This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.

RESEARCH REPORT

Measuring Deeper Learning Through Cognitively Demanding Test Items

Results from the Analysis of Six National and International Exams

Kun Yuan, Vi-Nhuan Le

Sponsored by the William and Flora Hewlett Foundation

The research described in this report was sponsored by the William and Flora Hewlett Foundation and was produced within RAND Education, a division of the RAND Corporation.

RAND OFFICES
SANTA MONICA, CA • WASHINGTON, DC
PITTSBURGH, PA • NEW ORLEANS, LA • JACKSON, MS • BOSTON, MA
CAMBRIDGE, UK • BRUSSELS, BE

Preface

In 2010, the William and Flora Hewlett Foundation's Education Program initiated its Deeper Learning Initiative, which focuses on students' development of *deeper learning skills* (the mastery of core academic content, critical-thinking, problem-solving, collaboration, communication, and "learn-how-to-learn" skills). As part of that initiative, the Foundation is interested in monitoring the extent to which deeper learning is assessed in the United States. A prior RAND study examined the extent to which selected state achievement tests measure aspects of deeper learning through cognitively demanding items. This related research assesses the cognitive demand of six nationally and internationally administered tests as a means for interpreting the results of the new generation of assessments being developed by the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers.

This report should be of interest to education policymakers, researchers, and practitioners whose work addresses assessment policy, the Common Core State Standards, and deeper learning.

RAND Education, a unit of the RAND Corporation, conducted the research described in this report. The William and Flora Hewlett Foundation provided the research funding.

Contents

Figures

Tables

Summary

In 2010, the William and Flora Hewlett Foundation's Education Program launched its strategic Deeper Learning Initiative, which focuses on students' development of deeper learning skills (i.e., the mastery of core academic content, critical-thinking, problem-solving, collaboration, communication, and "learn-how-to-learn" skills). As part of that initiative, the Foundation is interested in monitoring the extent to which deeper learning is assessed nationwide in the United States.

Although prior research indicates that state achievement tests have not been measuring deeper learning to a large degree (Polikoff, Porter, and Smithson, 2011; Yuan and Le, 2012), the Common Core State Standards (CCSS) initiative may increase the assessment of deeper learning nationwide. Forty-five states have adopted the CCSS, and two consortia—the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC)—are developing the next generation of assessments, which are designed to measure students' attainment of the standards. It is anticipated that these tests will emphasize deeper learning to a greater extent than other types of large-scale achievement tests, but there has been no systematic empirical examination of the extent to which other widely used achievement tests emphasize deeper learning. In this study, we examined the cognitive demand of six nationally and internationally administered tests. The results of this research will provide the Foundation with a benchmark understanding of the extent to which six these large-scale assessments—and, eventually, the CCSS assessments—measure students' deeper learning.[1]

About the Study

We Examined Six Nationally and Internationally Administered Tests

The six benchmark tests included in this study are administered as part of the Advanced Placement (AP), International Baccalaureate (IB), National Assessment of Educational Progress (NAEP), and Programme for International Student Assessment (PISA) test batteries and also include the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS). NAEP, administered nationally in the United States, is known as the nation's report card because it measures what U.S. students know and can do in core subjects. The other five tests are administered to students worldwide and are

[1] In this report, we refer to assessments designed to measure students' achievement according to the CCSS criteria as *CCSS assessments.* We refer to the six nationally and internationally administered tests examined here as *benchmark tests.*

used to compare students' educational achievement across countries (Provasnik, Gonzales, and Miller, 2009). In this study, we focused on mathematics and English language arts (ELA) tests.

We Applied Two Frameworks to Evaluate the Cognitive Demand of Benchmark Tests

We limited our analysis to three deeper learning skills: critical thinking, problem solving, and written communication. After reviewing multiple frameworks that have been used to describe the cognitive processes of test items and learning tasks, we chose two frameworks to evaluate the cognitive demand of released items from the six selected tests: Norman Webb's (2002b) Depth-of-Knowledge (DOK) framework, which was also used by Smarter Balanced to guide the development of its assessment, and PARCC's self-developed mathematics and ELA frameworks (PARCC, 2012a, 2012b).

Webb defines four levels of cognitive demand. Level 1 represents recall, level 2 represents the demonstration of a skill or understanding of a concept, level 3 represents strategic thinking, and level 4 represents extended thinking. In our analysis, we applied Webb's subject-specific descriptions for each of the DOK levels for mathematics, reading, and writing in our analysis.

PARCC provides two separate frameworks to describe the cognitive demand for mathematics and ELA, respectively. Cognitive demand is defined in terms of sources of cognitive complexity. Five sources of cognitive complexity contribute to the cognitive demand of mathematics items: mathematical content, mathematical practices, stimulus material (e.g., tables, graphs, figures, technology tools), response mode, and processing demand. Four sources of cognitive complexity contribute to the cognitive demand of ELA items: text complexity, command of textual evidence, response mode, and processing demand. We revised the ELA framework to include stimulus material to accommodate potential sources of cognitive complexity intrinsic to the technological component of the PISA ELA test.

Although the PARCC framework provides guidelines for combining the various dimensions to create an overall complexity score, we deviated from the recommended scoring mechanism. The scoring rubric gave relatively greater weight to the difficulty of the content and relatively less weight to cognitive processes, and we found that this approach did not work well for open-ended items, particularly in English. For example, a short writing prompt that asked for a sophisticated analysis of multilayered ideas rated as only moderately demanding under this scoring mechanism, despite being a complex task. To better capture the skills emphasized by the Deeper Learning Initiative, we revised the scoring mechanism to give 40-percent weight to mathematical practices, 25-percent weight each to mathematical content and response mode, and 5-percent weight each to stimulus material and processing demands. For ELA, we gave 40-percent weight to command of textual evidence, 25-percent weight each to text complexity and response mode, and 5-percent weight each to stimulus material and processing demands. Our modifications did

not result in appreciably different ratings, as the PARCC scoring mechanisms and our ratings were correlated at 0.91 in ELA and 0.93 in mathematics.

While the DOK ratings provided a straightforward classification of deeper learning (i.e., DOK ratings of 3 or higher were indicative of deeper learning), we did not have similar guidelines for the PARCC ratings. To increase the comparability of the two frameworks, we created cut scores for the PARCC ratings by examining the ratings' distribution and making holistic judgments about the cognitive demand of the items associated with each rating. We then converted the PARCC ratings to a four-category rating system. For the PARCC four-category classification, we interpreted a rating of 1 as representing a very low level of cognitive demand, 2 a low to medium level of cognitive demand, 3 a medium to high level of cognitive demand, and 4 a very high level of cognitive demand.

In examining the correspondence between the two frameworks' four-category ratings, we computed a weighted kappa value, which is a measure of rater agreement that takes into account of agreement due to chance. We observed a weighted kappa of 0.56 for ELA and 0.59 for mathematics. If we dichotomized the ratings and examined the correspondence between items considered indicative of deeper learning (i.e., ratings of 3 or higher) and those that were not, we observed a kappa of 0.74 for ELA and 0.67 for mathematics. Furthermore, we did not find that one framework gave systematically higher ratings to items. For the majority of the items, the PARCC and DOK frameworks classified a given item as demonstrating deeper learning (or not) in the same manner.

We analyzed the most recent version of the released test items for the six tests, with administration dates ranging from 2008 to 2011. In total, we analyzed 790 mathematics items and 436 ELA items, including 418 reading and 18 writing items. About half of the mathematics items required multiple-choice (MC) answers, and the other half required open-ended (OE) answers. About two-thirds of the reading items were MC items. All writing items were OE items.

Two researchers rated the cognitive demand of the released items from the six tests using the DOK and PARCC frameworks. The weighted kappa interrater reliability was high, ranging from 0.89 to 1 for both mathematics and ELA.

Findings

The Six Benchmark Tests Had Greater Cognitive Demand Than the State Tests

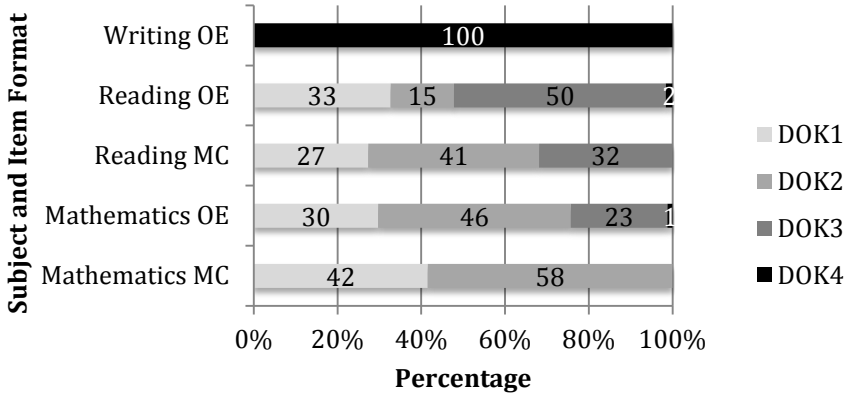
On average, the six benchmark tests demonstrated greater cognitive demand than did the state achievement tests in both subjects. The average share of items rated at or above DOK level 3 was

about 15 percent for mathematics and 40 percent for ELA across the six benchmark tests (see Figure S.1), compared with 2 percent for mathematics and 20 percent for ELA across the 17 state achievement tests included in our earlier study (see Yuan and Le, 2012).

The Cognitive Demand of Test Items Varied by Subject and Item Format

The overall composition patterns of the cognitive demand for the six benchmark tests were similar to what was observed for the state achievement tests (see Yuan and Le, 2012). In addition, the cognitive demand of the ELA tests was greater than that of the mathematics tests (see Figure S.1). Format is associated with the cognitive demand of items, with OE items being more cognitively demanding than MC items, as shown in the figure.

Figure S.1. Percentage of Test Items Rated at Each DOK Level, by Subject and Item Format



NOTE: Results were rounded up to integers.

