Raters for DOK	Pairwise: Standard/ Objective	Pairwise: Content Area/Strand
NAEP to NAEP Panel 1	0.89	0.90	0.25	0.85
EXPLORE 1 to NAEP Panel 1	0.92	0.91	0.30	0.95
EXPLORE 2 to NAEP Panel 1	0.90	0.90	0.31	0.95
NAEP to NAEP Panel 2	0.85	0.84	0.24	0.83
EXPLORE 1 to NAEP Panel 2	0.88	0.85	0.29	0.86
EXPLORE 2 to NAEP Panel 2	0.84	0.83	0.28	0.89

Table 23 Intra-class correlations, Winer reliability, and pairwise comparisons for the alignment analysis of the 2013 NAEP Grade 8 Mathematics Assessment and EXPLORE Mathematics Forms 1 and 2 mapped to ACT College Readiness Standards for EXPLORE Mathematics (Panel 1 N=8 and Panel 2 N=8)

Study and Panel	Shrout-Fleiss Intra-class Correlation for DOK	Winer Reliability: Mean of 8 Raters for DOK	Pairwise: Standard/ Objective	Pairwise: Content Area/Strand
NAEP to CRS Panel 1	0.89	0.88	0.25	0.77
EXPLORE 1 to CRS Panel 1	0.92	0.91	0.30	0.85
EXPLORE 2 to CRS Panel 1	0.90	0.90	0.31	0.81
NAEP to CRS Panel 2	0.85	0.84	0.24	0.76
EXPLORE 1 to CRS Panel 2	0.88	0.85	0.29	0.83
EXPLORE 2 to CRS Panel 2	0.85	0.83	0.28	0.82

Conclusions

Table 24 summarizes the alignment results for each of the four analyses using the percentage of the reporting categories (content area for NAEP and strand for the CRS) with a threshold level for each of the alignment criteria. The threshold levels here are those described on page 19. These minimum acceptable levels are somewhat arbitrary, but provide at least one gauge for comparing the results across the four analyses completed in this study. Of course, other threshold levels could be used that would result in either improved or lower degrees of alignment.

Table 24 Percent of content areas or strands with threshold levels for alignment between the NAEP 2013 Mathematics Framework for grade 8 and the ACT College Readiness Standards and the NAEP and EXPLORE Assessments

Assessment and Standards	Alignment Criteria			
	Categorical Concurrence (at least 6 items)	Depth-of-Knowledge Consistency (at least 50% match)	Range-of-Knowledge Correspondence (at least 50% of objectives hit)	Balance of Representation (Index value of 0.70 or more)
NAEP Assessment with NAEP Framework (N=5 categories)	100%	100%	100%	80%
EXPLORE with NAEP Framework (N=5 categories)	20%	100%	0%*	100%
NAEP Assessment with CRS (N= 7 Strands)	100%	100%	100%	28%
EXPLORE with CRS (N=7 Strands)	14% (75%)#	100%	71%*	100%

* The percentage is based on the composite of the two EXPLORE forms.
# Categorical Concurrence is 75 percent when analyzing only four content categories (pre-algebra, elementary algebra, geometry, statistics and probability) rather than seven strands. Note that EXPLORE is designed to be a general measure. It does not report on any of these categories or strands individually.

Process Outcomes and Alignment Results

The study was implemented very closely to the design as described in the Design Document (Appendix A). The process of content analysis at the Content Alignment Institute was carried out by mathematics teachers and leaders that were highly qualified and experienced, and as a group were representative of the population of mathematics teachers and leaders in the U.S. There were time pressures to complete all of the work at the five-day institute, however all of the panelists completed their content analysis and data code entry. Both panels completed all of the within-

group adjudications. Both between-group adjudications with the NAEP assessment were completed. The between-group adjudication after coding the two EXPLORE forms to the CRS was not performed because there was a reasonable agreement between the two panels. The overall agreement within each panel in assigning DOK levels to assessment items and items to content areas or strands was reasonably high. The agreement in assigning items to objectives or standards was lower. This lack of agreement at the objective or standard level was not considered to be significant because the results were reported at the content area and strand levels. Clearly, some panelists would have benefitted from having more time. However, the reasonably high agreement among panelists and between groups indicates the data are reliable and that time pressures did not critically influence the coding by panelists.

The NAEP Mathematics Framework and the CRS performance descriptors for mathematics were used in this study. Both included statement of performances of 8th grade students. The difference between the two documents used in this study is the purpose—i.e., why and how they were developed. The NAEP Framework was developed to guide the item writing and construction of a comprehensive test to be used to make inferences about the performance of a national population of students. The CRS were developed as a result of ACT's analysis of empirical evidence that represents the typical performance of students who scored within a given score range.

