FINAL REPORT

Alignment Between the 2013 NAEP Grade 8 Reading Assessment and the ACT EXPLORE Reading 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 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 reading content domains assessed by NAEP and EXPLORE?
2. To what extent is the emphasis of reading 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 reading 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. A 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 Reading 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 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. EXPLORE is a domain-sampled assessment. 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 Reading 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 reading 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 Reading Framework was used as the description of the content on the 2013 NAEP Grade 8 Reading assessment. The assessment included both literary and informational texts and assessed understanding according to three levels of cognitive targets:

Locate/Recall
Integrate/Interpret
Critique/Evaluate

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. CRS are used as the best available description of the content on the EXPLORE assessments. CRS categories were designed to communicate to educators and align to reading skill targets. 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 Reading used in this study were released in 2005 and have five strands:

MID: Main Ideas and Author's Approach
SUP: Supporting Details
REL: Sequential, Comparative, and Cause-Effect Relationships
MOW: Meanings of Words
GEN: Generalizations and Conclusions

A reading education content expert conducted an analysis of the frameworks for each of the reading assessments. The similarities and differences between the two frameworks were identified through these analyses and used to inform the training of panelists at the CAI. For reading, the CRS were found to distinguish between Uncomplicated and Complicated text whereas the NAEP Reading Framework did not; both frameworks varied in text passages selection with EXPLORE's giving more emphasis to a range of texts from across subject areas (literary prose, humanities, and social studies) and the NAEP Framework including a wider

range of passage types (e.g., poetry); the NAEP Framework included cross-text comparisons that were not mentioned in the ACT CRS; and the NAEP Framework explicitly expected a certain percentage of items to have a cognitive target of Critique/Evaluate whereas corresponding explicit standards targeting argumentative texts were not found in the CRS. (Argumentative text is addressed in other test documentation for EXPLORE.) 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 reading content domains assessed by each assessment; the emphasis of reading 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 and EXPLORE assessments 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 NAEP 2013 Reading Framework for grade 8 and assessment were analyzed in the Institute by the panelists. The 2013 NAEP Grade 8 Reading assessment is comprised of an item pool of 163 items, which were used in the content analysis. The NAEP Reading assessment took under

60 minutes for a student to complete, and was administered in "blocks" of items that differed from booklet to booklet. The NAEP Reading assessment for grade 8 had items written in three different formats: multiple-choice, short constructed-response, and extended constructed-response. The assessment 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 four points. The Institute analysis weighted the NAEP items by point value giving a total of 268 points for the NAEP Reading 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 EXPLORE Reading 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 a larger pool of items from the reading domain. Equivalence of forms was 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 Reading assessment is administered along with the English, Mathematics, and Science assessments. The complete set of assessments 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 Reading assessment only encounters a subset of the full NAEP reading item pool of 163 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 reading 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 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 Reading 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 Reading assessments and the NAEP Reading Framework was found to have low values on the four alignment criteria. The EXPLORE assessment had sufficient items to map well to two of the three NAEP categories, but a majority of the panelists did not find any items that mapped to the third category, Critique/Evaluate. The items that panelists mapped to Locate/Recall and Integrate/Interpret were consistent in the content complexity with the corresponding expectations in the NAEP Framework, but primarily targeted the objectives assigned a DOK 2 under Integrate/Interpret rather than the more complex objectives under this category. The ratings of Range of Knowledge

and Balance of Representation were good for the Locate/Recall category but were rated low for the Integrate/Interpret category. Only 25 percent or fewer of the underlying objectives had corresponding items for this category and the objective on making simple inferences was overemphasized. Panelists were not given precise instructions to differentiate between simple and complex inferences. Rather they were expected to follow the process and differentiate between these inferences by their assignment of DOK levels to the items. Range of Knowledge was not improved when the composite of the two EXPLORE Reading forms or a total of 60 items (composite analysis of two forms with 30 items each) was used. This comparison should be interpreted in light of the fact that since the NAEP uses matrix sampling, no single student ever takes all of the 163 items.

The alignment between the 2013 NAEP Grade 8 Reading assessment and the CRS was found to have mixed values on the four alignment criteria. The alignment was below 50 percent on Range-of-Knowledge Correspondence for one strand (REL) and below an index value of 0.70 for Balance of Representation for two strands (MID and MOW). The panels agreed on most criteria except one panel coded more items to the CRS strand Sequential, Comparative, and Cause-Effect Relationships (REL) while the other panel coded more items to the CRS strand Generalizations and Conclusions (GEN). What one panel interpreted as drawing a subtle generalization, the other panel interpreted as drawing a relationship, e.g., cause and effect. Both panels found NAEP items with over 13 point values that mapped to each of the five CRS content strands. The DOK Consistency was high for all five strands. Range of Knowledge was acceptable for four of the five strands. Panelists from both groups found items that targeted 50 percent or fewer of the standards under the CRS strand REL. The NAEP Reading assessment overemphasized one or two standards under two of the strands, CRS strands Main Ideas and Author's Approach (MID) and Meanings of Words (MOW). Similar to the alignment with the NAEP Framework, the NAEP assessment overemphasized the standards in the CRS that related to using context to determine the meaning of words.

The alignment between the 2013 NAEP Grade 8 Reading assessment and the NAEP 2013 Reading Framework was found to have high values for each of the four alignment criteria. Each of the three NAEP reading reporting categories had more than 25 point values (Locate/Recall; Integrate/Interpret; Critique/Evaluate). From 60 to 70 percent of the point values

were found to target the Integrate/Interpret category. The items had high DOK Consistency with the DOK levels of the assigned objectives, over 70 percent agreement. Items on the assessment targeted over 50 percent of the underlying objectives for each of the three reporting categories. The items were evenly distributed among the objectives for two of the categories, but emphasized the objective of interpreting the meaning of a word as used in text in the Integrate/Interpret category. This was purposefully done to have at least two vocabulary items for each of the 19 passages. This overemphasis on vocabulary items was not considered a major alignment issue because all of the other three alignment criteria were acceptably met. The two panels had good agreement in the summary results for each of the four alignment criteria and only differed in coding nine items to Integrate/Interpret and Critique/Evaluate.

The alignment between EXPLORE and the CRS was found to have mixed values on the four alignment criteria. Panelists mapped sixty percent or more of the items on each form to only one of the five strands, the CRS strand Supporting Details (SUP). The other four strands had an average across panels of fewer than five items. Having 30 items on a form, each worth one point, would make it difficult to have a large number of items for each of five strands if one strand was assigned over 60 percent of the items. Panelists only found one or two items that mapped to at least one of the five strands on both forms. The items compared favorably in DOK Consistency for both forms. However, the Range of Knowledge was below 50 percent for four strands with EXPLORE Form 1 and three strands with EXPLORE Form 2. Range of Knowledge was improved slightly when the composite of two forms was considered by having 50 percent or more of the standards under a strand with at least one corresponding item by both panels on one strand, by one panel on two strands, and by neither panel on two strands. Balance of Representation was generally good for all five strands, but this criterion is not very meaningful because of the low number of items for all except the SUP strand. The results for each of the forms were nearly the same as coded by both panels.

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 Reading and EXPLORE Reading assessments addressed many of the same topics, but not with the same concentration. The biggest difference found between the two assessments was the 23 percent of the NAEP point values, compared to no items on the EXPLORE forms, that targeted objectives under the Critique/Evaluate reporting category. This category represented more complex topics such as argumentation. The EXPLORE assessment had a greater proportion of its items than the NAEP assessment that targeted content under the Locate/Recall category. When the assessments were mapped to the CRS, panelists mapped nearly two-thirds of the EXPLORE Reading assessment to the CRS strand Supporting Details (SUP). The NAEP assessment only had about 25 percent of its point values that mapped to the CRS strand SUP, about the same percentage of point values that were found to map to the CRS strands Generalizations and Conclusions (GEN) and Meanings of Words (MOW). Thus, the distribution of items was different between the two assessments. The NAEP assessment included items targeting more complex topics (critiquing, evaluating, and forming generalizations) whereas EXPLORE placed greater emphasis on determining supporting details. Both assessments expected students to find the meaning of words in context and placed the lowest emphasis on sequential, comparative, and cause and effect relationships (CRS strand REL).

Depth-of-Knowledge Consistency. The NAEP Reading assessment had a higher average DOK level of items than did the EXPLORE assessment forms, 2.7 compared to 1.5 average DOK (levels 1 to 4), based on the panelists' item mappings. The NAEP assessment items tended to be DOK levels 2 or 3 whereas the EXPLORE assessment items all were DOK levels 1 or 2. The majority of reviewers coded 25 items on the NAEP assessment with a DOK 3 (15 percent). Even with the difference in the average DOK level between the two assessments, both assessments had reasonably high DOK Consistency with the corresponding objectives or standard.

Range-of-Knowledge Correspondence. The NAEP assessment covered more content than the EXPLORE assessment. An important reason for this was the large difference in the number of items between the two assessments—163 items on NAEP used to report on population performance and 30 items on EXPLORE used to report on individual performance. The NAEP

assessment had an acceptable Range of Knowledge on all three of the NAEP Framework reporting categories and four of the five CRS strands. EXPLORE only had an acceptable Range of Knowledge on the Locate/Recall NAEP reporting category and three of the five CRS strands. Neither assessment had high coverage of the CRS strand REL, and the EXPLORE forms had low coverage of the CRS strand GEN. Overall, the NAEP Reading assessment targeted a greater breadth of content in nearly all topics with the exception of forming relationships among ideas (sequential, comparative, and cause-and-effect). The EXPLORE reading forms were low in coverage on this topic, but also on integrating ideas, making generalizations, critiquing, and evaluating.

Balance of Representation. The items on both assessments were distributed fairly evenly among the framework objectives and standards. The one exception was the increased emphasis placed by the NAEP Reading assessment on determining the meaning of a word in context.

Assessment to Assessment Alignment

Research Question 1 (Correspondence between content). The NAEP Reading assessment and EXPLORE Reading had a large overlap in content coverage. EXPLORE, however, did not cover the more complex topics normally labeled as critiquing, evaluating, and generalizing. Otherwise, the two assessments aligned well in locating and recalling information with literary and informational texts and the less complex aspects of integrating and interpreting of text.

Research Question 2 (Proportionality between content). The two assessments differed in the proportion of items given to topics. About 15 percent of the NAEP Reading assessment items had an average DOK level 3 computed across all 16 panelists indicating these items were judged by the majority of the panelists to require more complex reasoning about the text, deeper inferences about the meaning from the text, and drawing inferences across tests. Panelists did not assign any of the items from the EXPLORE forms as a DOK level 3. Based on the panelists' item-to-framework mappings, EXPLORE placed more emphasis on locating details, making simple inferences, and identifying textually explicit information.

Research Question 3 (Systematic differences between content). Only one large difference was found between the NAEP and EXPLORE Reading assessments in that no items on the

EXPLORE forms corresponded to objectives under the NAEP Reading assessment's Critique/Evaluate reporting category. Otherwise, the two assessments differed mainly in the degree of emphasis and the level of complexity of the items.

Conclusions

In summary, the analysis results indicate moderate alignment at a very general level between the NAEP and EXPLORE reading assessments. The two assessments addressed similar content topics, but they differed on the degree of concentration of items on the topics. There was a marked difference in the content complexity of the two assessments. The NAEP Reading assessment had an average DOK level of items of 2.7 as compared to 1.5 average DOK for EXPLORE reading assessment (levels 1 to 4). The NAEP assessment items tended to be DOK levels 2 or 3 whereas EXPLORE items tended to be DOK levels 1 or 2. The NAEP Reading assessment with 163 items targeted over 50 percent of content descriptors in nearly all topics with the exception of forming relationships among ideas (sequential, comparative, and cause-and-effect). The EXPLORE Reading assessment with 30 items per form targeted over 50 percent of content descriptors for three CRS strands and the Locate/Recall NAEP reporting category. EXPLORE targeted fewer than one-third of the content descriptors for two NAEP reporting categories and two CRS strands. These were the content topics related to integrating ideas, making generalizations, and the NAEP Critique/Evaluate reporting category.

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 reading teachers that were highly qualified and experienced, and as a group were representative of the population of reading 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 Reading Framework and the CRS performance descriptors for reading were used in this study. The key difference between the two documents used in this study is the purpose, i.e., why and how they were developed. The NAEP Reading Framework was developed to guide the item writing and construction of a comprehensive assessment 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 (163 items), and the number of items on each EXPLORE form (30 items). It cannot be expected that the two assessments would cover the same degree of 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. 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 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 ACT EXPLORE assessments in reading and mathematics. The purpose of this contract from the National Assessment Governing Board is to evaluate the extent to which 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 reading content domains assessed by NAEP and EXPLORE?
2. To what extent is the emphasis of reading 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 reading 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 the 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; Appendix A). 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 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 Reading Framework, is referred to as a framework for convenience. When a specific reading framework is considered, the framework 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 domain. 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 Reading assessment only encounters a subset of the full NAEP item pool of 163 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 Reading assessment to the EXPLORE Reading assessment 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 reading are included as Appendices H.6 and H.7.