The Six Benchmark Tests Varied in Their Percentages of Cognitively Demanding Items

The six benchmark tests varied in their percentages of cognitively demanding items. IB and AP had higher percentages of cognitively demanding items than other benchmark tests in both subjects. TIMSS and PIRLS appeared to be less cognitively demanding than other benchmark tests. By and large, results were similar between the two frameworks in terms of the percentage of items rated at higher levels (3 and 4). There were some differences between the two frameworks in terms of the percentage of items rated at or above level 2. Several factors might have contributed to such differences, such as the sources of complexity considered, weights assigned to each source, and the features of each test that serve as key sources of complexity.

Only Two Benchmark Tests Met Both Criteria for High-Quality Measures of Deeper Learning

We used Darling-Hammond et al.'s (2013) framework that proposes a set of five criteria to determine whether a measure should be considered a high-quality assessment of higher-order

cognitive skills. We focused on the two criteria that could be assessed with the data from our study. Criterion I recommends that at least two-thirds of the test items be rated at or above DOK level 2. Criterion II recommends that at least one-third of mathematics items and half of ELA items be rated at or above DOK level 3. We extended these two criteria to the PARCC framework and examined the extent to which each of the six selected tests met the two criteria for high-quality measurement of higher-order cognitive skills under the two frameworks.

We found that the six benchmark tests varied in terms of the extent to which they met these two criteria (see Table S.1).

Table S.1. Whether a Benchmark Test Met Two Criteria for High-Quality Measures of Higher-Order Cognitive Skills Based on Two Frameworks

Subject	Test	DOK		PARCC	
		Criterion I	Criterion II	Criterion I	Criterion II
Mathematics	AP	✔		✔	
	IB	✔		✔	✔
	NAEP				
	PISA	✔			
	TIMSS				
ELA	AP	✔	✔	✔	✔
	IB	✔	✔	✔	✔
	NAEP	✔			
	PISA			✔	
	PIRLS				

NOTE: Criterion I indicates that at least two-thirds of the test items are rated at level 2 or higher. Criterion II indicates that at least one-third of the mathematics items and half of the ELA items are rated at level 3 or higher.

IB mathematics and ELA tests met both criteria under at least one framework. AP ELA tests met both criteria according to both frameworks. AP mathematics tests met Criterion I but not Criterion II according to both frameworks. PISA mathematics and ELA tests met Criterion I under one framework. Neither PISA's mathematics nor ELA tests met Criterion II under either framework. The NAEP mathematics test did not meet any of the criteria according to either framework. The NAEP ELA test met Criterion I according to the DOK framework but not the PARCC framework, and it did not meet Criterion II under either framework. Neither TIMSS nor PIRLS met the two criteria for high-quality assessments of higher-order cognitive skills.

Cognitive Demand Level Varied with Test Purpose and the Characteristics of Target Students

The findings also indicated that the percentage of cognitively demanding items on the six benchmark tests was associated with the purpose of the test and the characteristics of the targeted student population. The IB and AP tests assess students' readiness for postsecondary academic learning and target academically advanced high school students. In contrast, PISA, NAEP, TIMSS, and PIRLS assess what students know and can do, and these tests are administered to

students at all academic performance levels. Commensurately, PISA, NAEP, TIMSS, and PIRLS had proportionately fewer cognitively demanding items than the IB and AP tests.

Implications for The Foundation's Deeper Learning Initiative

This study has several implications for the Foundation as it gauges progress toward the Deeper Learning Initiative's goal of increasing the emphasis placed on deeper learning. First, although prior studies indicate that the CCSS assessments have the potential to place greater emphasis on deeper learning than most current state assessments, our results show that it is difficult to create high-quality deeper learning assessments in practice, especially when such tests will be used to measure the academic achievement of students at all performance levels. This suggests that it is necessary to analyze the operational forms of the CCSS assessments to understand the extent to which they will actually measure deeper learning when they are available in 2015.

Second, it is important to recognize that the tests differed with respect to their goals and targeted student populations, both of which affect the level of cognitive demand we can expect to observe. Measures such as the AP tests, which are intended to assess mastery of college-level content, can be expected to have a higher level of cognitive demand than measures such as NAEP, which is intended to assess the knowledge and skills that students at a given grade level should ideally demonstrate. The results from this study suggest that future analysis of the CCSS assessments should choose tests with similar purposes and targeted student populations as benchmark tests for comparisons. Given that the CCSS assessments will measure students at all performance levels, results pertaining to PISA, NAEP, TIMSS, and PIRLS arguably provide a better benchmark for future analysis of the CCSS assessments than do results from the IB and AP tests.

Third, future evaluations of the Deeper Learning Initiative may encounter the same types of challenges as this study, such that only a limited type of deeper learning skills can be examined. The CCSS assessments may not assess the intrapersonal and interpersonal competencies that are also part of the larger deeper learning construct advocated by the Foundation. Measures of intrapersonal and interpersonal skills are limited and have unknown validity and reliability (NRC, 2012; Soland, Hamilton, and Stecher, 2013). Given the current assessment landscape, the Foundation may have to make trade-offs with respect to psychometric properties, costs, and other considerations to assess the full range of deeper learning skills outlined in its Deeper Learning Initiative.

Fourth, our results indicate the need to develop frameworks that would allow an analysis of the mastery of core conceptual content as integrated with critical thinking and problem solving in each subject area. There is increasing evidence supporting the interdependence between critical-

thinking and problem-solving skills and fluency with the core concepts, practices, and organizing principles that constitute a subject domain (Schneider and Stern, 2010). Although the CCSS provides foundational knowledge and concepts for ELA and mathematics, it does not delineate skills and knowledge by grade level in the upper grades, so it is difficult to apply these standards to tests geared toward high school students, who constitute the majority of those who take the tests in our sample. Future studies examining the Foundation's Deeper Learning Initiative should consider using CCSS or other frameworks that define foundational concepts and knowledge for each subject area when assessing the cognitive demand of a given test item.

Study Limitations

There are several caveats worth noting when interpreting the results of this study. First, as a simplifying assumption, we treated cognitive demand as a fixed characteristic of the test item. However, it is important to recognize that the cognitive demand of an item as experienced by the examinee is a function of the interface between the individual's personal attributes, the testing environment, and the skills and knowledge being elicited by the test item (Kyllonen and Lajoie, 2003).

Second, we relied on released test items to examine the cognitive demand of the six benchmark tests. The degree to which these items are representative of the entire sample pool from which they are drawn varies across tests. Differences in the representativeness of released items among six benchmark tests might have introduced bias in the evaluation of the cognitive demand of these tests; however, the direction of this potential bias is unknown.

Finally, in our study, we defined a high-quality assessment in terms of the percentage of test items that assessed deeper learning. There are other ways to evaluate the extent to which a test emphasizes deeper learning, such as the proportion of the total score awarded for items that assess deeper learning, or the amount of time devoted to deeper learning items. We did not examine these alternative measures because we lacked the data to do so.

Acknowledgments

This study could not have been completed without assistance from the following individuals. We would like to thank Edys Quellmalz and P. David Pearson for their insightful comments on the analysis plan and choice of analytical framework. We also thank Joan Herman and Rebecca Buschang for sharing their expertise with the two analytic frameworks described in this report. We are grateful to the sponsor of this project, the William and Flora Hewlett Foundation, and the support and feedback we received from the Foundation's Education Program staff, especially Denis Udall.

We thank Christine Massey of the University of Pennsylvania and Anna Saavedra and Laura Hamilton at RAND for their thoughtful reviews and comments. We would also like to thank Lauren Skrabala for her edits and Diane Bronowicz and Sharon Koga for their administrative assistance.

Abbreviations

AP	Advanced Placement
CCSS	Common Core State Standards
DOK	Depth-of-Knowledge
ELA	English language arts
EOC	end-of-course
IB	International Baccalaureate
MC	multiple-choice
NAEP	National Assessment of Educational Progress
NRC	National Research Council
OE	open-ended
PARCC	Partnership for Assessment of Readiness for College and Careers
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
TIMSS	Trends in International Mathematics and Science Study

1. Introduction

The Deeper Learning Initiative

In 2010, the William and Flora Hewlett Foundation's Education Program launched its Deeper Learning Initiative, which emphasizes students' mastery of core academic content and their development of deeper learning skills. The initiative focuses on enabling students to attain the following types of deeper learning skills:

1. *Master core academic content:* Students will develop disciplinary knowledge, including facts and theories in a variety of domains—and the language and skills needed to acquire and understand this content.
2. *Think critically and solve complex problems:* Students will know how and when to apply core knowledge by employing statistical reasoning and scientific inquiry to formulate accurate hypotheses, offer coherent explanations, and make well-reasoned arguments, along with other skills. This category of competencies also includes creativity in analyzing and solving problems.
3. *Work collaboratively:* Students will cooperate to identify or create solutions to societal, vocational, and personal challenges. This category includes the ability to organize people, knowledge, and resources to achieve a goal and to understand and accept multiple points of view.
4. *Communicate effectively:* Students will be able to understand and transfer knowledge, meaning, and intention. This category involves the ability to express important concepts, present data and conclusions in writing and to an audience, and listen attentively.
5. *Learn how to learn:* Students will know how to monitor and direct their own learning.[1]

As part of its efforts to promote deeper learning, the Foundation also launched a corresponding research initiative intended to support the development of seven model school networks that embody the deeper learning approach, as well as assessments of deeper learning.

[1] These five skills represented the Deeper Learning framework at the start of this study and guided our analysis. The Foundation has since revised the initiative's framework. *Self-directed learning* replaced *learn how to learn* and emphasizes students' ability to set goals, monitor their own progress, and reflect on their strengths and areas for improvement. The Foundation also added *academic mindset*, which emphasizes students' belief in themselves and their ability to persist in the face of obstacles.

Guidelines on High-Quality Assessments of Deeper Learning

A recent report by the National Research Council (NRC) grouped deeper learning skills into three broad domains of competence: cognitive, intrapersonal, and interpersonal (NRC, 2012). The cognitive domain refers to cognitive processes and strategies, knowledge, and creativity. The intrapersonal domain represents an individual's capacity to manage his or her behavior and emotions to achieve a goal. The interpersonal domain refers to one's ability to communicate with others, such as expressing ideas and receiving, interpreting, and responding to messages. Reviewing the current status of assessments of deeper learning skills, the NRC found that existing standardized achievement tests mainly measure the cognitive aspects of deeper learning, and it concluded that skills in the cognitive domains have been studied more extensively than skills in the intrapersonal and interpersonal domains.