The Webb methodology used in this study was first developed to analyze the alignment between curriculum standards and assessments used to determine students' attainment of these standards. The alignment process was slightly modified to analyze the alignment between two assessments. As described in the Design Document (Appendix A), the four alignment criteria (Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation) are as applicable to judging the degree of alignment between two assessments as they are in judging the degree of alignment between an assessment and curriculum standards. What is different for the assessment-to-assessment comparison are the decision rules used to describe what acceptable alignment is. Since EXPLORE is a domain-sampled test, it may be reasonable for any one form to have only one or two items for any one CRS strand or to cover a low percentage of standards under a strand. Another difference in this study is the large difference in the number of items of the NAEP assessment, which uses matrix sampling (153 items), and the number of items on each EXPLORE form (30 items). It cannot be

expected that the two assessments would cover the same content range in all of the content domains of knowledge. The methodology examines the similarities and differences in content assessed by each test by considering the relationship of each to two different descriptions of performance, the NAEP framework and the CRS, enabling the findings to be grounded in more than one perspective of the content domain.

Comparison of NAEP with the Two Frameworks

Based on the summary results in Table 24, the NAEP assessment relative to the NAEP Framework and the CRS were fully aligned. The NAEP assessment compared to the standards in each of these analyses were found to have 100 percent alignment for three of the four alignment criteria (Table 24). The Balance of Representation was low for the Number Properties and Operations content area under the NAEP Framework and for five of the seven CRS strands. Neither of the low Balance indices are considered a major alignment issue since the minimum level on the other three alignment criteria were fully met. Over-emphasizing one or more standards does not have a detrimental impact on the Range-of-Knowledge addressed or the content complexity represented in the test. Even though the NAEP assessment had over 20 items (14 percent) that did not map precisely to any of the standards in the CRS (with these 20 items only mapping to generic objectives in CRS), the assessment did have a sufficient number of items that did precisely map to a sufficiently large number of standards for each of the CRS strands to have strong alignment.

Comparison of EXPLORE with the Two Frameworks

The alignment between the EXPLORE assessment and the NAEP Framework was weak. An important factor in this finding is the low number of assessment items (30 items per form) in relationship to the large number of NAEP objectives (N=101). An EXPLORE assessment form of 30 items only had six or more corresponding items for the Number Properties and Operations content area under the NAEP Framework. The other four content areas had five (Algebra) or fewer corresponding items. The EXPLORE assessment, even when the composite of two forms were considered, also had insufficient coverage of content of the NAEP Framework as indicated by the low Range-of-Knowledge Correspondence. The items on the EXPLORE assessment were consistent with the NAEP objectives on Depth-of-Knowledge, but the coverage of content over

the full domain expressed in the NAEP Framework was too sparse for the EXPLORE assessment and the NAEP Framework to be considered fully aligned.

The alignment between EXPLORE and the CRS was moderate when the analysis was done with four categories (pre-algebra, elementary algebra, geometry, and statistics/probability) rather than the seven strands from the CRS. The Categorical Concurrence and Range-of-Knowledge depends in part on the grain size of the content framework used in the analysis. Normally in alignment studies of state curriculum standards and state assessments, the reporting categories for the set of standards are used in the analysis and for reporting the results. In this study, the CRS framework with seven strands was used because it represented what was considered for the EXPLORE assessment as the document that most closely incorporated the features of a standards framework. However, ACT indicated in its documents four general content areas used to classify assessment items (ACT, 2013). These four general content areas represent content divisions used to stratify sampling of the larger domain when constructing test forms. None of the ACT documents indicated how the four content categories (pre-algebra, elementary algebra, geometry, and statistics and probability) were related to the seven CRS strands, particularly for pre-algebra and algebra. In reporting the Categorical Concurrence results, only one of the seven strands was found to have more than six corresponding items. This indicates there is low alignment between the CRS and EXPLORE, but this conclusion is somewhat unfair because the seven strands are not reporting categories for EXPLORE. EXPLORE reports a single mathematics score, representing overall mathematics achievement for an individual student. In this analysis, when using the four content categories and assuming that the three strands (GRE, PPF, and MEA) would be combined under geometry and that three strands (BOA, NCP, and XEI) would be evenly divided between pre-algebra and algebra, the alignment appears to be much improved and acceptable relative to the alignment criteria. Only the general area of statistics and probability (the PSD strand) would not have an acceptable level for Categorical Concurrence and a marginal Range-of-Knowledge Correspondence. The assessment only had four items corresponding to the PSD strand, below the six items used as the threshold level in this study. Still, the assessment items and the CRS framework would be considered aligned with respect to Depth-of-Knowledge, Range-of-Knowledge, and Balance of Representation.