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-of-Knowledge Correspondence 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 reading content domains assessed by each assessment; the emphasis of reading 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 Reading Framework and the ACT College Readiness Standards (CRS) for EXPLORE Reading. 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 reading content domain assessed by NAEP and that were assessed by EXPLORE?
- To what extent is the emphasis of reading 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 (CRS)? Are these differences such that entire reading 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 reading considered the

similarities and differences in the included cognitive targets, types of texts (e.g., literary, informational), and content strands. The NAEP framework listed content topics under three general categories and included specifications for the percentage of items by content complexity. The NAEP clustered items under cognitive targets. The CRS organized standards representing four score ranges under five content strands. The guide to the assessment (ACT, Inc., 2001, p. 11) listed the percent and number of items for each form of EXPLORE under three content areas (Prose Fiction, Humanities, and Social Sciences). The ACT, Inc., listed items under five content strands (e.g., main ideas and author's approach; supporting details; sequential, comparative, and cause-effect relationships; meaning of words; and generalizations and conclusions). 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 reading was conducted Dr. Karen Wixson (University of North Carolina—Greensboro). 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 Reading Framework Analysis Report can be found in Appendix B of this report.

Content Alignment Institute

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 grade 8 Mathematics and Reading assessments and the 2013 EXPLORE Mathematics and Reading assessments. 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 be 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, NAEP 2013 Reading Objective 2.3.b (Make complex inferences within and across texts to draw conclusions and provide supporting information) and Objective 2.3.c (Make complex inferences within and across texts to find evidence in support of an argument) both target finding supporting evidence or information by making inferences within and across texts. One panelist may appropriately decide that the main content knowledge to answer an item correctly requires complex inferences to “provide supporting information” (2.3.b) while another panelist may just as appropriately decide that the assessment item requires students to make complex inferences to “find evidence in support of an argument” (2.3.c). Similarly, the CRS MID 301 (Identify a clear main idea or purpose of straightforward paragraphs in uncomplicated literary narrative) may overlap with MID 401 (Infer the main idea or purpose of straightforward paragraphs in uncomplicated literary narratives), depending on the extent of inference judged to be required to answer a particular assessment item. 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 Reading Framework included 46 cognitive targets (objectives) and the CRS for EXPLORE Reading included 39 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; and
- 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 are White, non-Hispanic, 7 percent are Black, and 8 percent are Hispanic (NCES, 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 two were Ph.D. graduate students (both with teaching experience) and one was a university professor of reading education.

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 Reading Framework (NAGB, 2012) was used to create the 2013 NAEP Grade 8 Reading assessment. The assessment includes both literary and informational texts and assesses understanding according to three levels of cognitive targets:

- Locate/Recall
- Integrate/Interpret
- Critique/Evaluate

The same types of texts and the same three levels of cognitive targets are used to structure the assessments for all three grades—4, 8, and 12. Text sophistication, however, increases by grade. Objectives are identified according to two dimensions: by cognitive target level and by whether they target literary text, informational text, or both. The NAEP Reading Framework was written to guide the development of the main NAEP assessments at the national, state, and district levels. The NAEP Reading Framework was designed to specify what reading skills should be assessed by NAEP at grades 4, 8, and 12. The framework is not intended to be a curriculum framework for guiding instruction. All of the 46 grade 8 objectives to be assessed were included in this analysis. Panelists mapped the assessment items to the objectives identified within the NAEP 2013 Reading Framework's cognitive targets matrix.

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 Reading assessments used in this study were released in 2005, and the Standards relevant to EXPLORE are organized into five strands:

MID: Main Ideas and Author's Approach

SUP: Supporting Details

REL: Sequential, Comparative, and Cause-Effect Relationships

MOW: Meanings of Words

GEN: Generalizations and Conclusions

The CRS apply to three different types of text used in the EXPLORE Reading assessment: prose fiction, humanities, and social science passages. The CRS for EXPLORE Reading consists of 39 standards across the five strands, divided into four score ranges—13-15, 16-19, 20-23, 24-25. ACT documents state that although the strands overlap, each standard has been assigned to a primary strand. 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 Reading Frameworks

A reading content expert conducted an analysis of the CRS reading framework and the NAEP Reading Framework. The similarities and differences between the two frameworks were identified through these analyses and used to inform the training of panelists at the CAI. For reading, the CRS were found to distinguish between Uncomplicated and Complicated text whereas the NAEP Reading Framework did not; both frameworks varied in text passages selection with EXPLORE's giving more emphasis to a range of texts from across subject areas (literary prose, humanities, and social studies) and the NAEP Framework including a wider range of passage types (e.g., poetry); the NAEP Framework included cross-text comparisons that were not mentioned in the CRS; and the NAEP Framework explicitly expected a certain percentage of items to have a cognitive target of Critique/Evaluate whereas corresponding explicit standards targeting argumentative texts were not found in the CRS. (Argumentative text is addressed in other test documentation for EXPLORE.)

Assessments

NAEP

The 2013 NAEP Grade 8 Reading assessment 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 assessment and frameworks were analyzed for alignment. Students were given 50 minutes to complete the NAEP Reading test, which included two 25-minute blocks of items. Blocks

differed from booklet to booklet. One block generally consisted of one passage and its items. The number of items for a passage ranged from five to 11. Any one student only took two blocks. From this application of matrix sampling, inferences are made to the full population. No information is provided on individual students. The number of items and the amount of time provided 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 the two types of reading (literary and informational) for the total population and subpopulations.

The 2013 grade 8 NAEP Reading assessment included 163 items divided among 19 passages (between 400 and 1000 words in length), two general types of texts, three levels of complexity, and three item designs. Of the 19 passages and corresponding items, seven were administered only on the grade 8 assessment, eight were administered on both the grade 4 and grade 8 assessments, and four were administered on both the grade 8 and grade 12 assessments. Of the passages on the grade 8 assessment, according to the framework, 45 percent of the passages were to be literary and 55 percent of the passages were to be informational (NAGB, 2012, p. 11). The passages were selected to represent the type of texts students would experience both within and out of school. Some of the items even required integrating information across a pair of texts to assess students on the authentic task of reading and comparing multiple texts. It should be noted, however, clustering items by passage in itself should not have an impact on alignment as long as there is an appropriate variation of items from passage to passage.

Students were assessed on comprehension as well as meaning vocabulary, defined as the application of one's understanding of word meanings to passage comprehension (NAGB, 2012, p. 12). In multiple passages in the NAEP Reading assessment, students were expected to draw on their vocabulary knowledge. Gaining meaning of most words used in a paragraph or passage was viewed as a necessary condition for comprehension. Items devoted to vocabulary words were to target words linked to the central ideas of the passage. A student's understanding of the meaning of these words was to be assessed in the context of the passage. Students were not to be asked to draw upon prior knowledge of the definition of a word by selecting or producing such a definition. The vocabulary items could be multiple-choice or short constructed-response in format.

The 2013 NAEP grade 8 reading assessment had items written in three different formats: multiple-choice with 4 choices (40 percent), short constructed-response (45 percent), and extended constructed-response (15 percent). The assessment was divided in testing time between multiple-choice and the two types of constructed-response according to the percentages given above. Multiple-choice items were assumed to take students one minute to complete while short constructed-response items were assumed to take a student two to three minutes and extended constructed-response were assumed to take a student approximately five minutes to complete. Some short constructed-response items were scored at two levels: correct or incorrect. Other short constructed-response items were scored at three levels: correct, partially correct, or incorrect. Extended constructed-response items had multiple parts and required more than a short answer. They were scored at up to four levels. Scoring rubrics were used for all constructed-response items (NAGB, 2009, Chapter 3, p. 31).

The NAEP 2013 Reading Framework (NAGB, 2012) specified that the items on the assessment should be distributed among the three cognitive targets:

Locate/Recall	20 percent
Integrate/Interpret	50 percent
Critique/Evaluate	30 percent

Items corresponded to both literary (45 percent) and informational (55 percent) texts.

EXPLORE

Two forms of the EXPLORE Reading assessment were used in this study. Each form of EXPLORE had 30 multiple-choice items. 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 content areas, the cognitive distribution of the items, and the match to a difficulty distribution, as well as through fine-tuning using equivalent-population equating. The reading test was 30 minutes long and was administered along with the English, mathematics, 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.

Items either required referring or reasoning and applied evenly to three different domains of reading passages: prose fiction (e.g., short story and fiction), humanities (e.g., memoirs and personal essays, and in the content areas of architectures, art, dance, ethics, film, language, literary criticism, music, philosophy, radio, television, or theater), and social science (e.g., anthropology, archaeology, biography, business, economics, education, geography, history, political science, psychology, or sociology). Each of the three passages had 10 multiple-choice items. There is generally one item per passage that asks a student to define a word based on context.

Passages on EXPLORE are characterized as being uncomplicated and challenging (ACT, Inc., 2011, p. 12). Uncomplicated literary passages refer to excerpts from essays, short stories, and novels that tend to use simple language and structure, have a clear purpose and a familiar style, present straightforward interactions between characters, and employ only a limited number of literary devices such as metaphor, simile, or hyperbole. More challenging literary passages refer to excerpts from essays, short stories, and novels that tend to make moderate use of figurative language; have a more intricate structure and messages conveyed with some subtlety; and may feature somewhat complex interactions between characters. Uncomplicated informational passages refer to materials that tend to contain a limited amount of data, address basic concepts using familiar language and conventional organizational patterns, have a clear purpose, and are written to be accessible. More challenging informational passages refers to materials that tend to present concepts that are not always stated explicitly and that are accompanied or illustrated by more—and more detailed—supporting data, include some difficult context-dependent words, and are written in a somewhat more demanding and less accessible style.

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 reading, a large battery of items (N=163) 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 Reading assessment only encounters a subset of the full NAEP item pool of 163 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. One block generally consisted of one passage and its items. The number of items for a passage went from five to 11. Items were multiple-choice, short constructed-response (answered by one or two phrases or a sentence), and extended constructed-response (answered by one or two paragraphs). The point value for a correct response on a NAEP item went from one to four points. 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., 2013, p. 1).

The EXPLORE Reading assessment was designed to measure a student's level of reading comprehension as based on their skill in referring and reasoning. Assessment items required students to derive meaning from several texts by referring to what is explicitly stated and reasoning to determine implicit meanings. Items required students to use referring and reasoning skills to determine main ideas; locate and interpret significant details; understand sequences of events; make comparisons; comprehend cause-effect relationships; determine the meaning of context-dependent words, phrases, and statements; draw generalizations; and analyze the author's or narrator's voice or method (ACT, Inc., 2013, p. 6). Several forms were created, each with three prose passages representative of those that 8th and 9th graders would normally

experience. One passage was on social studies, one on prose fiction, and one on humanities. 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 Reading, a student took one form with three reading passages and 30 multiple-choice items, each assigned a score of one point and distributed 10 items per passage. The results of the EXPLORE Reading assessment, along with the English, Mathematics, and Science assessments, 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 each of these two assessments, the content represented in the framework for each assessment varied. The NAEP 2013 Reading 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 Framework, as used in this study, was organized by three cognitive levels: 1) Locate/Recall; 2) Integrate/Interpret; and 3) Critique/Evaluate. Under each of these cognitive levels, more specific statements of cognitive targets were listed by the type of text they were applicable to: both literary and informational text; literary text; and informational text. 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. The CRS organized knowledge and skills descriptions into five strands; within each strand, skills were grouped into four score-bands that corresponded to the ability level (as determined by the overall EXPLORE Reading score) of the student. Content standards were specified under each of the CRS strands to represent the performance of 8th and 9th graders at specific score ranges on the EXPLORE assessment (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 frameworks. Notably EXPLORE had 30 items for each form and the NAEP Framework had 46

cognitive targets. From the outset it was not clear how the two assessments would 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 163 reading items. The NAEP Reading Framework had 46 cognitive targets and the CRS had 39 standards. 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 46 cognitive targets clustered under three content areas, it would be unlikely to have one item mapped to each objective unless items were robust and could target more than one of the cognitive targets. This is different from when the NAEP assessment is mapped to the NAEP framework. The 163 items on the NAEP assessment could be distributed so that nearly each of the 46 cognitive targets would have three corresponding items. Both reading assessments had enough items to have the minimum number of items to satisfy the threshold level for each of the four alignment criteria. A primary difference between the two reading assessments is that the NAEP assessment has a larger number of items for mapping to any one of the NAEP cognitive targets or CRS standards.