Recently, a panel of curriculum, teaching, and assessment specialists provided guidelines for developing criteria for a high-quality assessment of deeper learning (Darling-Hammond et al., 2013). According to these experts, a high-quality assessment of deeper learning should meet the following five criteria:

- Assess higher-order cognitive skills.
- Measure critical abilities with high fidelity as they will be used in the real world.
- Benchmark with international assessments.
- Include items that are instructionally sensitive and educationally valuable.
- Provide a valid, reliable, and fair assessment.

This panel of experts also provided specific recommendations regarding the criterion for assessing higher-order cognitive skills. The panel used Norman Webb's Depth-of-Knowledge (DOK) framework as a metric to measure the cognitive demand of test items (see Webb, 2002b). The DOK framework uses a holistic rating scale, on items are categorized into four levels based on the complexity of thinking required to answer them. Level 1 represents recall, level 2 represents the demonstration of skill or concept, level 3 represents strategic thinking, and level 4 represents extended thinking. Items at DOK level 3 or 4 were considered to measure "higher-order" cognitive skills.

These experts set two criteria for high-quality measures of high-order cognitive skills. They recommended that at least two-thirds of items on high-quality mathematics and English language arts (ELA) tests be rated at or above DOK level 2. Moreover, at least one-third of items on high-quality mathematics tests should be rated at DOK level 3 or 4. For ELA, at least half of the items should be rated at DOK level 3 or 4.

Current Status of the Assessment of Deeper Learning

To assess the impact of the Deeper Learning Initiative, the Foundation is interested in monitoring changes in the assessment of deeper learning nationwide for the duration of the initiative. As the first step of this effort, the Foundation commissioned a RAND study to examine the extent to which deeper learning was assessed on state achievement tests at the outset of the initiative.

In that study, we used Webb's DOK framework to examine the cognitive demand of released test items from achievement tests administered in 17 states.[2] Adopting the standard that DOK level 4 was indicative of deeper learning, we found that the percentage of items rated at DOK level 4 was low: 0 percent for mathematics, 1 percent for reading, and 7 percent for writing (Yuan and Le, 2012).[3] Following the recommendations from Darling-Hammond et al. (2013) that DOK level 3 or higher is indicative of deeper learning, we reanalyzed the data and found that the extent to which deeper learning was assessed through the state achievement tests was still low: 2 percent of mathematics items, 22 percent in reading, and 21 percent in writing.

Using a different methodology that examined the alignment between the achievement tests and content standards in 19 states, Polikoff, Porter, and Smithson (2011) found similar evidence that the extent to which state achievement tests measured deeper learning was low. They reported that 80 percent of mathematics items and 52 percent of reading items assessed lower-level skills, whereas 7 percent of mathematics items and 33 percent of reading items assessed higher-order skills. Thus, neither study found that the current state tests approach the recommended levels of cognitive demand.

The Common Core Standards Initiative

The advent of the Common Core State Standards (CCSS) initiative has the potential to improve the degree to which deeper learning is assessed through state achievement tests. The CCSS initiative establishes a single set of educational standards for kindergarten through 12th grade that identifies the concepts and knowledge that students should acquire to show that they have

---

[2] These 17 states were chosen because they were reputed to have cognitively demanding achievement tests that addressed deeper learning. Choosing these 17 states might have introduced bias into our findings regarding the extent to which state achievement tests measured deeper learning *nationwide*. However, because project resources did not allow us to include all states in our earlier study (see Yuan and Le, 2012), we decided to focus on the state assessments with the highest likelihood of being cognitively demanding so as to provide an upper bound on the extent to which deeper learning is being assessed on state achievement tests.

[3] The percentage of cognitively demanding items is only one possible measure of the extent to which a test measures deeper learning skills. Other measures, such as the proportion of testing time devoted to cognitively demanding items or the proportion of test scores attributable to cognitively demanding items, should also be considered when assessing the extent to which a test measures deeper learning skills. However, a lack of information about the amount of testing time earmarked for each item type and the number of score points assigned to each item prevented us from taking these factors into account in our 2012 study.

attained the skills necessary for college and career success. The standards represent a significant departure from previous standards in that content is integrated across multiple subjects. For example, the CCSS for ELA define reading, writing, and oral language as tools for effective communication across the disciplines of literature, science (and other technical subjects), history and social studies (NRC, 2012). Similarly, the CCSS for mathematics define the use of mathematical skills in disciplines such as science, technology, and engineering (Darling-Hammond et al., 2013). In addition, the standards place greater explicit emphasis on college and career readiness, including research skills, textual analysis, and the ability to write and deliver logical arguments.

Forty-five states have adopted the CCSS standards, and two consortia, the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC), are currently developing the next generation of assessments that are designed to measure students' attainment of the CCSS. The consortia are developing a system of assessments that include both diagnostic or interim measures, as well as summative measures administered via computer. In the case of Smarter Balanced, the assessments will be computer-adaptive, in which the computer program adjusts the difficulty of the test items based on examinees' responses. The assessments are slated to be fully operational during the 2014–2015 academic year.[4]

The implementation of the CCSS assessments represents a unique opportunity for schools and districts to raise their expectations for student learning. Research has shown that assessments send a powerful signal of valued skills and knowledge (Hamilton, Stecher, and Yuan, 2008; Herman, 2004), and teachers will focus on content that is tested and deemphasize content that is not (Koretz and Hamilton, 2006; Faxon-Mills et al., 2013). If the Smarter Balanced and PARCC assessments emphasize deeper learning, there is the potential for the tests to greatly influence classroom instruction and student learning.

Because Smarter Balanced and PARCC have yet to release any operational test items, it is not possible to determine the cognitive demand of the CCSS assessments. However, Herman and Linn (2013) conducted a study of the Smarter Balanced and PARCC content specifications that can serve as a basis for understanding the extent to which these next-generation tests may promote deeper learning. Herman and Linn noted that both consortiums reorganized the standards into claims about student competency. For example, PARCC translated the reading standard as follows: "Students read and comprehend a range of sufficiently complex texts independently." Smarter Balanced translated the standard in this way: "Students can read closely

[4] In this report, we refer to assessments designed to measure students' achievement according to the CCSS criteria as *CCSS assessments.*

and analytically to comprehend a range of increasingly complex literary and informational texts." Each claim was then defined by specific evidence statements (PARCC) or assessment targets (Smarter Balanced), which identified the content domains to be assessed and the associated performance expectations. These expectations served as the targets for the content specifications that guided item development.

At the time of Herman and Linn's analysis, Smarter Balanced was further along in its test development, so much of the report focused on Smarter Balanced. A unique feature of the Smarter Balanced content specifications was that they listed the DOK levels at which each assessment target was to be assessed. Herman and Linn (2013) reviewed the assessment targets for grades 4, 8, and 11 in ELA and grades 3–8 and grade 11 in mathematics, noting the DOK level associated with each assessment target. The results are summarized in Table 1.1.

Table 1.1. Mean Percentage of Smarter Balanced Content Targets Rated at Each DOK Level

Level	ELA	Mathematics
Mean number of content targets	35	29
DOK 1	33%	46%
DOK 2	46%	79%
DOK 3	43%	49%
DOK 4	25%	21%

SOURCE: Herman and Linn, 2013.

NOTE: Because the assessment targets could be associated with multiple DOK levels, the results do not add to 100 percent.

As shown in Table 1.1, Herman and Linn (2013) found that 68 percent of the ELA targets and 70 percent of the mathematics targets had DOK ratings of 3 or above. A qualitative analysis of the content specifications for the PARCC assessments found similar levels of intended cognitive complexity (Herman and Linn, 2013). Assuming that the percentage of targets with DOK ratings of 3 or above provide an indication of deeper learning, at least initially, there appears to be some evidence suggesting that the Smarter Balanced test specifications may result in cognitively demanding assessments that have the potential to improve student learning. However, the ultimate determination as to whether the CCSS initiative results in high-quality assessment systems that promote deeper learning depends on the extent to which the final versions of the assessments faithfully represent the assessment targets.

Purpose of This Study

Given the potential impact of the CCSS initiative on the measurement of deeper learning nationwide, the Foundation is interested in assessing the extent to which the actual CCSS assessments developed by Smarter Balanced and PARCC measure deeper learning. However, an analysis of the CCSS assessments will not be available until the 2014–2015 academic year, when

the tests will be fully operationalized. Although it is anticipated that these tests will emphasize deeper learning to a greater extent than previously observed among other types of large-scale achievement tests, there has yet to be a systematic examination of the extent to which these other large-scale achievement tests emphasize deeper learning.

In this study, we examined the cognitive demand of six nationally and internationally administered tests. The results of this study will provide the Foundation with a benchmark understanding of the extent to which six large-scale assessments assess students' deeper learning.[5] Once the CCSS assessments are released and analyzed, the results of this study will help contextualize future analysis of the extent to which CCSS assessments measure deeper learning compared with the benchmark tests.

The six benchmark tests in our study are administered as part of the Advanced Placement (AP), International Baccalaureate (IB), National Assessment of Educational Progress (NAEP), and Program for International Student Assessment (PISA) test batteries and also include the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS). NAEP, administered nationally in the United States, is known as the nation's report card because it measures what U.S. students know and can do in core subjects. The other five tests are administered to students worldwide and are used to compare students' educational achievement across countries (Provasnik, Gonzales, and Miller, 2009). In this study, we focused specifically on mathematics and ELA tests.

This report addresses the following research questions:

1. On each selected test, what percentage of items are considered cognitively demanding?
2. To what extent did each selected test meet the two criteria for a high-quality assessment of deeper learning?

Information about the proportion of cognitively demanding items in these benchmark tests will be useful for the Foundation to set realistic expectations about the percentage of assessment items that can be reasonably expected to address higher-order skills. The CCSS initiative is intended to help students compete successfully in a global economy, so this study also lays the groundwork for understanding how cognitively demanding the CCSS assessments are compared with selected international tests. Moreover, findings regarding the cognitive demand of these tests at the outset of the Deeper Learning Initiative will be useful in assessing whether the CCSS initiative is associated with changes in the cognitive demand of the benchmark tests or state achievement tests over time.