Summary: Comparison of NAEP and EXPLORE

The content alignment results indicate that the 2013 NAEP Grade 8 Mathematics assessment and the EXPLORE assessment have a moderate level of alignment to each other. The set of items on each assessment targeted the same general content. EXPLORE covered the same content as the NAEP assessment to a lesser degree, in part, because it had one-fifth of the number of items on the NAEP assessment. As noted, one reason for the larger number of items on NAEP is the larger number of objectives to be addressed by the assessment and the matrix sampling design used to report the performance of student groups – hence, each NAEP mathematics examinee encounters a subset of the full NAEP item pool of 153 items, whereas each EXPLORE Mathematics test had 30 items and reported on the performance of individual students. Both assessments had similar proportions of items that corresponded to either the five content areas on the NAEP Framework or the seven strands on the CRS with only a few exceptions. EXPLORE had a higher proportion of items that targeted number and operations while the NAEP had a higher proportion of items that targeted algebra and data analysis, statistics, and probability. The NAEP Mathematics assessment and, to some degree, EXPLORE had at least some items in geometry, measurement, and data analysis. Topics such as lines of symmetry, similar figures, use of Pythagorean theorem, relationship between polygonal plane figures, coordinate geometry, mathematical reasoning in geometry, and sampling and experiments were assessed in part by NAEP but were not found on the two EXPLORE forms analyzed. EXPLORE assessed identifying, defining or describing geometric shapes whereas the NAEP did not. Considering the frameworks, the CRS content statements were expressed more generally than by the NAEP framework. As such, the NAEP assessment had some redundancy with respect to the CRS.

Both assessments had an acceptable Depth-of-Knowledge Consistency with their parent framework and each other's framework. More than half of the items on each assessment had a consistent DOK level with the corresponding objective or standard to which the item was coded. Even though the NAEP assessment had 38 constructed response items, the average DOK level on both the NAEP assessment and EXPLORE forms was essentially the same. The NAEP assessment with the NAEP Framework had a higher DOK Consistency on four of the five content areas than EXPLORE mapped to the NAEP Framework. The EXPLORE assessment had noticeably lower DOK Consistency than the NAEP assessment on the NAEP content areas of

measurement and data analysis/statistics/probability. Both assessments had the same DOK Consistency on all seven strands of the CRS except the NAEP assessment was more consistent in the CRS strand of Properties of Plane Figures (PPF) than EXPLORE. With one-fourth of its items in a constructed response format, the NAEP assessment approached measuring content knowledge in a wider variety of ways compared with EXPLORE. The NAEP assessment also had two or three items that were considered a DOK level 3 by most panelists whereas all panelists agreed that all EXPLORE items were classified at the DOK 1 or 2 level. The framework analysis indicated that the NAEP Framework used a boarder range of verbs, including such ones as analyze and verify, when compared to the CRS. There were, however, some verbs in the CRS not included in the NAEP Framework, i.e., locate, compute, manipulate, exhibit, substitute, and work with. Overall, the NAEP assessment had a slightly higher DOK Consistency than EXPLORE when analyzed with either framework, with only a few exceptions (e.g., geometry with the NAEP Framework).

The items on both assessments were distributed fairly evenly across the 101 objectives in the NAEP framework as indicated by the Balance of Representation, but across the 56 standards in the CRS, the NAEP items stacked up more on one or two of the standards under five of the seven strands rather than being spread more evenly over the underlying standards.

The analysis results addressed the three key research questions for this study. First, nearly all of the major topics targeted by the EXPLORE forms were also addressed by the NAEP assessment items. However, the NAEP assessments attended more to particular topics in measurement; geometry; and data analysis, statistics and probability. Both assessments targeted similar topics in number and algebra. Even though the NAEP assessment included constructed-response items, the average DOK level of items on both the NAEP assessment and EXPLORE was essentially the same.

Second, the two assessments had similar proportions of items that corresponded to the five content areas on the NAEP Framework and the seven strands on the CRS. However, there were several differences. EXPLORE had a higher proportion of items that targeted number and operations while the NAEP had a higher proportion of items that targeted algebra and data analysis, statistics, and probability. The NAEP Mathematics assessment covered about three

times the content of the EXPLORE assessment when compared to the NAEP Framework objectives. When compared to the CRS, the NAEP covered about twice the amount of content of the EXPLORE assessment. A composite of the two EXPLORE forms yielded increased content coverage, but still not at the breadth of the NAEP assessment. This result may also reflect the large difference in the number of items between the two assessments.