The data from the four studies can be used to determine the correspondence between the reading content domain of each assessment; the emphasis of reading content given by each assessment; and any systematic differences between the two assessments. If two assessments have exact agreement on the reading content domain, then there will be a strong correspondence when the two assessments are mapped to one framework, and a similar correspondence when the two 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 NAEP and EXPLORE and the content knowledge and skills unique to each

assessment. Because of the big differences in the number of items on each assessment, it is likely that 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 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 in the sample of the content domain represented on the EXPLORE assessment partially overlaps with the content covered by the NAEP assessment. In this case, there is a common set of content knowledge assessed by both assessments. In addition, each assessment targets 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 reading frameworks address the same general reading 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 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 Reading Framework and the CRS. The NAEP Framework included 46 objectives from the NAEP 2013 reading framework for grade 8. All of the NAEP Reading categories and objectives used in the analysis are listed in Appendix I. It should be noted that the NAEP Framework for reading does not label the expectations. In this report, these expectations will be referred to as objectives. The nomenclature for each objective was made up of three parts, and corresponded with those included in the framework preceded by a number indicating the cognitive process:

1. Locate/Recall
2. Integrate/Interpret
3. Critique/Evaluate

The second number indicates whether the objective applies to both literary and informational text (1), specific to literary text only (2), or specific to informational text only (3). Finally, a lower case letter identifies the objective. For example, 1.3.b represents the objective, "recognize rhetorical devices," found under content area of Locate/Recall (1), specific to informational text (3), and the second objective (b).

The reading framework directly states that "(t)he cognitive targets matrix is for illustrative purposes only and should not be considered an exhaustive list" (NAGB, 2012, p. 39, footnote 5). During the framework analysis, other statements of expectations were found mainly in the Achievement Levels that were used to fill in missing areas from the matrix. This was done by adding any statement under the Achievement Levels that did not have a corresponding expectation listed in the matrix. For example, the cognitive targets matrix does not explicitly note any expectations for vocabulary. The assessment of vocabulary is discussed in some detail in other parts of the framework document (NAGB, 2012, pp. 32-36) and is listed under the grade 8 Basic Achievement Level, "(t)hey should be able to interpret the meaning of a word as it is used in the text" (NAGB, 2012, p. 64). Language from the achievement levels then was used to supplement the cognitive targets matrix with Objective 2.1.d (interpret the meaning of a words as it is used in the text). A total of 17 statements were added to those provided in the cognitive targets matrix—1.3.b, 2.1.a, 2.1.b, 2.1.c, 2.1.d, 2.1.i, 2.1.j, 2.1.k, 2.1.l, 2.1.m, 2.2.f, 2.2.g, 2.2.h,

2.3.f, 2.3.g, 3.1.d, and 3.1.e. A judgment was made for each of the added statements as to the most appropriate mental process or cognitive target required and the type of text. For example, 1.3.b "recognize rhetorical devices" was placed under the general cognitive target of Locate/Recall and under both literary and informational text.

The analysis of the CRS included all of the 39 standards in that document (ACT, Inc., 2009) designated for EXPLORE. Numbers following the strand abbreviation —200s, 300s, 400s, and 500s—are used to identify a specific standard. All of the CRS Reading strands and standards used in the 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 know; those at the 400 level represent content knowledge and skills students who scored at the 20-23 level are likely to know, and those at the 500 level represent content knowledge and skills students who scored at the 24-25 level are likely to know. 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 describing their responsibilities as a facilitator (Appendix H.8). They were also sent the Framework Analysis for

their subject (Appendix B), other supporting materials such as the NAEP 2013 Reading 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, decision rules for reading include which types of passages qualify as literary vs. informational text; defining characteristics of "uncomplicated" vs. "more challenging" texts; and clarifying the difference between "simple" and "complex" inferences (see p.34 below).

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 Institute 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 Reading Framework for grade 8 and the CRS; and code 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 performed 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. At the CAI, the 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. 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 reading facilitators provided a few rules for assigning items to specific standards under the CRS. The reading facilitators indicated that simple and complex inferences were to be distinguished by the DOK level that was assigned. Simple inferences should be assigned a DOK level 2 and complex inferences should be assigned a DOK level 3. Panelists were told that the NAEP materials used the term “inference” in a similar way that the ACT materials used “generalization and conclusion.”

Panelists were not given precise instructions to differentiate between simple and complex inferences. Rather they were expected to follow the process and differentiate between these inferences by their assignment of DOK levels to the items. Panelists were instructed to use the NAEP classification of literary and informational texts.

Observers

National Assessment 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 framework 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 (CAI) 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 (DOK) 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 (CRS).

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 facilitators also provided a sample of assessment items to illustrate specific points in coding

assessment items to standards or objectives (Appendix H). 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. 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 Reading Framework, the next general level was either the type of text or the cognitive target. For coding to the CRS, the next level was the strand. This coding to a generic standard, or objective, sometimes indicates that the item targets 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 a set of standards or objectives. In this study, assessment items were mapped to a framework 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. However, the NAEP classification of texts was used when labeling passages from both assessments to standardize the coding process and to aid in interpreting the results.

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 (number of items, 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 (COR), 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 Tuesday, Day 2, the reading panelists took longer than planned to map the NAEP assessment items to the objectives in the NAEP Framework. As a result, a few panelists took additional time on Wednesday morning before 8:30 AM. 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 in the prior evening's debrief 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. This comment was made to have panelists 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 reading panelists were separated into two panels to continue the process. In their groups of eight, the panelists registered and logged onto the WATv2. Then the panels assigned DOK levels to the NAEP objectives. Each panelist assigned DOK levels to each of the 46 objectives in the NAEP Reading Framework individually by 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 reading panels and identified nine objectives for which the two panels' coding differed. The two groups of reading panelists then met together and resolved their differences on the disputed objectives. Of the nine objectives in dispute (20 percent of total), the consensus among the two panels was to take the lower DOK level (DOK 1 or 2) on four and the higher DOK level (DOK 3) on five of the objectives. The consensus process also was intended to increase the understanding of the DOK definitions by all panelists.

After analyzing a small sample of items of five to 10 assessment items from the NAEP assessment, each reading panel separately mapped the 163 NAEP Reading 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. Some panelists were slower than other panelists in completing the coding. For the first analysis, the facilitators waited until all eight panelists had completed their coding. However, the facilitators did some intervention with one or two slower members to help them increase their rate of coding. This invention included suggesting not reading each passage word for word and making a decision more quickly. When all of the members of a panel completed mapping all of the items, the 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.

The technical coordinator reviewed the analysis results for NAEP items coded to the NAEP objective for both reading panels after they had completed within-group adjudication. He identified 17 items (10 percent) that needed to be adjudicated between the panels. For these items, the objective assigned by the majority of the first panel did not coincide with the objective assigned by the majority of the other panel. He grouped the 17 items according to those for which the majority in each panel differed on the cognitive process (seven items) and those for which the majority in each panel agreed on the cognitive process but differed on the text type (10 items). Finally, several items were identified for which there was not a majority agreement in one panel. The two panels met as a group of 16 with the two facilitators and adjudicated the identified items. The panels were asked to spend no more than one hour adjudicating between-groups. They were asked to start with the group of seven items and then go to the group of 10 items. After this adjudication discussion, panelists could change their codes if they felt there was a compelling reason. Panelists were not, however, required to change codes and could maintain their original code even if it differed from the majority codes. After this adjudication, only one or two panelists changed their codes on seven of the 17 items. On a few items, one or two panelists changed their codes to agree with the majority of their own panel. One observed difference between the two panels was in their interpretation of how to assign a code for "evaluation." One panel felt that "evaluation" required some value statement or criterion whereas the other panel did not have as strict an interpretation of "evaluation." This difference between the panels was brought up in the adjudication session. The two groups only differed by 0.5 or more on the average DOK on eight of the 163 items (5 percent). The agreement between the two groups on DOK was considered to be sufficient so that adjudication of DOK was not needed.

Days 3-5

On day three, the two reading panels coded the two forms of the EXPLORE Reading assessment to the NAEP Framework objectives. Two EXPLORE forms were analyzed to have some comparison of the variation among the forms. After the panels completed the within-group adjudication, the technical coordinator identified items from each form that needed to be adjudicated between the groups. The two reading panels only varied on four items on Form 1 and

five items on Form 2. For seven of these nine items the two panels only differed on the text type and if the item related to both literary and text passages or if the item related to only one. It was judged that the panels had sufficient agreement and that adjudication was not necessary. The two groups differed on the average DOK assigned to one item by 0.50 or more. The agreement on DOK also was sufficiently high to not warrant using time for adjudication.

The two panels next analyzed the NAEP Reading assessment and the two EXPLORE forms in comparison to the CRS with 39 standards. First each panel assigned a DOK level to each standard and then reached a consensus within their group. The group only differed on the assignment of DOK on six standards (15 percent). Between-group consensus was facilitated by group leaders, as needed, for any standards for which the two groups differed in original DOK codes. The two panels decided on the higher of the DOK values for two of the items and the lower of the DOK values for four of the items. Following reaching consensus across groups on the DOK values for the CRS standards, the separate panels first mapped the 163 NAEP items and then the 30 items from each form of the EXPLORE Reading assessment to the standards in the CRS.

After the NAEP assessment was mapped to the CRS, the technical coordinator identified 24 items (15 percent) on which the majority of the two panels differed on the strand to which the item was assigned. Because of time pressures, the two reading panels were limited to one hour to adjudicate as many of these items as possible. They were able to discuss 15 of the 24 items. This resulted in the two groups varying on only five percent of the items that were not discussed. This was the acceptable percentage as specified in the Design Document. The overall results between panels were not greatly affected by the shortened adjudication process because panelists were reluctant to change the coding of items from the majority within their group. Any differences between the two groups represent reasonable variations in how items could be assigned to different objectives. One or two panelists made changes on eight items. On one item, four panelists from one group changed their results. On at least two items, panelists changed their codes to that of the majority of the panelists within their panel. The differences that remained with codes for these 15 items are likely due to the assessment item not precisely fitting any one standard or strand. The average DOK level assigned to items across all eight panel members only

differed by 0.5 or more on two of the 163 items (1 percent). As a result, the DOK values were not adjudicated.

Finally, the two reading panels coded the two forms of the EXPLORE Reading assessment to the CRS. (Note: Two EXPLORE forms were analyzed to have some comparison of the variation among the forms). After the panels completed the within-group adjudication, the technical coordinator identified items from each form that needed to be adjudicated between groups. The panels repeated the between-group adjudication process for these items. Only three items on each EXPLORE form were identified for adjudication on topic. These items were discussed, but no changes were made by either group. The two panels' results differed by 0.5 or more on the average DOK for one item on Form 1 and two items on Form 2. Although the study design included adjudication when necessary, the 95 percent agreement between the two panels on DOK level was sufficient that the technical coordinator determined that adjudication was not needed and would not change the alignment results to any great extent. Per the Design Document for the study methodology, the codes were not adjudicated because of reasonable between-group agreement.

Variations in the Process

In general, the content analysis process in the reading panels was performed in the intended sequence as outlined in the agenda. There were a few variations to the intended agenda. On Day 2, both panels spent the entire day in training and coding the NAEP assessment items to the NAEP standards. There was no time to do any adjudication as scheduled. Reviewers arrived early on Day 3, as early as 7:30 AM, to complete their coding. The coding however was not completed until after lunch on Wednesday, shifting the schedule by one-half day. The reading panelists spent more time than planned with the NAEP Reading passages because some of them were quite long. Rather than conducting between-panel adjudication, both reading groups went immediately to coding the two EXPLORE forms to the NAEP Framework. Panel 1 completed this coding well before Panel 2, but both panels completed their within-panel adjudication before leaving on Wednesday. At the beginning of Day 4, the two reading panels spent one hour to discuss their differences on the coding of the NAEP assessment to the NAEP Framework they had completed the day before. In this short period of time, one panelist from each group

described the rationale for group's coding results on each item with between-group differences. This was followed by a very brief discussion. Then the 16 panelists went on to the next item. In this session the panels were not able to discuss their analysis codes in the full depth generally included in adjudication. The schedule as planned was continued with the two groups assigning DOK levels to the CRS standards and then conducting a between-group adjudication of the results. However, when the two reading panels began coding the NAEP assessment to the CRS standards a question arose about how to code the passages from a *biography* that was identified by ACT as a literary text. A sidebar discussion was held with the technical coordinator, the COR, the two reading facilitators, and the two ACT observers. The ACT staff indicated that they considered the *biography* as informational. However, according to the classification from the NAEP Framework a biographical sketch was considered as literary non-fiction. The technical coordinator reemphasized the instructions given to the reading facilitators prior to the Institute and referred to the passage types included with the decision rules on how the NAEP classifications are to be used. After consulting with specialist observers and the reading facilitators, it was reaffirmed that *biography* needed to be coded in this study as literary nonfiction. The between-panel adjudication for the NAEP assessment comparison to the CRS was done on Friday morning. As for the NAEP to NAEP adjudication, the adjudication was curtailed after one hour so that only 15 of 24 items were discussed. Adjudicating the 15 items brought the non-adjudicated items (nine) within the five percent margin as specified in the Design Document as not needing adjudication. The between group adjudication of codes for analysis of the two EXPLORE forms with the CRS did not require a considerable amount of time because there were differences on only three items on each form. The process was completed by 1:30 PM on Friday, but some panelists continued to make changes to their codes until 2:00 PM. Overall, the reading groups experienced some pressure to complete the work in the required time. As a result, there are some differences in the alignment data between the two groups, and the implications are discussed in the findings section below.