[5] In this report, we refer to the six nationally and internationally administered tests examined here as *benchmark tests*.

Structure of This Report

We conducted our study in three steps. Because these benchmark tests were developed for different purposes and target different student populations, differences among them might affect findings regarding their respective percentages of cognitively demanding test items. The six tests might also have their own frameworks that categorize the cognitive level of test items in a particular way. Moreover, the representativeness of items that could feasibly be included in this project has implications for the utility of our findings. Therefore, we began by conducting an online search for information about each benchmark test and any prior research comparing the tests. Chapter Two provides descriptions of each benchmark test and a review of the relevant literature.

After developing an understanding of each test to be included in the analysis, we identified frameworks that could be used to analyze the extent to which test items measure deeper learning skills. We then applied the selected analytical frameworks to the six tests to assess the extent to which they did so. In Chapter Three, we describe the alternative frameworks we considered, how we chose the frameworks for analysis, and our methods for applying the selected analytical frameworks to analyze the selected tests.

Chapter Four presents our findings regarding the percentage of items on each benchmark test classified as cognitively demanding and the extent to which each test met the two criteria for the high-quality assessment of deeper learning. Chapter Five discusses the implications for testing policy and the limitations of this project.

This report also includes six appendixes. Appendixes A and B provide additional information about the distributions of NAEP and TIMSS items as specified by their frameworks. Appendixes C and D present sample test items rated at each level in our analytical frameworks. Appendix E gives the distribution of modified PARCC ratings, and Appendix F presents results for each dimension of the PARCC framework.

2. Tests Included in This Study

In this chapter, we provide a brief description of each test, including its purpose, the framework used to guide its design, test item formats, subjects and grades tested, the number of released items analyzed, and the degree to which analyzed items were representative of their entire sample pools. We also present a review of prior research that compared benchmark tests.

Advance Placement Programs and Exams

AP offers college-level courses to advanced high school students around the world. AP reaches a large number of schools and students worldwide, with an enrollment of more than 2 million students from 18,000 schools around the world. In the United States, 32 percent of public high school students in the class of 2012 took at least one AP exam during high school.

In 2012–2013, AP courses in mathematics included Calculus AB, Calculus BC, Computer Science A, Computer Science AB, and Statistics. AP ELA courses included English Language and Composition, and English Literature and Composition. In this study, we focused on the end-of-course (EOC) exams for three mathematics courses (Calculus AB, Calculus BC, and Statistics) and both ELA courses.

Panels of subject-matter experts and college faculty develop the AP curricula and exams. AP exams are administered annually in May. Each of the five exams we studied includes two sections, one consisting of only multiple-choice (MC) items and the other including only open-ended (OE) items (College Board, 2014). The Calculus AB and BC exams have 45 MC items and six OE items. The statistics exam has 40 MC items and six OE items. MC and OE items contribute equally to the total score on the exam. The English Language and Composition test and the English Literature and Composition test have 54 and 55 MC items, respectively. Both tests have three essays. MC items contribute to 45 percent of the total score, and essays contribute to 55 percent.

Complete AP test forms were available for analysis. We examined all items in the most recently released versions of these five AP exams, including 193 items from the 2008 Calculus AB and BC tests and the 2007 Statistics test, along with six writing items from the 2007 English Language and Composition test and the 2009 English Literature and Composition test.

International Baccalaureate Programmes and Exams

IB offers three worldwide education programs for children aged 3–19, including the Primary Years Programme (ages 3–12), the Middle Years Programme (ages 11–16), and the Diploma Programme (ages 16–19). More than 1.1 million students from 3,662 IB schools in 146 countries have enrolled in these programs (International Baccalaureate, undated). In the United States, there are 1,465 IB schools, with 394 Primary Years Programs, 487 Middle Years Programs, and 800 Diploma Programs (International Baccalaureate, undated).

In our study, we focused on the mathematics and language EOC exams in the Diploma Programme. The Diploma Programme is a two-year course of study for advanced high school students. Its curriculum includes six academic areas: language, second language, experimental sciences, individuals and societies, mathematics and computer science, and the arts. In each area, courses are categorized into two levels based on the recommended teaching times of 240 hours for courses at the high level and 150 hours for courses at the standard level.

Students take EOC exams administered in May and November after they finish each course. We analyzed the EOC exams for mathematics and English courses at both the high and standard levels. All exams analyzed consist of only OE items. All exams have two sections with the exception of mathematics exams at the high level, which have three sections. The third section of the mathematics exam at the high level is divided into four subsections.

Each section of the standard-level exams has a potential total raw score of 90. Each section of the high-level exams has a potential total raw score of 120, except that the third section of the mathematics exam at the high level has a total raw score of 240. These raw scores are converted to a seven-point scale grade (1 = lowest; 7 = highest), which is combined with results on other required components of the program and used for college admission and placement.

Complete IB test forms were available for analysis. We analyzed all 157 mathematics items and six writing items in the most recently released IB exams, including mathematics exams used in November 2011 and language exams used in November 2010.

National Assessment of Educational Progress

NAEP is the largest national assessment of what U.S. students know and can do in core subjects, such as mathematics, reading, science, and writing. There are two types of assessments: the main NAEP and the long-trend NAEP. For this study, we focused on main NAEP mathematics, reading, and writing tests, which are administered to students at all performance levels in grades 4, 8, and 12 across the country. The National Assessment Governing Board develops frameworks

to guide the development of NAEP tests. In the following sections, we briefly review the framework and test for each subject.

NAEP Mathematics Framework and Test

The framework for mathematics defines five content areas in which students are assessed, including number properties and operations, measurement, geometry, data analysis, statistics and probability, and algebra (National Assessment Governing Board, 2010a).[1] The framework also defines three levels of mathematical complexity, including low, moderate, and high, with items at the low level focusing on recalling and recognizing and items at the high level emphasizing the use of higher-order thinking skills. The National Assessment Governing Board specified that the share of mathematics test items at each level should be 25, 50, and 25 percent, respectively.

The NAEP mathematics test uses both MC and OE items. Testing time for the mathematics test is divided evenly between the two types of items. The mathematical complexity of an item is expected to be unrelated to its format. Moreover, the context of items should be kept balanced between purely mathematical ideas and concepts and real-world problems.

NAEP Reading Framework and Test

The framework for reading specifies two types of texts: literary texts and informational texts (National Assessment Governing Board, 2010b). The former includes fiction, literary nonfiction, and poetry. The latter includes exposition, argumentation and persuasive text, and procedural text and documents.[2] The NAEP reading framework also specifies three groups of cognitive skills that reading items should measure: locate/recall, integrate/interpret, and critique/evaluate.

The reading test includes both MC and OE items. The percentage of each type of item differs by grade level. Students in grade 4 are expected to split their assessment time about evenly between MC and OE items. Students in grades 8 and 12 spend more time on OE items.

NAEP Writing Framework and Test

NAEP writing tests for grades 4, 8, and 12 consist of two extended-response writing tasks, each of which takes 30 minutes to complete (National Assessment Governing Board, 2010c). Students need to write both essays in English. For each writing task, students may be asked to complete three types of writing: persuasive, informative, and narrative. Informative writing stresses the subject matter that is being explained. Persuasive writing emphasizes making an impact on the reader. Narrative writing focuses on students' experiences, perceptions, and imagination.

[1] See Appendix A for the distribution of NAEP mathematics items by grade and content area.

[2] See Appendix A for the distribution of literary and information passages on the NAEP reading test.

Collectively, the three types of writing items aim to assess students' ability to persuade, explain, and convey real or imagined experiences in school and in the workplace. [3]

NAEP Released Test Items Analyzed

In this study, we analyzed 190 mathematics items, 147 reading items, and six writing items released from the 2009 and 2011 NAEP administrations. These items are only a selection of those used to compile the NAEP test forms and may not fully represent the complete range of content coverage, cognitive skills, and difficulty levels of the entire NAEP item pool for a subject.

Progress in International Reading Literacy Study

PIRLS is an international comparative study of the reading achievement, behavior, and attitudes of fourth graders. PIRLS was administered originally in 2001 and every five years since then. The most recent administration was in 53 education systems (including countries and subnational entities) in 2011.

Students at all performance levels in the sampled schools take the PIRLS tests. Students complete a reading test and a series of surveys of about their reading behaviors and attitudes toward reading. For the purposes of this study, we focused on the reading test. This test assesses students' reading skills in two areas: reading for literary experience and reading to acquire and use information. It also measures students' skills in constructing meaning from a text in four different ways: retrieving information, making inferences, interpreting and integrating, and examining and evaluating. The 2011 PIRLS framework specifies that the reading texts used in the test should be divided evenly between literary experience and informational texts, and the share of items addressing the four types of comprehension processes should be 20, 30, 30, and 20 percent, respectively.

The PIRLS test includes five literary and five information passages in total. Each passage comes with approximately 12 questions, half MC questions and half OE questions. Each passage and its associated questions are considered a "block". The total score on each block is about 15 points, with one point per MC item, one or two points per short OE item, and three points per extended OE item. Each student answers two blocks of questions during the test.

In this study, we analyzed four blocks of passages and 54 associated questions released from the 2006 and 2011 PIRLS administrations. Although these items are only a portion of those used to

[3] See Appendix A for the distribution of NAEP writing items by grade and writing goals.

compile the PIRLS test forms, these questions are representative of the PIRLS item pool (Martin and Mullis, 2012).

Program for International Student Assessment

PISA is an international assessment that measures 15-year-old students' reading, mathematics, and science literacy. It was first conducted in 2000 and has been administered every three years thereafter. More than 70 countries have participated in PISA.

PISA organizes items in units around a common stimulus for all three subjects (OECD, 2009). Each unit has up to five MC or OE questions. All test units are organized into 13-item clusters. The number of item clusters for a subject varies with each administration because PISA focuses on a particular subject during each administration. Each item cluster takes 30 minutes to complete. These item clusters are assigned to 13 test booklets, with each booklet containing four clusters. Students at all performance levels in the sampled schools take the PISA tests. They are randomly assigned to complete one of the 13 test booklets in two hours. We focused on the mathematics and reading tests in this study.