Third, regarding significant differences between the two assessments, no large differences represented by major holes within a content domain were found between the NAEP assessment and EXPLORE forms other than the sizeable difference in the number of items. The two mathematics assessments differed in degree rather than substance.

In sum, considering all four alignment criteria, there is a moderate degree of alignment between the NAEP Mathematics assessment and EXPLORE for mathematics. The NAEP assessment, however, covered more content and was more consistent in matching expectations with regards to content complexity. In particular, the analysis showed that the NAEP Framework and the assessment had more content coverage of topics in measurement, geometry, and data analysis, statistics and probability. Both assessments targeted similar topics in number and algebra. Nearly all of the topics targeted by EXPLORE forms were also addressed by the NAEP assessment items. Even though the NAEP assessment included constructed-response items, the average DOK level of items on both the NAEP assessment and EXPLORE was essentially the same.

References

ACT, Inc., 2009. *Connecting College Readiness Standards to the Classroom for Mathematics Teachers*. Iowa City, IA: Author.

ACT, Inc., 2011. *Your Guide to EXPLORE: What It Measures, Its Purposes and Foundations, How It is Developed*. Iowa City, IA: Author. Downloaded October, 2014, http://www.act.org/explore/pdf/YourGuideEXPLORE.pdf

Cronbach, L. J. 1971. "Test validation." In R. L. Thorndike (Ed.), *Educational measurement* (2nd edition), Washington, DC: American Council on Education, 443-507.

Fields, R. 2014. *Towards the National Assessment of Educational Progress (NAEP) as an Indicator of Academic Preparedness for College and Job Training*. Washington, DC: National Assessment Governing Board.

Horizon Research, Inc. (2012) National Survey of Science and Mathematics Education. Chapel Hill: Author. http://www.horizon-research.com/2012nssme/wp-content/uploads/2013/02/2012-NSSME-Chapter-2.pdf

Kane, M. (2006). Validation. In R. L. Brennan (Ed), *Educational measurement* (4th edition). Washington, DC: American Council on Education/Praeger, 17-64.

Messick, S. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23(2): 13-23.

Moss, P. A. 1992. "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research* 62(3): 229-58.

National Assessment Governing Board (no date). Technical Report: NAEP 12th grade preparedness research. Washington, D.C., downloaded March 15, 2015 (https://www.nagb.org/what-we-do/preparedness-research.html).

National Assessment Governing Board (2009). *Design of Content Alignment Studies in Mathematics and Reading for 12th grade NAEP and other Assessments to be Used in Preparedness Research Studies*. Washington, D.C. Downloaded March 15, 2015.

National Assessment Governing Board (NAGB), 2010. *NAEP 2011 Mathematics Framework for the National Assessment of Educational Progress*. Washington, D.C: U.S. Government Printing Office. Downloaded October 2014. https://www.nagb.org/publications/frameworks/mathematics/2011-mathematics-framework.html

National Assessment Governing Board (NAGB), 2012. *Mathematics Framework for the 2013 National Assessment of Educational Progress*. Washington, D.C: U.S. Government Printing Office. Downloaded October, 2014. https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/mathematics/2013-mathematics-framework.pdf

National Center for Education Statistics (2012) Schools and Staffing Survey, 2001-12. Washington, DC: U.S. Department of Education, downloaded May 2015. http://nces.ed.gov/surveys/sass/tables_list.asp#2012

NORC, 2014. *A Comparison of NAEP Grade 8 Mathematics Framework and ACT EXPLORE College Readiness Standards for Mathematics*. Chicago, IL, author.

Shrout, P. E., and J.L. Fleiss. 1979. "Intra-class Correlations: Uses in Assessing Rater Reliability." Psychological Bulletin 86(2): 420-28.

Subkoviak, M. J. 1988. "A Practitioner's Guide to Computation and Interpretation of Reliability Indices for Mastery Tests." Journal of Educational Measurement, 25(1): 47-55.

Webb, N. L. 1997. Criteria for Alignment of Expectations and Assessments in Mathematics and Mathematics Education. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research. Downloaded March 16, 20 15.

Webb, N. L. (2002). An analysis of the alignment between mathematics standards and assessments for three states. Paper presented at the American Educational Research Association Annual Meeting, New Orleans, Louisiana, April 1-5.

Webb, N. L. Identifying content for student achievement tests. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publisher, pp. 155 -180, 2006.