Feedback Survey

Panelists were requested to complete a brief online survey at the end of the day on 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 Appendix 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 Reading assessment was composed of 163 items of which about two-thirds were multiple-choice items. The other nearly one-third of the NAEP items were constructed-response (short or extended), many of which allowed for partial credit when scored. The EXPLORE items were distributed evenly between passages within three content areas: prose fiction, humanities, and social sciences. In contrast, the NAEP items were almost equally divided between passages categorized as literary or informational. EXPLORE prose fiction content area corresponds to the literary category of the NAEP Framework. The humanities and social science passages of EXPLORE are most likely to correspond to the informational category of the NAEP Framework. Hence, the EXPLORE assessments are likely to have a greater proportion of items focused on informational passages than does the NAEP assessment. EXPLORE also differentiates between “uncomplicated” and “more challenging” passages while NAEP does not. Any greater proportion of items on EXPLORE that target informational text and the challenge level of different texts have the potential of influencing the degree of alignment between EXPLORE and NAEP Reading assessments. However, differences in these areas are important for characterizing the NAEP assessment and judging its viability for providing information about preparedness of students in grade 8. The alignment information will be valuable for reporting on the distribution of passages and reading items between literary and informational text, and the distribution between less and more complicated texts. The alignment results will thus inform decisions about using the NAEP grade 8 assessment for reporting relevant to being on track for academic preparedness for college level work.

Table 2 lists the items by the number of points given for a correct answer. Fifty-two of the NAEP assessment items were given a maximum of two to four points. The analysis weighted items by point value, giving a total of 268 points for the NAEP Reading assessment and a total of 30 points for EXPLORE. Weighting constructed-response 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. The large difference between the two assessments in the total number of items and point values could 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.

Both assessments varied the type of items by content or cognitive complexity. EXPLORE items emphasize referring or reasoning. Referring questions ask about materials explicitly stated in a passage and are designed to measure literal reading understanding. Reasoning questions ask about "meaning implicit in a passage and require cogent reasoning about a passage" (ACT, Inc., 2009, p. 15). The NAEP Framework identified three cognitive levels represented on the assessment (proportion of items in parentheses): Locate/Recall (20 percent), Integrate/Interpret (50 percent), and Critique/Evaluate (30 percent).

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

NAEP Gr, 8 Reading		
Point Value	**Number of Items**	**Total Point Value**
1	111	111
2	11	22
3	29	87
4	12	48
Total	163	268
EXPLORE Form 1		
1	30	30
Total	30	30
EXPLORE Form 2		
1	30	30
Total	30	30

The difference in complexity of the NAEP objectives and the CRS is shown in terms of DOK in Tables 3 and 4 below. The panelists judged that the majority (77 percent) of the 39 standards in the CRS had a DOK level of 2, related to application of skills and concepts, whereas the majority

(61 percent) of the 46 objectives in the NAEP Reading Framework were judged as DOK 3, related to reasoning about and evaluating text. Only two standards (5 percent) of the CRS were judged to address content at the level of 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 reading 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 answer 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 by DOK levels represents the intended DOK levels and not necessarily the actual distribution of reading 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 Reading 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
Locate/Recall	4	1 2 3	3 1 0	75 25 0
Integrate/Interpret	28	1 2 3	0 14 14	0 50 50
Critique/Evaluate	14	1 2 3	0 0 14	0 0 100
Total	46	1 2 3	3 15 28	6.5 32.6 60.9

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

CRS Strands	Total Number of Standards	DOK Level	Number of Standards by DOK Level	Percent within Strands by DOK Level
Main Ideas and Author's Approach (MID)	8	1 2 3	2 6 0	25 75 0
Supporting Details (SUP)	8	1 2 3	3 4 1	37.5 50 12.5
Sequential, Comparative, and Cause-Effect Relationships (REL)	12	1 2 3	2 10 0	16.7 83.3 0
Meanings of Words (MOW)	5	1 2 3	0 5 0	0 100 0
Generalizations and Conclusions (GEN)	6	1 2 3	0 5 1	0 83.3 16.7
Total	39	1 2 3	7 30 2	18 77 5

Alignment of Assessments to the Frameworks

An important step in judging the alignment between the NAEP and EXPLORE assessments was mapping each assessment to both frameworks. The results of these mappings were used to draw conclusions on the alignment between the two assessments. Within each of their panels, the reading panelists individually coded the assessment items to a framework. Guided by the facilitator, the panelists adjudicated the codes for any items with a large variation in the assignment of objectives or standards to the item or in the assignment of a DOK level to the item. In general, variations between the panels were adjudicated by all 16 panelists if the number of items needing adjudication was five or more. The final results for the reporting categories (e.g., content areas or strands) 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 framework was if each panelist found an appropriate objective or standard within the framework that corresponded to what an item was measuring. If no corresponding objective or standard was found, then the panelists were asked to match the item to a generic standard, defined as a more general content category such as a

subtopic, content area, or strand. When panelists code a large number of items to generic standards, it indicates that, for these items, the match to the framework is only very general. In this report, items coded to generic standards are only reported in the analysis if two or more panelists coded an item as such. If a panelist could not find even a generic standard that corresponded to an item, then the panelist was to enter "uncodeable" for the item. No items were considered uncodeable for either the NAEP or EXPLORE reading assessments.

None of the panelists from either reading panel coded any of the NAEP assessment items to a generic standard when the items were mapped to the NAEP Framework (Table 5). Similarly, none of the panelists from either reading panel coded any EXPLORE items from each form to a generic standard when the items were mapped to the NAEP Framework (Table 6). Only one panelist mapped five items to the generic standard MID and three items to the generic standard REL when coding the NAEP assessment to the CRS (Table 7). The general rationale for coding an item to the CRS strand MID was that no standard under this strand required determining the main idea of a passage, rather the item only required it for a paragraph. The panelist who mapped items to the CRS strand REL noted that although the items related to this general topic, the panelist did not see a specific standard that included the combination of required components for these items. For example, CRS standard REL 503 states *Identify clear relationships between characters, ideas, and so on in more challenging literary narratives* and CRS standard REL 505 states *Identify clear cause-effect relationships in more challenging passages* but no standard includes challenging non-literary texts with relationships that are not cause-and-effect. None of the panelists from either reading panel coded any of the assessment items on the EXPLORE forms to a generic standard when the items were mapped to the CRS (Table 8). Aside from the few items that were judged to fit generic standards by only one panelist, panelists found specific matches for all items overall.

Summary of coding to generic standard: With the exception of one panelist, no items were coded to generic standards in any of the analysis. Thus, nearly all of the panelists were able to find a match for each NAEP Reading assessment item and for each EXPLORE item when mapping either to the NAEP Reading Framework or to the CRS for EXPLORE Reading.

Summary of coding items as uncodeable: None of the reading panelists marked any reading items on either of the assessments as uncodeable.

Table 5. Items assigned to generic content expectations by one or more panelists by panel and number of reviewers for the 2013 NAEP Grade 8 Reading Assessment mapped to the NAEP 2013 Reading Framework for grade 8

Assessment Panel	Generic Content Expectation
NAEP Panel 1	No generic objectives coded by more than one panelist
NAEP Panel 2	No generic objectives coded by more than one panelist

Table 6. Items assigned to generic content expectations by one or more panelists by panel and number of reviewers for EXPLORE Reading Forms 1 and 2 mapped to the NAEP 2013 Reading Framework for grade 8

Assessment Panel	Generic Content Expectation
EXPLORE 1 Panel 1	No generic objectives coded by more than one panelist
EXPLORE 1 Panel 2	No generic objectives coded by more than one panelist
EXPLORE 2 Panel 1	No generic objectives coded by more than one panelist
EXPLORE 2 Panel 2	No generic objectives coded by more than one panelist

Table 7. Items assigned to generic content expectations by one or more panelists by panel and number of reviewers for the 2013 NAEP Grade 8 Reading Assessment mapped to the ACT College Readiness Standards for EXPLORE Reading

Assessment Panel	Generic Content Expectation	Number of items by number of panelists
NAEP Panel 1	MID	five items by one panelist
	REL	three items by one panelist
NAEP Panel 2	None	No generic objectives coded by more than one panelist

Table 8. Items assigned to generic content expectations by one or more panelists by panel and number of reviewers for EXPLORE Reading Forms 1 and 2 mapped to the ACT College Readiness Standards for EXPLORE Reading

Assessment Panel	Generic Content Expectation
EXPLORE 1 Panel 1	No generic objectives coded by more than one panelist
EXPLORE 1 Panel 2	No generic objectives coded by more than one panelist
EXPLORE 2 Panel 1	No generic objectives coded by more than one panelist
EXPLORE 2 Panel 2	No generic objectives coded by more than one panelist

Study 1: Alignment of the NAEP Assessment and the NAEP 2013 Reading 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 Reading Framework and the 2013 NAEP Grade 8 Reading assessment is high compared to the alignment of other standards and assessments analyzed in prior studies. One contributing factor is the large number of items and point values, which contributes to the ease of meeting the criterion of categorical concurrence. In this analysis, each item is weighted by its maximum point value, which ranges from one to four points.

Categorical Concurrence. Table 9 shows a summary of each panel's codes by the 3 NAEP assessment areas, totaling the point values of each item that mapped to each area. The point values for each of the three content areas are from 29 (Locate/Recall) to 189 (Integrate/Interpret). This is far in excess of the Categorical-Concurrence criterion's threshold level (six items) used in this study to make a reliable judgment on a student's performance on a content area, as discussed on pages 7 and 8. Note that assessments may not report on performance by individual content areas, and so Categorical Concurrence should not be interpreted as a reflection of the assessment itself. Even if the items were not weighted by point value, the number of items mapped to the respective content areas for each of the two panels is well above the minimum level. Among the three content areas, the greatest emphasis is given to Integrate/Interpret—120 items (74 percent) by Panel 1 and 111 items (68 percent) by Panel 2. Lower emphasis by the total number of items

is given to the other two categories. The category of Locate/Recall accounts for 27 items (16 percent) by Panel 1 and 30 items (18 percent) by Panel 2. The category of Critique/Evaluate accounts for 16 items (10 percent) by Panel 1 and 23 items (14 percent) by Panel 2. The category of Locate/Recall accounted for a larger number of items than Critique/Evaluate, but the Locate/Recall items tended to be 1-point or 2-point items. The percentage of weighted NAEP items corresponding to each NAEP content category, as determined by the panelists, corresponds closely to the percentages of items specified in the NAEP framework (NAGB, 2012).

Each of the three NAEP reading reporting categories had more than 25 point values (Locate/Recall; Integrate/Interpret; Critique/Evaluate). From 60 to 70 percent of the point values were found to target the Integrate/Interpret category. Items on the assessment targeted over 50 percent of the underlying objectives for each of the three reporting categories. The items were evenly distributed among the objectives for two of the categories, but emphasized the objective of interpreting the meaning of a word as used in text in the Integrate/Interpret category. This was purposefully done to have at least two vocabulary items for each of the 19 passages. This overemphasis on vocabulary items was not considered a major alignment issue because all of the other three alignment criteria met the threshold level.

Depth-of-Knowledge Consistency. The Depth-of-Knowledge Consistency was high for all three NAEP content areas for both panels. Panel 1 found that 71 percent to 91 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 76 percent to 95 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. Overall, the results between the two panels had strong agreement on the level of complexity of the NAEP items and on whether the complexity of the items matched the complexity as expected by the objectives. The DOK value assigned to an item correlated with the point value of the item (Table 10). The average DOK increased as the point value increased. Most of the multiple-choice items were assigned a DOK 1 or 2. Two-point items were generally assigned DOK 2, although panelists also coded some NAEP items as DOK 1 or DOK 3. Three-point items were generally assigned DOK 2 or 3. Four-point items were generally assigned DOK 2 or 3, leaning more heavily toward DOK 3 than the three-point items.