Similar to NAEP and PIRLS, PISA has a detailed framework to guide the development of tests. In the following section, we briefly describe the mathematics and reading frameworks.

PISA Mathematics Framework

The PISA mathematics framework categorizes mathematics test items along three dimensions: the situations or contexts in which the problems are located, the mathematical content used to solve the problem, and the mathematical competencies used to solve the problem (OECD, 2009). The situation in a mathematical problem has four levels: personal, educational/occupational, public, and scientific. The mathematics content dimension has four content areas used to solve a problem: space and shape, change and relationships, quantity, and uncertainty.

The PISA mathematics assessment measures three clusters of mathematical competencies: the knowledge and skills to recall facts and perform routine procedures, to solve nonroutine problems, and to develop and implement problem-solving strategies in complicated and unfamiliar settings. Half of the test items should be in the connection cluster, and one-quarter of the items should be in each of the other two clusters.

PISA Reading Framework

The PISA reading framework also describes test items along three dimensions: situations, text, and reading processes (OECD, 2009). Situations represent the use for which the text is constructed. PISA categorizes the situations of an item into four groups: personal, public,

occupational, and educational. The PISA reading assessment uses six types of texts to measure reading skills: description, narrative, argumentation, instruction, and transaction. The test also measures three major aspects of reading processes: accessing and retrieving, integrating and interpreting, and reflecting and evaluating.

Analyzed PISA Released Items

In this study, we analyzed 50 units of mathematics test items and 30 units of reading test items released by the time this study was conducted. Although these items are only a portion of those used to compile the PISA test forms, they are representative of the PISA item pool (Kelly, 2012).

Trends in International Mathematics and Science Study

TIMSS conducts an international assessment of the mathematics and science achievement of students in grades 4 and 8. It was administered first in 1995 and every four years thereafter. More than 50 countries and education systems have participated in TIMSS. We focused on the TIMSS mathematics test in this study.

The TIMSS assessment framework defines two dimensions for the TIMSS mathematics test (Mullis, Martin, Ruddock, et al., 2009). The content dimension specifies three domains (i.e., number, geometric shapes and measures, and data display) for grade 4 and four domains (i.e., number, algebra, geometry, and data and chance) for grade 8. The cognitive dimension describes three cognitive processes to be assessed: knowing, applying, and reasoning.[4]

TIMSS test items, including both MC and OE items, are grouped into item blocks, with about ten to 14 items in each block for grade 4 and 12–18 items per block for grade 8. Each MC item is worth one score point. An OE item is worth one or two score points. The distribution of items across the content and cognitive domains within each block matches the corresponding distribution across the full item pool. Blocks of items are combined to form test booklets. Students at all performance levels in sampled schools take the TIMSS tests. Each booklet contains two blocks of mathematics items and two blocks of science items. During the test, each student answers questions contained in one test booklet.

In this study, we analyzed 161 TIMSS mathematics items released after the 2011 administration: 73 items for grade 4 and 88 items for grade 8. Although these released items are only portion of those used to compile the TIMSS test, they are representative of the entire item pool of items (Martin and Mullis, 2012).

[4] See Appendix B for the content and cognitive domains assessed at each grade level and the expected percentage of test items in each cognitive dimension in the 2011 administration.

Prior Research Comparing Benchmark Tests

Table 2.1 summarizes a few key features of the six benchmark tests included in this study, such as the purpose, scope, and tested subjects. Prior research offers more detailed comparisons of these tests. In the following section, we review prior studies on the similarities and differences among the six tests included in this study.

Comparing NAEP, PISA, and TIMSS on Mathematics

The U.S. Department of Education (Nohara, 2001) conducted a study to compare the NAEP, PISA, and TIMSS mathematics and science tests. It convened expert panels in mathematics and science and asked participants to examine the content, context, response type, requirements for multistep reasoning, and other characteristics of items from three tests.

This study reported that the three mathematics tests differed in most areas examined, and differences in the purpose of each assessment contributed substantially to other differences among the tests. In terms of content coverage, NAEP and TIMSS had more items on number sense, properties, and operations, while PISA had more data analysis items. Regarding context, the majority of PISA items (97 percent) presented students with real-life situations or problem-solving scenarios, compared with less than 50 percent of the items on the NAEP and TIMSS tests.

These three tests also differed in the proportion of item types. MC items were the most common on the NAEP (60 percent) and TIMSS (77 percent) tests, while 50 percent of PISA items were short OE items. The proportion of extended OE items was low on all three assessments, at 10 percent on the NAEP, 3 percent on the TIMSS, and 12 percent on the PISA tests.

In terms of the proportion of items requiring multistep reasoning, 41 percent of NAEP items and 44 percent of PISA items required this skill, compared with 31 percent of TIMSS items. Moreover, more than 90 percent of PISA items required the interpretation of figures or graphical data, compared with 56 percent of NAEP and 45 percent of TIMSS items. In addition, most NAEP and TIMSS items tended to focus on a single, identifiable piece of knowledge, skill, or concept.

Table 2.1. Comparisons of the Six Benchmark Tests on Key Characteristics

Test	Purpose	Scope	Tested Subjects Analyzed[1]	Tested Grades/Age	Tested Student Population	Item Type	Partial or Complete Test Forms Analyzed
AP	End-of-course exams for AP students	Worldwide	Mathematics and ELA	Grades 10–12	Advanced students	MC and OE	Complete
IB	End-of-course exams for IB students	Worldwide	Mathematics and ELA	Grades 10–12	Advanced students	OE	Complete
NAEP	Assessing national educational achievement	Nationwide	Mathematics and ELA	Grades 4, 8, and 12	Students at all performance levels	MC and OE	Partial
PIRLS	International assessment and comparison	Worldwide	ELA	Grade 4	Students at all performance levels	MC and OE	Partial[2]
PISA	International assessment and comparison	Worldwide	Mathematics and ELA	15 years old	Students at all performance levels	MC and OE	Partial[2]
TIMSS	International assessment and comparison	Worldwide	Mathematics	Grades 4 and 8	Students at all performance levels	MC and OE	Partial[2]

1. The table cites only the subjects included in this study, where applicable.
2. Although items available for analysis for these tests came from partial test forms, the test developers confirmed that these released items were representative of the entire item pool.

Only a small portion of items required a combination of topic areas or focused more on students' thinking abilities than on concept topics. Citing differences among the tests on multiple dimensions, the study concluded that PISA was the most difficult of the three assessments.

Comparing NAEP and PIRLS Fourth-Grade Reading Tests

Several studies have examined the similarities and differences between the NAEP reading test and the PIRLS test (Binkley and Kelly, 2003; Stephens and Coleman, 2007). Thompson and colleagues (2012) compared the NAEP and PIRLS reading tests administered in 2011. They found that although NAEP and PIRLS were similar in terms of the types of texts featured, test items used, and reading processes assessed, they differed in several respects. In particular, the average length of PIRLS reading passages was shorter than on the NAEP test. The readability of PIRLS reading passages was about one grade level lower than the NAEP passages. In terms of reading processes assessed, PIRLS focused more on locating and recalling text-based information, while NAEP placed greater emphasis on integrating and interpreting, and critiquing and evaluating. Based on these findings, the authors concluded that PIRLS 2011 was cognitively less challenging than the NAEP 2011 reading test for fourth graders.

Advanced Placement and International Baccalaureate Tests

Research on the cognitive demand of AP and IB exams is limited compared with that on other tests included in our study. One report from the NRC on advanced studies of mathematics and science in U.S. high schools compared AP and IB exams (NRC, 2002). The authors reviewed the development and scoring process of two exams but did not examine the cognitive demand of test items. They highlighted differences between the two programs, such as the main goal of the program. AP exams assess what students should know and be able to do in a typical college-level course in the United States, while IB exams measure students' readiness for postsecondary education in many countries. They also noted that these differences in program goals might have contributed to observed differences between two exams.

Summary

Prior research has compared the cognitive demand of NAEP, PIRLS, PISA, and TIMSS exams but not that of AP and IB exams. This research found that the PISA mathematics test was more cognitively demanding than the NAEP and TIMSS mathematics tests, and the NAEP reading test was more cognitively demanding than PIRLS for fourth graders. These studies also acknowledged that these tests differ in many ways, such as in terms of the goals of the test or program and the target student population, and these differences should be taken into account when making comparisons.

3. Cognitive Demand Frameworks and Ratings for the Benchmark Tests

In this chapter, we describe the types of deeper learning skills that the six benchmark tests allowed us to analyze, how we chose the frameworks for our analysis of the extent to which these types of deeper learning skills were assessed by the benchmark tests, and the rating process.

Aspects of Deeper Learning Skills Assessed in This Study

The Deeper Learning Initiative focuses on enabling students to emerge from their schooling with the ability to master core academic content knowledge, think critically and solve complex problems, work collaboratively, communicate effectively, and learn how to learn. Ideally, a deeper learning assessment would assess elements from each of the cognitive, intrapersonal, and interpersonal domains.

However, similar to the state tests analyzed in our earlier study (Yuan and Le, 2012), the tests included in this study were limited in the types of deeper learning skills they measured. Only skills in the cognitive domain were measured, and skills measuring intrapersonal and interpersonal skills were absent or not explicitly assessed. For example, because the tests included in this study were intended as measures of individual students' competencies, none of the tests assessed the ability to work collaboratively. In a similar vein, although students may have engaged in self-reflective learning processes or other "learning how to learn" skills while answering the test items, none of the selected tests explicitly set out to measure learning how to learn. The tests included in this study also measured effective communication in terms of written skills but not in terms of oral skills, so they could measure only limited aspects of effective communication.

Finally, we did not examine the extent to which the tests assessed the mastery of core academic content. As defined by the Foundation, mastery of core academic content entails learning a set of facts and theories within a specific subject area or domain. Such a conception stems from learning theories suggesting that mastery of skills in one subject does not necessarily mean that students will be able to transfer or apply those same skills to another subject (NRC, 2011). Thus, to assess the mastery of core academic content, we would need to define the specific skills, concepts, and procedures that would be considered foundational knowledge for each subject at each grade level. Although the CCSS defines foundational knowledge and concepts for ELA and mathematics, it does not delineate the skills and knowledge by grade level at the upper grades, so

it is difficult to apply the CCSS to tests geared toward high school students, who constitute the target population for the majority of the tests in our sample. Because it was beyond the scope of this study to define the specific facts and theories in each subject area that would be considered foundational knowledge at each high school grade level, we did not assess the mastery of core academic content, though it should be noted that all tests included in this study included items that assessed core content within a domain.

Frameworks Used to Assess the Cognitive Demand of Benchmark Tests

We examined existing frameworks for measuring cognitive demand, all of which adopted a unidimensional formulation such that the conceptual content within each subject was generically defined (i.e., there were no attempts to define core concepts that would be considered foundational for each subject). It should also be noted that, as a simplifying assumption, we chose frameworks that treated cognitive demand as an inherent characteristic of the test item. However, it is important to recognize that the cognitive demand of an item as experienced by the examinee is a function of the interface between the personal aptitude of the examinee (such as motivational and affective characteristics), the testing environment, and the skills and knowledge being elicited by the test item (Kyllonen and Lajoie, 2003). For example, examinees who have more interest in the content on which a reading passage or writing prompt is based, or who have previously studied the content of the item, can be expected to draw more quickly upon their schema or conceptual system that interrelates their knowledge about the topic. Their familiarity or interest in the topic may free up their working memory capacity, allowing them to more efficiently organize their thoughts and respond more insightfully to the task at hand. Under these circumstances, the item is likely to be of lower cognitive demand for these examinees than would be the case for other similarly proficient examinees who have less interest in the topic or who are encountering the topic for the first time. Because we could not take into account the interactions among an examinee's attributes, the features of the test item, and the performance setting, our conception of cognitive demand assumes that it is fixed across examinees and situations.

In total, we considered five frameworks for educational objectives, cognitive processes, and learning standards: Norman Webb's (2002b) four-level DOK framework; Andrew Porter's (2002) five-level cognitive demand framework; Karin Hess et al.'s (2009) matrix that combines Webb's DOK framework and Bloom's Taxonomy of Educational Objectives; Newmann, Lopez, and Bryk's (1998) set of standards to evaluate the cognitive demand of classroom assignments and student work; and Lindsay Matsumura and colleagues' (2006) instructional quality assessment toolkit to measure the quality of instruction and the cognitive demand of student assignments.

We chose Webb's DOK framework because it is the most widely used framework to assess the cognitive demand of test items (Rothman, 2003; Webb, 1999, 2002a, 2002b, 2007) and because it provides separate descriptions for reading and writing, which made it easier for us to apply the framework to the six selected tests that administer separate measures for reading or writing (see Yuan and Le, 2012, for a more detailed discussion of each of the alternative frameworks). Notably, Smarter Balanced also uses the DOK framework to guide the development of its assessments (Measured Progress and ETS Collaborative, 2012). However, PARCC developed its own framework, in which the cognitive demand of an item depends on multiple sources, such as the complexity of the subject matter and the level of scaffolding afforded by the response option. We used both the DOK and PARCC frameworks to evaluate the cognitive demand of a given item. Each of these two frameworks is described in more detail below.

Webb's DOK Framework

Webb's DOK framework defines four levels of cognitive demand, where level 1 represents recall, level 2 represents the demonstration of a skill or concept, level 3 represents strategic thinking, and level 4 represents extended thinking (Webb, 2002b). Subject-specific descriptions for each of the DOK levels are as follows:

- Mathematics
    - DOK1: Recall of a fact, term, concept, or procedure.
    - DOK2: Use information, conceptual knowledge, and procedures in two or more steps.
    - DOK3: Requires reasoning and developing a plan or sequence of steps; has some complexity and more than one possible answer.
    - DOK4: Requires an investigation, time to think and process multiple conditions of the problem, and nonroutine manipulations.
- Reading
    - DOK1: Receive or recite facts or demonstrate basic comprehension.
    - DOK2: Engagement of some mental processing beyond recalling or reproducing a response, such as for predicting a logical outcome based on information in a reading selection or identifying the major events in a narrative.
    - DOK3: Requires abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts.
    - DOK4: Requires an extended activity in which students perform complex analyses of the connections among texts. Students may be asked to develop hypotheses or find themes across different texts.

- Writing
    - DOK1: Write simple facts, use punctuation marks correctly, identify standard English grammatical structures.
    - DOK2: Engagement of some mental processing, such as constructing compound sentences, using simple organizational strategies, or writing summaries.
    - DOK3: Requires higher-level processing, including supporting ideas with details and examples, using an appropriate voice for the intended audience, and producing a logical progression of ideas.
    - DOK4: Requires an extended activity in which students produce multiparagraph compositions that demonstrate synthesis and analysis of complex ideas (adapted from Webb, 2002b).

We applied Webb's subject-specific descriptions for each of the DOK levels to the mathematics, reading, and writing items in our analysis. Appendix C provides a sample of items rated at each DOK level in each subject.

PARCC's Cognitive Complexity Framework

PARCC uses two separate frameworks for mathematics and ELA/literacy. Cognitive demand is defined in terms of sources of cognitive complexity. For mathematics, the five sources of cognitive complexity are described as follows:

- *Mathematical content:* At each grade level, there is a range in the level of demand in the content standards. PARCC categorizes the least challenging content as low complexity and the most challenging as high complexity, with the remainder categorized as moderate complexity. Categorizations are determined based on typical expectations for mathematical knowledge at the grade level. New mathematical concepts and skills that require large shifts from previously learned concepts and skills are expected to be more complex than those that require small shifts.
- *Mathematical practices:* This source of complexity reflects the level of mathematical cognitive demand in items and tasks and the level of processing of the mathematical content. It involves what students are asked to do with mathematical content, such as apply and analyze the content.
- *Stimulus material:* This dimension accounts for the number of pieces of stimulus material in an item, as well as the role of technology tools.
- *Response mode:* The way in which examinees are required to complete assessment activities influences an item's cognitive complexity. In general, selecting a response from among given choices often is less cognitively complex than generating an original response. This

difference is due, in part, to the response scaffolding (i.e., response choices) in selected-response items that is absent from constructed-response items.

- *Processing demands:* Reading load and linguistic demands in item stems, instructions for responding to an item, and response options contribute to the cognitive complexity of items (adapted from PARCC, 2012a).

For ELA, there were four sources of cognitive complexity:

- *Text complexity:* A text will be assigned to one of three categories of test complexity: readily accessible, moderately complex, or very complex.
- *Command of textual evidence:* This source of cognitive complexity is defined as the amount of text that an examinee must process (i.e., select and understand) to respond correctly to an assessment item. The amount of text to be processed is not a reference to the length of a text or the volume of reading that is required. Instead, this category focuses on the number of details in one or more texts that must be processed.
- *Response mode:* The way in which examinees are required to complete assessment activities influences an item's cognitive complexity. In general, selecting a response from among given choices is less cognitively complex than generating an original response. This difference is due, in part, to the response scaffolding (i.e., response choices) in selected-response items that is absent from constructed-response items.
- *Processing demands:* Linguistic demands and reading load in item stems, instructions for responding to an item, and response options contribute to the cognitive complexity of items. Linguistic demands include vocabulary choices, phrasing, and other grammatical structures (adapted from PARCC, 2012b).

For each of the five dimensions in mathematics and four dimensions in ELA, the PARCC framework provides criteria for an item to be classified as having low, moderate, or high complexity. Using those descriptions, we identified exemplar items for each rating and each dimension. The exemplar items are presented in Appendix D and were used to guide the coding of each item for the purposes of this study.

Although the PARCC framework identified only four sources of cognitive complexity for ELA tests, we modified the framework so that it also included the stimulus material dimension. We made this decision based on the fact that some of the PISA reading items were administered via computer and required students to use technology to respond to nontext items.

Applying the DOK and PARCC Frameworks to the Six Benchmark Tests

To promote consistency in the ratings of the cognitive demand of the items, we needed to ensure that there was a shared understanding of the frameworks. We facilitated this process by holding

multiple meetings with the research team at the National Center for Research on Evaluation, Standards, and Student Testing, which had analyzed the cognitive demand of the content specifications of the CCSS assessments. We also met with the developers of the PARCC framework and with our external panelists. During the meetings, we discussed details of the frameworks, including our understanding of the descriptions of deeper learning at each DOK and modified PARCC level. We also reviewed various scenarios that were encountered in our prior study on state achievement tests and discussed potential ratings for each of these scenarios. After these meetings, we used the two frameworks to rate the cognitive demand of the items.

To ensure that we coded the items in a consistent manner, we conducted two rounds of calibration ratings. The calibration process involved coding randomly selected MC and OE items from each test, with a total of 70 mathematics and 80 ELA items selected for calibration. In the first round of calibration, we rated items independently and reconvened to discuss our ratings and resolve any discrepancies. In the second round, we rated the remaining calibration items independently to check the interrater reliability. Results showed good interrater reliability for both subjects under both frameworks. For DOK results, the weighted kappa coefficient, which is a measure of rater agreement that takes into account agreement due to chance, was 1 for ELA and 0.91 for mathematics. The average weighted kappa coefficient across dimensions was 0.95 for ELA and 0.89 for mathematics for the PARCC results. Given that the interrater reliability was high for both subjects and frameworks, we then independently analyzed the released items for each subject, with one rater coding mathematics and the other rater coding ELA.

Modifying the PARCC Scoring System to Better Align with the Deeper Learning Initiative

Although the PARCC framework provides guidelines for combining the various dimensions to create an overall complexity score, we deviated from the recommended scoring mechanisms to better capture the skills emphasized by the Deeper Learning Initiative. The PARCC guidelines proposed giving 30-percent weight to mathematical content, 40-percent weight to mathematical practices, and 30-percent weight to a processing complexity index, which equally weighted stimulus material, response mode, and processing demands. (Mathematically, this means that the PARCC guidelines give 10-percent weight to each of the index categories.) However, in coding the mathematics items, we encountered situations in which the mathematical content was complex (e.g., imaginary numbers) but the mathematical procedure used to solve the item was relatively straightforward. Similarly, some items assessed difficult mathematical content, but because the item was presented in MC format, it was possible to leverage the response options to identify the correct response with minimal knowledge of the content being tested. Thus, we revised the scoring mechanism to give 25-percent weight to mathematical content, 40-percent weight to mathematical practices, 5-percent weight to stimulus material, 25-percent weight to

response mode, and 5-percent weight to processing demands. The correlation between the ratings under the PARCC scoring system and the ratings obtained under our modified PARCC scoring system was very high, at 0.93.