Table 9. Item numbers and percentages on four alignment criteria by panel for the 2013 NAEP Grade 8 Reading Assessment mapped to the NAEP 2013 Reading Framework for grade 8

NAEP Content Areas	NAEP Assessment Items Mapped to NAEP Framework Reading 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 objectives 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
1 Locate/ Recall	29.6	32.5	71	76	75	81	0.84	0.75
2 Integrate/ Interpret	189	168	73	78	70	65	0.60	0.64
3 Critique/ Evaluate	51	72	91	95	54	60	0.75	0.76
Total Point Value	269.6	272.5						

Table 10. Average depth-of-knowledge level of 2013 NAEP grade 8 reading items by item type, panel and average across panels

Item Type	Number of Items	Average DOK Panel 1	Average DOK Panel 2	Average DOK Across Panels
Multiple Choice (1 Point)	111	1.8	1.8	1.8
Constructed-response (2 Points)	11	2.0	2.2	2.1
Constructed-response (3 Points)	29	2.6	2.6	2.6
Constructed-response (4 Points)	12	2.6	2.8	2.7

Range of Knowledge. The Range-of-Knowledge Correspondence indicates that at least one item on the NAEP assessment corresponded to 54 to 81 percent of the framework objectives, with the percentage varying by content area and panel. The Range of Knowledge in the NAEP assessment, compared with the NAEP Framework, far exceeds the threshold of 50 percent of the framework objectives with at least one corresponding item. The two panels had high agreement on Range of Knowledge for each of the given content areas, only differing by up to about 6 percent for any of the content areas. Both panels found Locate/Recall with the highest coverage. Across content areas, each group found at least one item that corresponded to 29 of the 46 objectives. In other words, 63 percent of the objectives, overall, were targeted by at least one item on the assessment.

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 more that items represent a monomial or a binomial distribution, the lower the index will be. Only the Integrate/Interpret category did not have a Balance Index that reached the threshold level of 0.70. The majority of panelists in both reading groups coded a total of 56 (Panel 1) or 57 (Panel 2) items to objective 2.1.d *Interpret the meaning of a words as it is used in the text*. One reading group (Panel 2) mapped 17 items to objective 2.1.a *Make simple inferences from texts*. The other reading panel, however, mapped 13 items to this same objective. Panel 1 coded nine items to 2.2.f *Recognize character actions and infer and support character feelings* and Panel 2 coded five items to this objective. All other objectives to which at least one item was coded by a majority of panelists had between one and five items mapped to each. The relatively high emphasis on interpreting the meaning of words used in a text lowered the Balance Index for the Integrate/Interpret category. This was by design with the intent to have two vocabulary items for each of the 19 passages. Because the other alignment criteria were met for this content area, this over-emphasis on interpreting the meaning of words is not considered a deficit in alignment and can be considered a choice to emphasize vocabulary more on a grade 8 assessment as it is an important topic for the middle grades. The Balance Index values for both groups were very close and only differed by 0.01 to 0.09 for any content area.

Panelists Responses to Debriefing Questions for Study 1. In the debriefing on Study 1, panelists' comments were consistent with the findings from the Balance Index. Based on their expertise as educators, several panelists noted that there was a good balance in the variety of types of texts found in the NAEP items, including literary and informational passages.

[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.)

Summary of Study 1 Findings:

- **Categorical concurrence**. The alignment between the NAEP assessment and the NAEP framework met this criterion, with point values for all three NAEP content areas far in excess of the threshold level (six items) to make reliable judgments about student performance on a content area.
- **The Depth-of-Knowledge Consistency** was high for all three content areas as reviewed by both reading panels.
- **The Range-of-Knowledge Correspondence** indicates that at least one item on the NAEP assessment corresponded to 54 to 81 percent of the framework objectives.
- **Balance of Representation Index**. Only the Integrate/Interpret category did not have a Balance Index that reached the 0.7 threshold index level. The Locate/Recall and Critique/Evaluate content areas each exceeded the acceptable level for the Balance Index.
- **Overall,** the 2013 NAEP Grade 8 Reading assessment and the 2013 NAEP Reading Framework were found to have acceptable levels on all four alignment criteria. This was true for both panels. From 60 to 70 percent of the point values were found to target the Integrate/Interpret category. The items had high DOK Consistency with the DOK levels of the assigned objectives, with over 70 percent agreement. Items on the assessment targeted over 50 percent of the underlying objectives for each of the three reporting categories. Only the Integrate/Interpret category did not have a Balance Index that reached the 0.7 index level threshold.

Study 2: Alignment of EXPLORE Forms 1 and 2 and the NAEP 2013 Reading Framework

The alignment between EXPLORE Forms 1 and 2 and the NAEP 2013 Reading Framework was weak overall, according to the threshold levels for the four alignment criteria. The pattern of degree of alignment was similar for both forms and consistent between the two panels.

Categorical Concurrence. Both EXPLORE forms had 30 assessment items. The two panels agreed that about half of the items on each of the forms corresponded to expectations under the Locate/Recall NAEP reporting category and about half of the items corresponded to expectations under the Integrate/Interpret NAEP reporting category (Tables 11 and 12). The main alignment deficit was a lack of items that corresponded to the Critique/Evaluate NAEP category. Neither

form had any items that more than one reviewer coded to any objectives within the Critique/Evaluate category. This absence of items corresponding to the expectations within the Critique/Evaluate category suggests a different focus for the EXPLORE assessments compared with the NAEP Framework. The items on the EXPLORE forms targeted recall of knowledge and routine reading skills as well as application of these skills, including making simple inferences. The items on EXPLORE did not target objectives that expected complex inferences, critique, or evaluation of text as expressed by the NAEP categories.

Depth-of-Knowledge Consistency. The Depth-of-Knowledge Consistency criterion between the two EXPLORE assessment forms and the NAEP Reading Framework was acceptably met for two of the three NAEP reporting categories, Locate/Recall and Integrate/Interpret. For these two reporting categories, over 70 percent of the items had a DOK level that was at least as high as the DOK level of the assigned objective, higher than the threshold of 50 percent. Depth-of-Knowledge Consistency was not met for the third NAEP reporting category, Critique/Evaluate, because there were an insufficient number of items to make a judgment. It should be noted that with the exception of one item on one EXPLORE form as judged by one panelist between the two panels, all items mapped to Integrate/Interpret targeted objectives that had been assigned a DOK 2, reflecting the use and application of reading skills. Recall that within the Integrate/Interpret NAEP reporting category, 14 of the 28 objectives (50 percent) expect DOK 3 level items, which demand deep inference and reasoning. However, all EXPLORE items that mapped to Integrate/Interpret objectives were assigned a DOK 1 or 2. Although DOK Consistency could be considered acceptable, according to the threshold for the criterion, it is important to recognize that all of the items identified as corresponding to Integrate/Interpret were assigned a DOK 1 or 2. These items mapped favorably to the subset of about half of the objectives under the Integrate/Interpret category with these DOK levels. However, half of the objectives under Integrate/Interpret were judged to expect DOK 3 engagement and none of the EXPLORE assessment items were consistent with the content complexity of these objectives. None of the items on either of the EXPLORE forms were judged by any of the reviewers as DOK 3.

Range-of-Knowledge Correspondence. Data from both panels indicated the reporting category of Locate/Recall to be acceptably aligned with each EXPLORE form for all alignment criteria,

including Range of Knowledge and Balance of Representation. Data from both reading panels supported that over 70 percent of the underlying objectives for the NAEP Locate/Recall reporting category had at least one corresponding item. Both groups found that for each form zero items mapped to 22 to 24 of the NAEP Integrate/Interpret objectives, meaning that only four to six of the objectives in this reporting category were targeted by at least one assessment item. Even when a composite[6] form, the aggregation of the two forms into one form of 60 items, of both of the EXPLORE assessment forms was considered, Panel 1 mapped at least one item to only eight of the 28 objectives (28 percent) and Panel 2 mapped at least one item to only four of the objectives (14 percent) (Table 14). Thus, the Range-of-Knowledge Correspondence was not considered met for the NAEP reporting category Integrate/Interpret. As for the other criterion, there was an insufficient number of items to make any determination on Range of Knowledge for Critique/Evaluate.

Balance of Representation. The items mapped to objectives under the Locate/Recall reporting category were fairly evenly distributed as indicated by a Balance Index of over 0.85. This was not the case for the Integrate/Interpret NAEP reporting category. For each form, both panels found that most of the items within Integrate/Interpret mapped to NAEP objective 2.1.a *Make simple inferences from texts*. Both groups found seven or eight items on both forms that mapped to this objective. Only a maximum of three items was found to map to any other of the Integrate/Interpret objectives. Thus, the Balance Index was slightly below what was considered as the threshold, 0.70.

Panelists Responses to Debriefing Questions for Study 2. Panelists' comments to the debriefing questions supported the findings. Many comments focused on the difference in the complexity of content on NAEP compared with the EXPLORE forms. Panelists noted that items on the EXPLORE forms required recall or simple inferences only and that in general, the cognitive complexity was lower than what was required by the NAEP Reading Framework. Panelists also noted the EXPLORE forms did not include poetry passages.

[6] The composite EXPLORE form will referred 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.

Summary of Study 2 Findings:

- **Categorical concurrence**. The two panels agreed that half of the EXPLORE items corresponded to expectations under the Locate/Recall NAEP content area and half of the items corresponded to expectations under the Integrate/Interpret content area. However, a majority of the panelists did not find any items that mapped to the third NAEP category, Critique/Evaluate.
- **The Depth-of-Knowledge Consistency** was acceptable for the EXPLORE items that mapped to Locate/Recall and Integrate/Interpret reporting categories. The items mapped to these categories primarily targeted NAEP objectives assigned a DOK level of 2 rather than the more complex objectives under these categories. All EXPLORE items that mapped to Integrate/Interpret objectives were assigned a DOK 1 or 2. When compared to just the multiple-choice items on the NAEP assessment (Table 10), the 60 multiple-choice items across the two EXPLORE forms had a lower average DOK level as coded by both panels (Table 13).
- **The Range-of-Knowledge Correspondence** indicates that the acceptable 50 percent level was reached for only one of the three content areas of EXPLORE Reading. The Range-of-Knowledge Correspondence level was not improved when the composite of the two EXPLORE Reading forms with a total of 60 items was used.
- **Balance of Representation Index**. Only the Locate/Recall content area had a Balance Index that reached the 0.7 minimum level.
- **Overall,** EXPLORE had sufficient items to map well to two of the three NAEP categories, but a majority of the panelists did not find any items that mapped to the third category, Critique/Evaluate. The items coded to the Locate/Recall and Integrate/Interpret categories primarily targeted the objectives assigned a DOK 2 rather than the more complex objectives under this category. The levels for Range of Knowledge and Balance of Representation exceeded the threshold levels for the Locate/Recall NAEP category but were rated low for the Integrate/Interpret category. Range of Knowledge was not improved when the composite of the two EXPLORE Reading forms (a total of 60 items) was analyzed instead of the two 30-item forms.

Table 11. Item numbers and percentages on four alignment criteria by panel for EXPLORE Reading Form 1 mapped to the NAEP 2013 Reading Framework for grade 8

NAEP Content Areas	EXPLORE Form 1 Mapped to NAEP Framework Reading 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 objectives 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
1 Locate/ Recall	14.9	17.4	72.0	77.4	71.9	75	0.87	0.86
2 Integrate/ Interpret	15	12.6	88.9	93.75	25	16.1	0.68	0.64
3 Critique/ Evaluate	0.12	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Total Point Value	30	30						

Table 12. Item numbers and percentages on four alignment criteria by panel for EXPLORE Reading Form 2 mapped to the NAEP 2013 Reading Framework for grade 8

NAEP Content Areas	EXPLORE Form 2 Mapped to NAEP Framework Reading 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 objectives 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
1 Locate/ Recall	14.8	15.4	73.8	74.1	75	75	0.85	0.79
2 Integrate/ Interpret	15.2	14.6	89.4	91	22.3	15.6	0.62	0.68
3 Critique/ Evaluate	0	0	N/A	N/A	N/A	N/A	N/A	N/A
Total Point Value	30	30						

Table 13. Average depth-of-knowledge level of EXPLORE Reading items across two forms by item type, panel and across panels

Item Type	Number of Items (Across both forms)	Average DOK Panel 1	Average DOK Panel 2	Average DOK Across Panels
Multiple Choice (1 Point)	60	1.5	1.5	1.5

Table 14. Percent of objectives under a reporting category with at least one corresponding item from the composite of two EXPLORE Reading forms (1 and 2) mapped to the NAEP 2013 Reading Framework for grade 8

NAEP Reporting Categories	Composite of EXPLORE Forms 1 and 2 Mapped to NAEP Reading Framework by Panels 1 and 2	
	Range of Knowledge (percent of objectives with at least one hit)	
	Panel 1* %	Panel 2* %
1 Locate/Recall	75	75
2 Integrate/Interpret	29	14
3 Critique/Evaluate	NA**	NA

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

** NA indicates insufficient number of items to compute the Range of Knowledge

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

The alignment between the 2013 NAEP Grade 8 Reading assessment and the CRS was moderate relative to the four alignment criteria, based on the content analysis data yielded from the panelists' review (Table 15).