Similarly, we deviated from PARCC's recommended scoring weights for ELA, partly because of the addition of the stimulus material dimension. The PARCC framework suggests an overall ELA composite that gives 50-percent weight to the text complexity dimension and 50-percent weight to a processing complexity index, defined as a weighted composite in which command of textual evidence is given 45-percent emphasis, response mode is given 45-percent emphasis, and processing demand is given 10-percent emphasis. (Mathematically, this means that the PARCC guidelines give 22.5-percent weight to command of textual evidence, 22.5-percent weight to response mode, and 5-percent weight to processing demands.) However, we found this weighting mechanism to be less applicable for OE writing items, for which students may be asked to respond to a provocative writing prompt that elicits an in-depth analysis of sophisticated and multifaceted ideas, yet the writing prompt itself is readily accessible in terms of its text complexity. In such cases, the item may receive a lower score than warranted according to its emphasis on assessing deeper learning. Thus, we revised the scoring mechanisms for ELA such that we gave 25-percent weight to text complexity, 40-percent weight to command of textual evidence, 25-percent weight to response mode, 5-percent weight to processing demands, and 5-percent weight to stimulus material. Similar to the mathematics results, the correlation between the ratings under the proposed PARCC scoring system and the ratings under our modified scoring system was very high, at 0.91 for ELA.

Correspondence Between the DOK and the Modified PARCC Frameworks

A key challenge in using the DOK and modified PARCC frameworks was determining whether the two frameworks classified an item's cognitive complexity in the same manner. While the DOK ratings provided a straightforward classification of deeper learning (i.e., DOK ratings of 3 or higher were indicative of deeper learning), we did not have similar guidelines for the PARCC ratings. To increase the comparability of the two frameworks, we set out to create a four-category rating system with the PARCC ratings. This entailed examining the distribution of the ratings, setting a series of preliminary cut scores, and making holistic judgments about the cognitive demand of the items in the categories delineated by the cut scores. We used an iterative process in which we made adjustments to the cut scores until the majority of the items in a given category represented roughly the same level of cognitive demand. We interpreted a rating of 1 to represent a very low level of cognitive demand, 2 to represent a low to medium level of cognitive demand, 3 to represent a medium to high level of cognitive demand, and 4 to represent a very high level of cognitive demand.

We then converted the PARCC ratings to a four-category scale. For mathematics, we assigned weighted PARCC scores between 1 and 1.15 a "1" rating, scores between 1.15 and 1.95 were assigned a "2" rating, scores between 1.95 and 2.8 were assigned a "3" rating, and scores greater than 2.8 were assigned a "4" rating. For ELA, weighted scores between 1 and 1.3 were assigned a "1" rating, scores between 1.25 and 1.65 were assigned a "2" rating, scores between 1.65 and 2.0 were assigned a "3" rating, and scores greater than 2.0 were assigned a "4" rating. (See Appendix E for the distribution of the cut scores.) Similar to the criterion applied to DOK levels as recommended by a panel of curriculum, teaching, and assessment specialists (Darling-Hammond et al., 2013), we identified deeper learning items as those that had PARCC ratings of 3 or higher.

In examining the correspondence in four-category ratings between the two frameworks, we observed a weighted kappa of 0.56 for ELA and 0.59 for mathematics. We then collapsed the four-category ratings into two-category ratings, corresponding to whether or not the items were considered indicative of deeper learning. Items with an initial four-category rating of 3 or higher were considered indicative of deeper learning. Conversely, items with an initial four-category rating of 2 or lower were considered not indicative of deeper learning. We observed a fair degree of correspondence between the classification of items as being indicative of deeper learning (or not), with a kappa value of 0.74 for ELA and 0.67 for mathematics.

We further examined our frameworks to see whether one framework gave systematically higher ratings to items than the other, and we did not observe any systematic differences. In 5 percent of the ELA ratings, the DOK framework classified the item as indicative of deeper learning (i.e., the DOK rating was at least 3), whereas the PARCC framework did not (i.e., the PARCC rating was less than 3). In 6 percent of the ELA ratings, the reverse was true (i.e., the PARCC ratings were at least 3 and the DOK ratings were less than 3). In 3 percent of the mathematics ratings, the DOK framework classified the item as being indicative of deeper learning, whereas the PARCC failed to do so, and in 5 percent of the mathematics ratings, the PARCC framework indicated that the item was indicative of deeper learning, whereas the DOK framework failed to do so.

We examined whether there were particular item features that could account for the discrepancies between the two frameworks in the classification of an item as indicative of deeper learning. In examining the content of the discrepant test items, it appeared that the DOK framework placed relatively greater emphasis on the types of cognitive processes elicited, whereas the PARCC framework placed relatively greater emphasis on the difficulty of the content being tested. For example, in ELA, an item that required examinees to summarize the abstract theme of a moderately accessible reading passage was classified as a 3 under the DOK framework but a 2 under the PARCC framework. In mathematics, an item that required students

to find the mean of a Poisson distribution was rated as a 3 under the PARCC framework but a 2 under the DOK framework. In this case, the discrepancy arose because the item assessed an advanced mathematical topic, but it required only fundamental knowledge about Poisson distributions. These discrepancies notwithstanding, for the majority of the items, the PARCC and DOK frameworks classified a given item as deeper learning (or not) in the same manner.

4. Findings

In this chapter, we present the results of our analysis. The findings represent the PARCC ratings under the modified weighting system, but the results and interpretations remained robust for the ratings obtained using the weighting system proposed by the PARCC developers.

Overview of Analyzed Test Items

In total, we examined 790 mathematics items and 436 ELA items (418 reading and 18 writing items). Of the mathematics items analyzed, 45 percent were MC items, and 55 percent were OE items. The proportion of MC mathematics items varied by test, with 67 percent for AP and NAEP, 47 percent for TIMSS, 23 percent for PISA, and none for IB. Among all reading test items analyzed, 68 percent were MC items, with 100 percent for AP, 60 percent for NAEP, and 55 percent for PIRLS and PISA. All writing items were OE items. All IB ELA items were writing items. (See Tables 4.1 and 4.2.) Only AP, IB, and NAEP had released writing items for analysis.

Table 4.1. Number of Released Mathematics Test Items Analyzed, by Test, Form, Grade, and Year

Test	Subject	Grade/Age	Year	Number of MC Items	Number of OE Items	Total Items
AP	Mathematics—calculus AB	9–12	2008	45	22	67
	Mathematics—calculus BC	9–12	2008	45	22	67
	Mathematics—statistics	9–12	2007	40	19	59
IB	Mathematics—high level	10–12	2011	0	104	104
	Mathematics—standard level	10–12	2011	0	53	53
NAEP	Mathematics	4	2009	19	12	31
	Mathematics	4	2011	36	15	51
	Mathematics	8	2009	22	12	34
	Mathematics	8	2011	35	12	47
	Mathematics	12	2009	17	10	27
TIMSS	Mathematics	4	2011	34	39	73
	Mathematics	8	2011	47	71	88
PISA	Mathematics	15 years old	2009	21	68	89

Table 4.2. Number of Released ELA Test Items Analyzed by Test, Form, Grade, and Year

Test	Subject	Grade/Age	Year	MC Items	OE Items	Total Items
AP	Language and composition	9–12	2009	52	3*	55
	Literature and composition	9–12	2009	55	3*	58
IB	English—high level	10–12	2010	0	3*	3
	English—standard level	10–12	2010	0	3*	3
NAEP	Reading	4	2009	15	9	24
	Reading	4	2011	23	13	36
	Reading	8	2009	16	9	25
	Reading	8	2011	22	14	36
	Reading	12	2009	16	10	26
	Writing	8	2011	0	3*	3
	Writing	12	2011	0	3*	3
PIRLS	Reading	4	2011	30	24	54
PISA	Reading	15 years old	2009	60	50	110

NOTE: * indicates writing items analyzed in this study.

Cognitive Demand of Analyzed Test Items

As described in Chapter Three, to assist comparisons between DOK and PARCC results, we calculated a composite PARCC score and categorized it into four levels. In the following sections, we present the percentage of items at each DOK level and each rescaled PARCC level by subject, item format, and test. Appendix F shows the results by PARCC cognitive dimension.

The Six Benchmark Tests Had Greater Cognitive Demand Than the State Tests

On average, the six benchmark tests had greater cognitive demand than that of the state achievement tests in both mathematics and ELA (Yuan and Le, 2012). The average share of items rated at or above DOK level 3 was about 15 percent for mathematics and 40 percent for ELA on the six national and international tests included in this study (see Figure 4.1), compared with 2 percent for mathematics and 20 percent for ELA on the 17 state achievement tests included in the previous study (Yuan and Le, 2012).

The Cognitive Demand of Test Items Varied by Subject and Item Format

The overall composition pattern of the cognitive demand in the six benchmark tests is similar to what was observed in the state achievement tests (Yuan and Le, 2012). The cognitive demand of the ELA tests is greater than that of the mathematics tests. About 40 percent of the reading items and all writing items were rated at or above a DOK or PARCC level of 3, compared with about 15 percent for mathematics items (see Figures 4.1 and 4.2).