Categorical Concurrence. The Categorical Concurrence between the NAEP assessment and the CRS for EXPLORE was high. Both panels found a number of items and point values that exceeded the threshold for each of the five CRS strands. The CRS strand Sequential, Comparative, and Cause-Effect Relationships (REL) had the lowest total point value based on panelist codes, according to both panels' data. The other four strands had 50 point values or more, based on averages from both panels. The NAEP point values corresponding to each strand was consistent between the two panels for the CRS strands Main Ideas and Author's Approach (MID) and Meanings of Words (MOW). Panel results varied by nearly 12 points for the CRS strand Supporting Details (SUP), by nearly 13 points for the CRS strand Sequential,

Comparative, and Cause-Effect Relationships (REL), and by about 28 points for the CRS strand Generalizations and Conclusions (GEN). Approximately 90 percent of the NAEP items were found to target standards at the 400 or 500 score levels representing items likely to be solved by students scoring higher on EXPLORE scoring scale. The CRS strand Supporting Details (SUP) was the only strand that included items judged by both panels to correspond to standards at the 200 or 300 score level.

Depth-of-Knowledge Consistency. Depth-of-Knowledge Consistency was high. The DOK levels of the NAEP items compared favorably with the DOK levels of the corresponding CRS standards and a reasonable number of standards under four of the five strands had over half of the items with a DOK that was the same as or higher than the DOK of the corresponding standard. For the CRS strand Supporting Details (SUP), around 75 percent of the items had a DOK level that was at least as high as the DOK level of the assigned standard. For the other four strands, the DOK Consistency was around 90 percent or higher. The DOK Consistency for the CRS strand SUP was lower largely because the majority of the items that panelists coded to CRS standard SUP 502 were judged to have a DOK level 1 in contrast with the DOK 2 assigned to the standard. Another contributing factor was that the majority of items that panelists coded to CRS standard SUP 503 were judged to have a DOK level 2 in contrast with the DOK 3 assigned to the standard.

Range-of-Knowledge Correspondence. The Range-of-Knowledge Correspondence criterion was met for the assessment for four of the five CRS strands. In other words, the NAEP assessment had items that mapped to more than half of the standards underlying a CRS strand for four of the five strands. Data suggest the criterion was not met for Range of Knowledge for items with respect to the CRS strand Sequential, Comparative, and Cause-Effect Relationships (REL). For the other four strands the NAEP items mapped to 60 percent to 100 percent of the underlying standards. The results of each panel varied from less than 3 percent (CRS strands MID and MOW) to nearly 20 percent (CRS strand REL). The Range of Knowledge was lower for the CRS strand REL because of weighted items mapping to CRS standards REL 503 and 505 (for Panel 1) and CRS standards REL 502 and 503 (for Panel 2). Neither panel had a majority of panelists mapping any of the items to the standards for the lower score levels (200 and 300) for the CRS strand REL. Overall, the Range of Knowledge between the assessment and the CRS indicated

that the NAEP assessment items corresponded to a high percentage of the reading standards in the CRS, particularly those representing the higher score intervals.

Balance of Representation. Balance of Representation between the NAEP assessment and the CRS was lower than the threshold level (0.70) for two of the CRS strands, i.e., Main Ideas and Author's Approach (MID) and Meanings of Words (MOW). One group's data show low Balance of Representation for the CRS strand Generalizations and Conclusions (GEN). Even when Range of Knowledge was fairly high for a strand, panelists in both groups found that most of the items corresponding to the strand mapped to one standard. The most severe case of this was for the CRS strand Meanings of Words (MOW). Reviewers found NAEP items worth from 62 to 66 in total point value (mainly one-point items) that mapped to this strand, but nearly all of these NAEP point values mapped to two standards, CRS standards MOW 501 and MOW 502. Both of these standards addressed using context to determine the meaning of a word. This finding is comparable to the finding when the NAEP assessment was mapped to the NAEP Framework. About 60 of the items ask students to determine the meaning of a word in context. The lower Balance of Representation between the NAEP assessment and the CRS also could be partly due to the cumulative structure of the standards. Overall, panelists chose to map items to the standards representing the higher score intervals. Considering the Balance-of-Representation and Range-of-Knowledge criteria together, the NAEP assessment was found to target a reasonably high number of the CRS standards. This was the only study in which a panelist coded any reading item to a generic standard (Table 7). However, no item received such an assignment by two or more panelists. This indicates that panelists were able to fit each of the NAEP items to at least one of the CRS reading standards.

Panelists Responses to Debriefing Questions for Study 3. In their comments, panelists described some challenges of mapping the NAEP items to the CRS. A panelist observed, "[t]he ACT standards didn't reach the level of rigor expected on the NAEP assessment. For example, there is no category for evaluation, only generalization and drawing conclusions. There is no allowance for figurative language outside of uncomplicated literary narrative. For that matter, all literary text is not literary narrative, e.g., poetry. There is no means to identify an argument. No place for a discussion of text features. No standard for identification of theme. No place to code simple conclusions of challenging texts." Another panelist noted that "[t]here were topics and

skills that were hard to match. [For example], the ACT standards do not directly address poetry." The data indicate alignment, but these and other item-level panelists' comments suggest that these findings should be considered in conjunction with these content differences between NAEP items and the CRS.

Summary of Study 3 Findings:

- **Categorical concurrence**. The number of point values for items from the NAEP Reading assessment was high for each of the five CRS strands, more than 26 points for any one strand. Both panels found NAEP items with over 13 point values that mapped to each of the five CRS content strands. One panel coded more items to the CRS strand Sequential, Comparative, and Cause-Effect Relationships (REL), while the other panel coded more items to the CRS strand Generalizations and Conclusions (GEN). What one panel interpreted as drawing a subtle generalization, the other panel interpreted as drawing a relationship, e.g., cause and effect.
- **The Depth-of-Knowledge Consistency** was rated as high. For the CRS strand Supporting Details (SUP), 75 percent of the items had a DOK level that was at least as high as the DOK level of the assigned standard. For the other four strands, the DOK Consistency was 90 percent or higher.
- **The Range-of-Knowledge Correspondence** criterion was met for the NAEP Reading assessment relative to four of the five CRS strands. The results indicate a low Range of Knowledge level for the NAEP items relating to the CRS strand Sequential, Comparative, and Cause-Effect Relationships (REL). Panelists from both groups found items that targeted 50 percent or fewer of the standards under the CRS strand REL. For the other four strands the NAEP assessment items mapped to 60 to 100 percent of the underlying standards. The NAEP Reading assessment overemphasized one or two standards under two CRS strands— Main Ideas and Author's Approach (MID) and Meanings of Words (MOW). Similar to the alignment with the NAEP Framework, the NAEP assessment overemphasized the standards in the CRS that related to using context to determine the meaning of words.

- **The Balance of Representation** between the NAEP Reading assessment and the CRS was lower than the threshold level (0.70) for two of the CRS strands—Main Ideas and Author's Approach (MID) and Meanings of Words (MOW). The Balance Index level was acceptable for the other three strands.
- **Overall,** the NAEP Reading assessment and the CRS were found to have mixed alignment analysis results. Both panels mapped NAEP items with over 13 point values to each of the five CRS content strands, and the DOK Consistency was high for all five strands. Range-of-Knowledge Correspondence was acceptable for four of the five strands. Range-of-Knowledge Correspondence was below 50 percent for one strand (REL), and the Balance Index was below a value of 0.70 for two strands (MID and MOW).

Table 15. Item numbers and percentages on four alignment criteria by panel for the 2013 NAEP Grade 8 Reading Assessment mapped to the ACT College Readiness Standards for EXPLORE

CRS Strands	NAEP Grade 8 Reading Assessment Mapped to College Readiness Standards Reading 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
Main Ideas and Author's Approach (MID)	51	48.5	98.3	98.5	67.7	70.3	0.69	0.67
Supporting Details (SUP)	78.4	66.6	79.8	71.2	100	87.5	0.78	0.82
Sequential, Comparative, and Cause-Effect Relationships (REL)	26.6	13.75	98.6	92.8	48.8	29.2	0.73	0.73
Meanings of Words (MOW)	62.3	66	93.7	88.3	60	60	0.65	0.65
Generalizations and Conclusions (GEN)	54.9	82.5	88.1	91.6	60.4	72.9	0.79	0.66
Total Point Value	273.2	277.4						

Study 4: Alignment of the EXPLORE Reading Assessment with College Readiness Standards for EXPLORE

The alignment analysis results for EXPLORE in relation to the CRS met acceptable levels for two of the four alignment criteria. Panelists found only slight differences between the two EXPLORE Reading forms relative to the CRS.

Categorical Concurrence. With five strands of the CRS and a 30-item assessment, exactly six items should target standards within each strand to attain acceptable alignment according to the threshold level used in this analysis. For both forms and according to data from both panels, this minimum level of six items per strand was only attained for the CRS strand SUP (Tables 16 and 17). Form 1 had 22 or 23 items and Form 2 had 18 or 19 items that mapped to the CRS strand SUP. Each form, according to at least one of the panels, had one or more strands to which zero items were judged by a majority of panelists to correspond. For Form 1, Panel 1 found no items that more than one reviewer coded as corresponding to the CRS strand REL. A slim majority (five of eight reviewers) on Panel 2 identified one item corresponding to the CRS strand REL but no items corresponding to the CRS strand GEN. For Form 2, Panel 1 identified one item as corresponding to the CRS strand GEN but Panel 2 disagreed, finding no items corresponding to the CRS strand GEN. The two panels disagreed on three items. On these three items, Panel 2 members felt a student is required to locate details or make a simple inference whereas Panel 1 members felt that a student is required to make a generalization about ideas or people. Panel 2 tended not to assign these items to generalizations because they thought a typical 8th grader would answer the questions by making a simple inference or attending to details.

Depth-of-Knowledge Consistency. Depth-of-Knowledge Consistency between the two EXPLORE forms and the CRS was acceptable. Over 90 percent of the items for most of the strands had a DOK level that was the same as or higher than the DOK level of the corresponding standard. For Form 1, the panels did not agree on the DOK levels of the items mapped to the CRS strand REL. Panel 1 shows 100 percent of items at or above the DOK of the corresponding standard, while Panel 2 shows 68 percent of items at or above the corresponding DOK of the standard. For Form 2, there was some disagreement for the CRS strand SUP (85 percent vs. 70 percent). It is important to note, however, that for each form there were only four or fewer items judged by a majority of panelists to correspond to strands other than the CRS strand SUP. Even

if DOK Consistency is technically met for an assessment, that is, if only two items, for example, correspond to a strand then DOK Consistency relates only to these two assessment items and cannot be extended beyond this limited sample. The data support DOK Consistency for the CRS strand SUP but for the other strands, the lack of Categorical Concurrence needs to be taken into account when considering the technical attainment of DOK Consistency.

Range-of-Knowledge Correspondence. Range-of-Knowledge Correspondence was acceptable for SUP but low or unmet for the other four strands. The threshold level is matching items to at least 50 percent of the standards under a strand. The one exception is the results from Panel 1, Form 2, in which case a majority of panelists mapped items to half of the MOW standards. The two EXPLORE forms varied some in the standards targeted by each. When a composite of both forms was considered, the Range of Knowledge improved for Panel 1 for the MID strand (five of eight standards with corresponding items) and SUP strand (six of eight standards with corresponding items). Panel 2 results did not vary when considering a composite form of 60 items created by aggregating the items from each of the two EXPLORE assessment forms (Table 18). Panel 1 increased for the CRS strand MID to 62 percent when considering the composite of the two forms. In this composite analysis, the Range of Knowledge for the CRS strand SUP increased to seven of eight standards (88 percent) for Panel 2 with at least one corresponding item and the Range of Knowledge for the CRS strand MOW increased to three standards with corresponding items (60 percent) for Panel 2. The other three strands were well below the 50 percent threshold when considering the composite form.

Balance of Representation. Because only one or two standards within strands other than the SUP strand had corresponding items, the Balance Index was not met for either of the EXPLORE forms. The CRS strand SUP had an acceptable balance as determined by data from both panels and assessments.

Panelists Responses to Debriefing Questions for Study 4. Several panelists noted that there was minimal complexity within both the passages and the assessment items. Panelists also noted an absence of items probing abstract thought, which is consistent with the DOK levels assigned to the EXPLORE items. Panelists expressed surprise at the high proportion of items that mapped to the CRS strand SUP, and the very small set of items that corresponded to the other strands.

Summary of Study 4 Findings:

- **Categorical Concurrence.** The threshold level of six items per strand was only attained for the CRS strand SUP. The average number of items coded to a strand for the other four strands varied from zero to three items. **The Depth-of-Knowledge Consistency** was rated as acceptable. Over 90 percent of the items had a DOK level that was the same as or higher than the DOK level of the corresponding standard. The data support DOK Consistency for the SUP strand, but for the other strands, the lack of Categorical Concurrence needs to be considered in relation to the technical attainment of DOK Consistency.
- **Range-of-Knowledge Correspondence** was acceptable for the CRS strand SUP but low or unmet for all other strands, and they were rated by panelists to have items corresponding to at most two standards per strand.
- **The Balance of Representation** results exceeded the 0.70 level for the CRS strand SUP. For the other four strands, the Balance Index had little meaning because of the low number of items that were found corresponding to any of these strands.
- **Overall,** the alignment between EXPLORE and the CRS was found to be weak based on the threshold levels used for each of the four criteria. Sixty percent or more of the items on each form mapped to only one of the five CRS strands, the Supporting Details (SUP) strand. The other four strands were mapped with an average of fewer than five items. The acceptable level was met for the Depth-of-Knowledge Consistency for the two forms analyzed. However, the Range-of-Knowledge Correspondence was below 50 percent for four strands with EXPLORE Form 1 and three strands with EXPLORE Form 2. The Range-of-Knowledge Correspondence was improved slightly when the composite of two forms was considered. Note that EXPLORE is not designed to report performance on individual strands.

Table 16. Number of items and percentages on four alignment criteria by panel for EXPLORE Reading Form 1 mapped to the ACT College Readiness Standards

CRS Strands	EXPLORE Reading Form 1 Mapped to College Readiness Standards Reading 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
Main Ideas and Author's Approach (MID)	1.9	2.3	93.8	100	12.5	25	1.00	0.96
Supporting Details (SUP)	22	23.3	97.7	92.5	57.8	62.5	0.70	0.71
Sequential, Comparative, and Cause-Effect Relationships (REL)	0.1	1.5	100	68.3	1.04	10.4	N/A*	0.98
Meanings of Words (MOW)	3.	2.8	91.7	100	22.5	40	0.98	0.90
Generalizations and Conclusions (GEN)	3.1	0.6	100	100	16.7	8.3	1.00	1.00
Total Point Value	30.1	30.5						

* N/A indicates an insufficient number of items was assigned to this strand in order to compute balance.

Table 17. Number of items and percentages on four alignment criteria by panel for EXPLORE Reading Form 2 mapped to the ACT College Readiness Standards for EXPLORE

CRS Strands	EXPLORE Reading Form 2 Mapped to College Readiness Standards Reading 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
Main Ideas and Author's Approach (MID)	4.5	3.9	95	100	48.4	32.8	0.90	0.88
Supporting Details (SUP)	18.9	18	84.6	69.6	56.3	75	0.83	0.74
Sequential, Comparative, and Cause-Effect Relationships (REL)	2.1	3.3	100	91.7	17.7	21.9	1.00	0.95
Meanings of Words (MOW)	3.3	5.1	90.6	90.4	47.5	50	0.85	0.87
Generalizations and Conclusions (GEN)	1.4	0.5	75	50	22.9	8.3	1.00	1.00
Total Point Value	30.1	30.8						

Table 18. Percent of standards under a reporting category with at least one corresponding item from the composite of two EXPLORE Reading forms (1 and 2) mapped to the College Readiness Standards

CRS Strands	Composite of EXPLORE Forms 1 and 2 Mapped to CRS by Panels 1 and 2	
	Range of Knowledge (percent of strands with at least one hit)	
	Panel 1* %	Panel 2* %
Main Ideas and Author's Approach (MID)	62	38
Supporting Details (SUP)	75	88
Sequential, Comparative, and Cause-Effect Relationships (REL)	17	17
Meanings of Words (MOW)	40	60
Generalizations and Conclusions (GEN)	17	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 Reading assessment with the two forms of the EXPLORE Reading assessment, the data were aggregated across panels for a test (Tables 17 and 18). For most categories the two reading panels had very similar results for both NAEP and EXPLORE. The data in Tables 19 and 20 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. A composite of the two EXPLORE forms was used to compute the Range of Knowledge. It is important to note again that each item on the EXPLORE forms was equivalent to one point whereas the items on the NAEP assessment had point values of 1, 2, 3, or 4.

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 and EXPLORE assessments addressed many of the same topics, but not with the same concentration. The biggest difference found between to the two assessments was the 23 percent of the NAEP point values, compared to no items on the

EXPLORE Reading assessment, that targeted objectives under the NAEP Critique/Evaluate reporting category. This category represented more complex topics, such as argumentation. The EXPLORE assessment had a greater proportion of its items than the NAEP assessment that targeted content under the Locate/Recall category.

When the NAEP and EXPLORE assessments were mapped to the CRS, 68 percent of the items on the EXPLORE forms targeted the CRS strand Supporting Details (SUP), whereas the NAEP assessment only had about 25 percent of its point values that mapped to the CRS strand SUP, about the same percentage of NAEP point values that were found to map to the CRS strands Generalizations and Conclusions (GEN) and Meanings of Words (MOW). Only about 5 percent of items on the EXPLORE forms targeted the CRS strand GEN, while around 25 percent of items on the NAEP assessment targeted this same strand. The proportions of items targeting the other three strands were relatively similar, ranging from 1 percent to 11 percent. Thus, the distribution of items was different between the two assessments. The NAEP assessment included items targeting more complex topics (critiquing, evaluate, and forming generalizations) whereas EXPLORE placed greater emphasis on determining supporting details. Both assessments expected students to find the meaning of words in context and placed the lowest emphasis on sequential, comparative, and cause and effect relationships (CRS strand REL).

Depth-of-Knowledge Consistency. The NAEP Reading assessment had a higher average DOK level of items than did the EXPLORE assessment forms, 2.7 compared to an average DOK level of 1.5 (DOK levels range from 1 to 4). The NAEP assessment items tended to be DOK levels 2 or 3, whereas all EXPLORE items were DOK levels 1 or 2. The NAEP Framework had a much higher percentage of objectives with a DOK level of 3 (61 percent) than the percentage of standards in the CRS with a DOK level of 3 (5 percent). Even with the difference in the average DOK level between the two assessments, both assessments had reasonably high DOK Consistency with the corresponding objectives or standard. One explanation for this result is that the EXPLORE Reading assessment targeted the less complex objectives, those assigned DOK 1 or 2, as compared to the NAEP Framework. The NAEP assessment targeted many of the same objectives as well as other objectives classified as DOK 3.

Although DOK Consistency for the NAEP content area of Integrate/Interpret appears higher for the EXPLORE forms (92 percent) than for the NAEP assessment (76 percent), all of the objectives within Integrate/Interpret that were targeted by EXPLORE items were DOK 2 objectives. DOK Consistency for EXPLORE and the NAEP Integrate/Interpret category is, therefore, limited to DOK 2-level objectives. It is important to note that 50 percent of the Integrate/Interpret objectives were considered DOK 3. Although the NAEP items had relatively lower DOK Consistency, NAEP items targeted objectives that were considered DOK 3 as well as objectives that were considered DOK 2.

Range-of-Knowledge Correspondence. The NAEP assessment covered more content than the EXPLORE assessment. This could be expected because of the large difference in the number of items between the two assessments. The NAEP assessment had an acceptable Range of Knowledge on all three of the NAEP Framework reporting categories and on four of the five CRS strands. EXPLORE only had an acceptable Range of Knowledge on the Locate/Recall NAEP reporting category and on three of the five CRS strands. Neither assessment had high coverage of the CRS strand REL, and EXPLORE had low coverage of the CRS strand GEN. Overall, the NAEP Reading assessment targeted a greater breadth of content in nearly all topics with the exception of forming relationships among ideas (sequential, comparative, and cause-and-effect). The EXPLORE Reading assessment was low in coverage on this topic, but also on integrating ideas, making generalizations, critiquing, and evaluating.

The NAEP assessment targeted at least 57 percent of the objectives under each of the given content areas on the NAEP Framework, while EXPLORE emphasized the objectives under Locate/Recall and did not include any items targeting Critique/Evaluate. When the EXPLORE forms were compared to the CRS, the items targeted a greater number of standards when a composite of the two forms was used. The two EXPLORE forms did vary some in the content targeted so that when the two forms are considered as one form the Range of Knowledge is improved. When averaged across the two panels, an acceptable Range of Knowledge was attained for three of the five strands (CRS strands MID, SUP, and MOW). In contrast, NAEP met Range of Knowledge acceptable levels for all but one CRS strand (REL). Overall, the NAEP assessment measured a slightly larger domain of content than EXPLORE.

Balance of Representation. An acceptable Balance of Representation for an assessment-to-assessment analysis implies that neither assessment has a large proportion of items that targets one specific topic that is not over emphasized in content coverage by items on the other assessment. The items on both assessments were distributed fairly evenly among the framework objectives and standards. The one exception was the greater emphasis placed by the NAEP Reading assessment on determining the meaning of a word in context. This was because of the relatively higher number of items on vocabulary (word meaning in context).

Summary of Alignment between the Two Assessments: Overall, the analysis results indicate moderate alignment at a very general level between the NAEP and EXPLORE reading assessments. The NAEP and EXPLORE Reading assessments aligned well in locating and recalling information with literary and informational texts and the less complex aspects of integrating and interpreting of text, but EXPLORE did not cover the topics rated with higher DOK levels. For example, the Panels found no EXPLORE items that corresponded to objectives under the NAEP Critique/Evaluate reporting category. The NAEP assessment had items that sought to measure students' knowledge on all of the topics that EXPLORE did.

- The two assessments differed on DOK Consistency. The NAEP Reading assessment had 28 items that were judged by panelists to be a DOK 3. All 28 items were constructed-response items. The panelists did not find any item on EXPLORE to be judged as a DOK 3.
- The two assessments differed mainly in the degree of emphasis and the level of complexity of the items.

Table 19. Mean number of items and percentages on four alignment criteria for content areas of the NAEP 2013 Reading Framework for grade 8 mapped to the 2013 NAEP Grade 8 Reading Assessment and two forms of EXPLORE Reading

NAEP Content Areas	Assessments mapped to NAEP 2013 Reading 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 objectives with at least one hit)		Balance of Representation (balance index)	
	NAEP	EXP	NAEP	EXP	NAEP	EXP	NAEP	EXP
1 Locate/ Recall	30.6 (11%)	15.6 (52%)	73.3	74.3	78.1	75	0.80	0.84
2 Integrate/ Interpret	178.6 (66%)	14.4 (48%)	75.7	91.8	67.2	23.2	0.62	0.66
3 Critique/ Evaluate	61.3 (23%)	0.04 (0%)	92.7	N/A	56.7	N/A	0.76	N/A
Total Point Value	270.5	30						

*The Range of Knowledge for the composite of the two forms was computed for each panel by counting the standards with at least one item coded by four or more panelists and then dividing by the total possible standards under the strand including the generic standard if appropriate. Then, the composite Range of Knowledge is the average across the two panels.

Table 20. Mean number of items and percentages on four alignment criteria for strands of the ACT College Readiness Standards for EXPLORE mapped to the 2013 NAEP Grade 8 Reading Assessment and two forms of EXPLORE

CRS Strands	Assessments mapped to for grade 8 ACT College Readiness Standards for EXPLORE Reading 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	EXP	NAEP	EXP
Main Ideas and Author's Approach (MID)	49.8 (18%)	3.1 (10%)	98.4	97.2	69.0	50	0.68	0.94
Supporting Details (SUP)	72.5 (26%)	20.5 (68%)	75.5	86.1	93.8	81.3	0.80	0.75
Sequential, Comparative, and Cause-Effect Relationships (REL)	20.2 (7%)	1.8 (6%)	96.4	90	39	16.7	0.73	0.98
Meanings of Words (MOW)	64.1 (23%)	3.5 (12%)	91.0	93.2	60	50	0.65	0.90
Generalizations and Conclusions (GEN)	68.7 (25%)	1.4 (5%)	89.9	81.3	66.7	33.3	0.73	1.00
Total Point Value	275.3	30.3						

The Range of Knowledge for the composite of the two forms was computed for each panel by counting the standards with at least one item coded by four or more panelists and dividing by the total possible standards under the strand including the generic standard if appropriate. Then the average percentage of the composite Range of Knowledge across the two panels was computed.

Reliability of Data

Based on a recommendation from a 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) intra-class correlation 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 researcher who was not a member of the immediate project. A pairwise comparison was used to determine the degree of agreement among reviewers coding items to objective/standards and content area/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 comparison 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 reading panelists' assignment of DOK levels to items was high for each of the 2 panels of eight reviewers for all 12 analyses (Tables 21 and 22). This was true for 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. For all 12 analyses, the intra-class correlations for assigning DOK levels to items 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 were all above 0.81 for each panel and analysis which is reasonably high for most alignment studies. The pairwise agreement in assigning items to specific objective or standards varied from 0.29 to 0.35. 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. Panelists were not required to change their coding unless they found a compelling reason. A few checks were made comparing the results before adjudication and after adjudication. One analysis was done to provide some information on the impact of adjudication for reading. Both reading panels made changes in their codes through the within and between-panel adjudication process when mapping the NAEP Reading assessment to the NAEP Reading Framework. The data for both panels were recorded just after all of the panelists in each group completed their initial coding. The data were compared to the final codes for each panel. At least one member of Panel 1 made changes in their coding of items to the NAEP objectives on 13 items (8 percent of the total number of items). At least one member of Panel 2 made changes in their coding of items to the NAEP

objectives on 14 items (9 percent of the total number of items). For Panel 1, only one panelist changed their coding on eight items, two panelists changed on four items, and three panelists changed on one item. For Panel 2, only one panelist changed their codes on nine items and five panelists changed their codes on five items. Half of the changes made by those in Panel 1 caused a panelist to change to the majority coding within panel and half of changes were influenced by the other panel. However, members of Panel 2 were influenced by the other panel for one of the 14 items.

More changes were made in coding DOK levels to items when coding the NAEP Reading assessment to the NAEP Reading Framework. Panelists from Panel 1 made changes in the DOK levels assigned to items on 18 (11 percent) to 36 (22 percent) of the items. Two panelists from Panel 2 differed significantly from the other six panelists in their assignments of DOK by coding a number of the items as a DOK 3. After adjudication, these two panelists made changes on 65 and 66 items (40 percent), respectively. Another five panelists made changes on the assigned DOK level on 15 (9 percent) to 47 (29 percent) of the items. The pre-adjudication data was not captured for one of the panelists from Panel 2.

Another analysis was performed comparing the impact of adjudication on Panel 2 when coding EXPLORE Form 1 to the NAEP Reading Framework. Panelists made changes in assigning items to objectives for eight items. Two panelists made changes on four items, three panelists made changes on two items, and four panelists made changes on two items. There was much higher agreement on assigning DOK levels to items showing some improvement on understanding the DOK definitions and how to apply them. Over the eight panelists and the 30 items on EXPLORE Form 1, the DOK level was changed on from 1 (3 percent) to 11 (37 percent) of the items.

Summary of Reliability Results. The panelists with adjudication had strong agreement in assigning DOK levels to items and for assigning assessment items to objectives (NAEP) or standards (CRS). The panelists had lower agreement in assigning items to specific expectations underlying the NAEP reporting categories and the CRS strands. These findings were true for all 12 analyses performed—NAEP assessment to NAEP Framework, the two EXPLORE forms to the NAEP Framework, NAEP assessment to the CRS, and the two EXPLORE forms to the CRS.

Table 21. Intra-class correlations, Winer reliability, and pairwise comparisons for the alignment analysis of the 2013 NAEP Grade 8 Reading Assessment and EXPLORE Reading Forms 1 and 2 mapped to NAEP 2013 Reading Framework for grade 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	0.89	0.34	0.94
EXPLORE 1 to NAEP Panel 1	0.92	0.91	0.33	0.89
EXPLORE 2 to NAEP Panel 1	0.90	0.90	0.35	0.89
NAEP to NAEP Panel 2	0.85	0.84	0.35	0.88
EXPLORE 1 to NAEP Panel 2	0.88	0.85	0.32	0.81
EXPLORE 2 to NAEP Panel 2	0.84	0.83	0.29	0.82

Table 22. Intra-class correlations, Winer reliability, and pairwise comparisons for the alignment analysis of the 2013 NAEP Grade 8 Reading Assessment and EXPLORE Reading Forms 1 and 2 mapped to ACT College Readiness Standards for EXPLORE Reading

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	0.89	0.34	0.86
EXPLORE 1 to CRS Panel 1	0.92	0.91	0.33	0.96
EXPLORE 2 to CRS Panel 1	0.92	0.90	0.35	0.95
NAEP to CRS Panel 2	0.85	0.84	0.35	0.90
EXPLORE 1 to CRS Panel 2	0.88	0.85	0.32	0.89
EXPLORE 2 to CRS Panel 2	0.85	0.83	0.29	0.85

Conclusions

In sum, the content analysis results indicated moderate alignment at a general level between the NAEP and EXPLORE reading assessments. The two assessments addressed similar content topics, but they differed on the degree of concentration of items on the topics. There was a marked difference in the content complexity of the two assessments. The NAEP Reading assessment had an average DOK level of items of 2.7 as compared to a 1.5 average DOK for EXPLORE Reading items (DOK levels range from 1 to 4). The NAEP assessment items tended to be DOK levels 2 or 3 whereas the EXPLORE assessment items tended to be DOK levels 1 or 2. The NAEP Reading assessment targeted a greater breadth of content in nearly all topics with the exception of forming relationships among ideas (sequential, comparative, and cause-and-effect). The EXPLORE Reading assessment was low in breadth of content on this topic, but also on integrating ideas, making generalizations, critiquing, and evaluating.

Table 23 summarizes the alignment results for each of the four analyses using the percentage of the reporting categories (cognitive level for NAEP and strand for the ACT College Readiness Standards) with a threshold level for each of the alignment criteria. The threshold levels are those described on pages 7 and 8. These threshold 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 acceptable levels could be used that would result in either improved or lower degrees of alignment.

Table 23. Percent of cognitive levels or strands with acceptable levels for alignment

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=3 categories)	100%	100%	100%	100%
EXPLORE with NAEP Framework (N=3 categories)	67%	100%*	50%*	50%*
NAEP Assessment with College Readiness Standards (N= 5 Strands)	100%	100%	80%	60%
EXPLORE with College Readiness Standards N=5 Strands	20%	100%	60%#	100%

* One of the three cognitive levels for the NAEP Framework, critique/evaluate, did not have sufficient number of items to judge an acceptable level for the category. The percentage is based on two categories rather than three.

Percentage is based on the composite of the two forms.

Process Outcomes and Alignment Results

The study was implemented very closely to the design as described in the Design Document. The process of content analysis at the Content Alignment Institute was carried out by reading teachers that were highly qualified and experienced, and the group was representative of the population of reading teachers 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. All adjudications required by the methodology specified in the Design Document were completed, which included within-group adjudications as well as between-group adjudications. 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 assessment 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 Reading Framework and the CRS performance descriptors for reading 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 (163 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 23, the NAEP Reading assessment and the NAEP Reading Framework are fully aligned. The NAEP assessment and the CRS are moderately aligned. The alignment could be improved by increasing the items on the NAEP Reading

assessment that related to the CRS strand Sequential, Comparative, and Cause and Effect Relationships (REL) and by less emphasis on authors' approach (CRS standards MID 402 and 504) and word vocabulary in context (CRS standard MOW 501).

Comparison of EXPLORE with the Two Frameworks

EXPLORE and the NAEP Framework were weakly aligned. A large majority of the panelists did not find any EXPLORE item that mapped to any of the objectives under the Critique/Evaluate reporting category. EXPLORE also had low coverage on the NAEP Integrate/Interpret category and only targeted the objectives with lower content complexity under this category. As for the NAEP Reading Framework, EXPLORE had an overemphasis on word vocabulary in context. The alignment between EXPLORE and the CRS was moderate. The main alignment issue between EXPLORE and the CRS was the low number of items that targeted four of the five strands. Two-thirds of the items (about 20 items on a form) of the EXPLORE Reading assessment targeted only one strand, CRS strand Supporting Details (SUP). Otherwise, the DOK Consistency, composite Range-of-Knowledge Correspondence, and Balance of Representation met the acceptable criteria applied in each of these areas.

Summary: Comparison of NAEP and EXPLORE

When considering the alignment between the two assessments, the 2013 NAEP Grade 8 Reading assessment and EXPLORE were found to have some degree of alignment, but not to the degree that the two assessments can be considered to be fully aligned. The EXPLORE Reading forms, however, targeted a smaller proportion and breadth of content, in part because of the limitations of having only one-fifth of the number of items as were 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 Reading examinee encounters a subset of the full NAEP item pool of 163 items, whereas each EXPLORE Reading test had 30 items and reported on the performance of individual students.

The two assessments differed in the proportions of items that corresponded to the three content areas of the NAEP Framework. The EXPLORE forms targeted essentially none of the objectives

within the Critique/Evaluate reporting category of the NAEP Framework in contrast with about 25 percent of items on the NAEP assessment that targeted this same content area. Correspondingly, about 50 percent of the items on the EXPLORE forms targeted the Locate/Recall reporting category of the NAEP Framework in contrast with only about 11 percent of the NAEP assessment items.

The two assessments also differed in the proportions of items that corresponded to the five strands of the CRS. The greatest differences were for the CRS strand Supporting Details (SUP) strand and the CRS strand Generalizations and Conclusions (GEN). The EXPLORE forms heavily emphasized standards under the CRS strand SUP, with around 68 percent of items corresponding to this strand. In contrast, only about 26 percent of items on the NAEP assessment corresponded to the CRS strand SUP. Only about 5 percent of items on the EXPLORE Reading forms targeted the CRS strand GEN, while around 25 percent of items on the NAEP assessment targeted this same strand. The proportions of items targeting the other three strands were relatively similar, ranging from 1 percent to 11 percent.

Both assessments had acceptable Depth-of-Knowledge Consistency with both frameworks. More than half of the items had a DOK level at or above the corresponding objective or standard. The NAEP Reading assessment, with nearly 100 constructed-response items, had an average DOK level that was higher than the average DOK level found on the EXPLORE forms. The EXPLORE forms had no items that mapped to the Critique/Evaluate category whereas the NAEP assessments had 25 percent of its items that did. So the two assessments did not have any DOK Consistency with Critique/Evaluate assessment items. For the Locate/Recall category, the DOK Consistency was similar for both assessments. DOK Consistency was greater than the acceptable level for both assessments with the NAEP Integrate/Interpret category, but the EXPLORE assessment had an even greater proportion of items than the NAEP assessment that were at or above the corresponding DOK levels of the objectives within this strand. However, all of the items identified as corresponding to Integrate/Interpret on the EXPLORE forms were assigned a DOK 1 or 2. Even though the EXPLORE assessment had high DOK Consistency with the Integrate/Interpret strand, when the low Range of Knowledge is considered (fewer than 25 percent of the objectives with corresponding items), the EXPLORE assessment had DOK Consistency only with fewer than half of the underlying objectives and none of the objectives

judged to have a DOK level 3. The NAEP form, in contrast, included 28 items that were judged as DOK 3 with many targeting DOK 3 objectives.

Both assessments had similar DOK Consistency on all five strands of the CRS with a variation ranging from about 1 percent to about 11 percent by strand. One caveat is that the EXPLORE forms had very few items corresponding to the four CRS strands other than the CRS strand SUP, and so DOK Consistency applies in some cases to a single item mapped to a strand. Thus, the replacement of just one or two items on an assessment could change the DOK Consistency for an entire strand. In general, the DOK values of EXPLORE assessment items were lower than the DOK values of assessment items on the NAEP assessment.

The results show that the NAEP assessment addresses more reading content than EXPLORE. The large number of items on the NAEP assessment contributes to this difference. One quarter of the NAEP Reading assessment covers more complex content than does the EXPLORE assessment including items that target critiquing and evaluation of text. The one possible area where the NAEP Reading assessment does not fit very well with the EXPLORE Reading assessment is in the coverage of those standards that represent the lower scale scores for the CRS strand Sequential, Comparative, and Cause and Effect Relationships (REL).

The two EXPLORE forms did not cover exactly the same objectives or standards. A composite for the two EXPLORE forms would yield increased content coverage, but still not at the level of the Range of Knowledge of the NAEP assessment. Both assessments emphasized the skill of determining the meaning of a word from a context compared to other objectives and standards.

The analysis results addressed the three key research questions for the study. First, the NAEP and EXPLORE Reading assessments were found to have a large overlap in content coverage. EXPLORE, however, did not cover the more complex topics normally labeled as critiquing, evaluating, and generalizing. Otherwise, the two assessments aligned well in locating and recalling information with literary and informational texts and the less complex integrating and interpreting of text.

Second, the two assessments differed in the proportion of items given to topics. About 15 percent of the NAEP Reading assessment items required more complex reasoning about the text, deeper

inferences about the meaning from the text, and drawing inferences across tests. None of the items on EXPLORE reached this level of complexity. The EXPLORE assessment placed more emphasis on locating details, making simple inferences, and identifying textually explicit information.

Third, regarding significant differences between the two assessments, only one large difference was found between the NAEP and EXPLORE Reading assessments. The findings showed there is a complete absence of any items on EXPLORE that corresponded to objectives under the NAEP Critique/Evaluate reporting category. Otherwise, the two assessments differed mainly in the degree of emphasis and the level of complexity of the items.

In sum, considering all four alignment criteria, a moderate degree of alignment was found between the NAEP and EXPLORE Reading assessments. The NAEP assessment covered all of the content addressed by EXPLORE and at a similar level of complexity, but in addition targeted more content and at a higher content complexity than did EXPLORE.

References

ACT, Inc., 2009. Connecting College Readiness Standards to the Classroom for Language Arts Teachers/Reading. 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).

ACT, Inc., 2013. ACT EXPLORE technical manual, 2013|2014. Iowa City, IA: Author.

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

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.

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, March). 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), 229-258.

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), 2012. *Reading 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/reading/2013-reading-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. “Intraclass 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, 2015.

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.