Figure 4.1. Percentage of Items Rated at Each DOK Level, by Subject

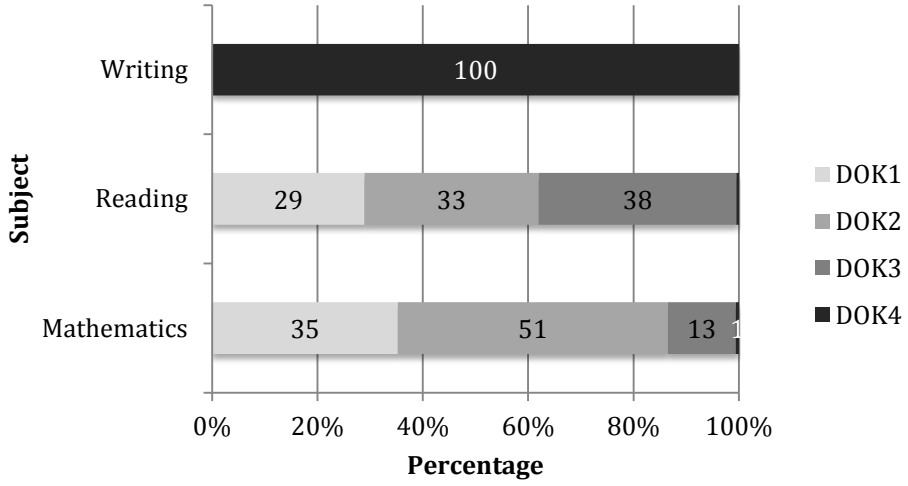
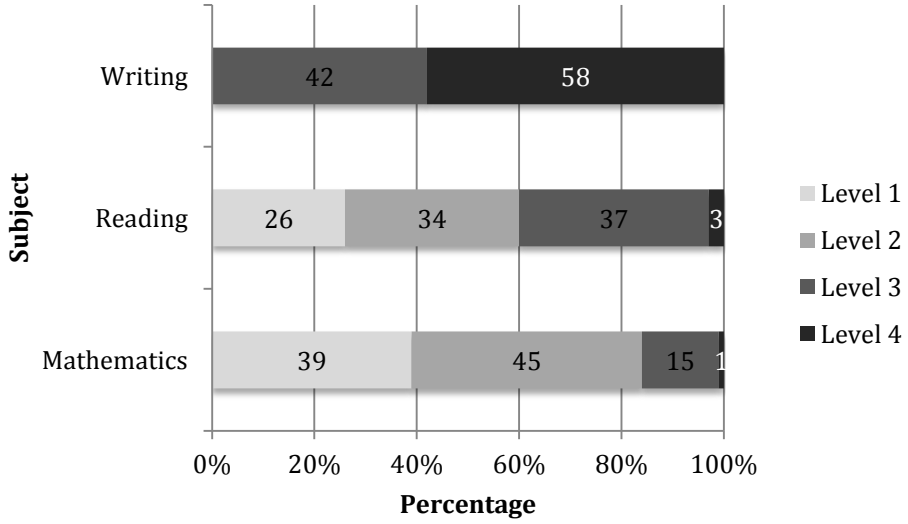


Figure 4.2. Percentage of Items Rated at Each Rescaled PARCC Level, by Subject



Results also showed that item format was associated with DOK levels. Greater proportions of OE items were rated at higher DOK or PARCC levels (3 and 4) compared with MC items for both mathematics and ELA (see Figures 4.3 and 4.4). Moreover, both types of reading items were rated as having greater cognitive demand than were corresponding types of mathematics items. None of mathematics MC items rated at or above DOK or PARCC level 3, while one-third of reading MC items rated at or above DOK or PARCC level 3. Although the cognitive demand levels of mathematics and reading OE items were distributed across the four DOK or PARCC levels, only one-quarter of mathematics OE items rated at or above DOK or PARCC level 3, compared with more than 50 percent of reading OE items.

Figure 4.3. Percentage of Items Rated at Each DOK Level, by Subject and Item Format

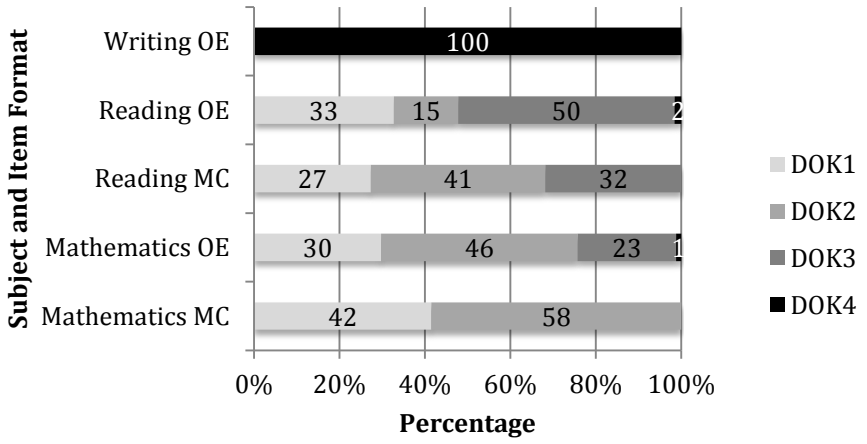
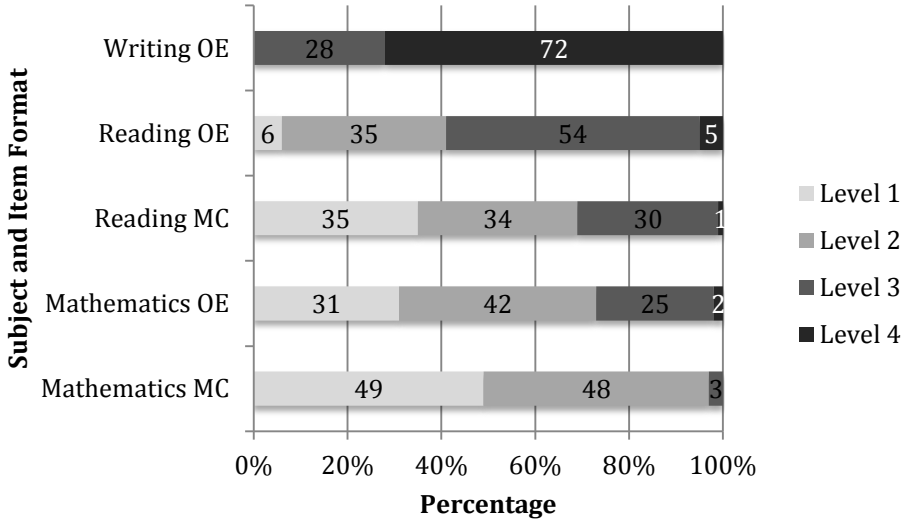


Figure 4.4. Percentage of Items Rated at Each Rescaled PARCC Level, by Subject and Item Format



The percentage of items rated at each level was similar between two frameworks for mathematics and reading. While all writing items were rated at higher levels (i.e., levels 3 and 4) under both frameworks, 100 percent of writing items rated DOK level 4, compared with 72 percent rated under the PARCC framework. Although we revised the scoring mechanism for the PARCC ELA framework to better capture the importance of different sources of complexity, the PARCC framework may still give relatively greater weight to text complexity than the DOK framework, which might have contributed to the differences in ratings for writing items between two frameworks.

The Six Benchmark Tests Varied in Their Proportion of Cognitively Demanding Items

When we look at results across both types of items, more than 60 percent of the items on the five mathematics tests were rated at or above DOK or PARCC level 2 (see Tables 4.3 and 4.4). More than or close to two-third of AP, IB, and PISA mathematics items rated at or above DOK or PARCC level 2. The share of items rated at or above DOK level 2 was over 60 percent for NAEP and about 50 percent for TIMSS. On both tests, slightly more than 30 percent mathematics items rated at or above PARCC level 2.

Across the five mathematics tests, only about 15 percent of mathematics items were rated at or above DOK or PARCC level 3 (see Tables 4.3 and 4.4). More than or close to one-third of IB mathematics exam items rated at or above DOK or PARCC level 3. About one-seventh of AP mathematics items rated at or above DOK level 3, and about one-quarter rated at PARCC level 3. One-sixth of PISA's mathematics items rated at or above DOK or PARCC level 3. This percentage was below 10 percent for both NAEP and TIMSS.

On average, more than 70 percent of items on the five reading tests rated at or above DOK or PARCC level 2 (see Tables 4.5 and 4.6). More than two-thirds of AP reading items rated at or above level 2 on both frameworks. More than two-thirds of NAEP and PISA items rated at or above level 2 on one framework and but not on the other framework. Fewer than half of PIRLS items rated at or above level 2 on both frameworks.

Across the five reading tests, about 40 percent of the reading items rated at or above DOK or PARCC level 3 (see Tables 4.5 and 4.6). More than half of AP reading items rated at or above level 3 of both frameworks. About one-third of NAEP and PISA reading items rated at or above level 3 on both frameworks. About one-sixth of PIRLS reading items rated at or above level 3 on both frameworks.

Results were similar between the two frameworks in terms of the percentage of items rated at higher levels (3 and 4) for both subjects. The results showed some differences between two frameworks in the percentage of items rated at or above level 2 for both subjects, and these differences varied by test. For instance, on the AP mathematics and reading tests, greater percentage of items rated at or above level 2 under the PARCC framework than under the DOK framework. In contrast, NAEP mathematics and reading tests had a smaller percentage of items rated at or above level 2 under PARCC than under DOK. The PISA mathematics test had a slightly smaller percentage of items rated at or above level 2 under PARCC than under DOK, while the PISA reading test had a higher percentage of items rated at or above level 2 under PARCC than under DOK.

Table 4.3. Percentage of Mathematics Test Items Rated at Each DOK Level, by Test and Item Format

	All Items			MC Items					OE Items					Above	Above
Test	N	MC	OE	N	DOK 1	DOK 2	DOK 3	DOK 4	N	DOK 1	DOK 2	DOK 3	DOK 4	DOK 1	DOK 2
AP	193	67%	33%	130	34%	66%	0%	0%	63	24%	30%	43%	3%	69%	15%
IB	157	0%	100%	0	—	—	—	—	157	21%	50%	28%	1%	79%	29%
NAEP	190	68%	32%	129	45%	55%	0%	0%	61	25%	52%	21%	2%	61%	7%
PISA	89	24%	76%	21	19%	81%	0%	0%	68	37%	41%	22%	0%	67%	17%
TIMSS	161	50%	50%	81	55%	45%	0%	0%	80	49%	48%	3%	0%	48%	2%
Total	790	46%	54%	361	42%	59%	0%	0%	429	30%	46%	23%	1%	65%	13%

Table 4.4. Percentage of Mathematics Test Items Rated at Each Rescaled PARCC Level, by Test and Item Format

	All Items			MC Items					OE Items					Above	Above
Test	N	MC	OE	N	Level 1	Level 2	Level 3	Level 4	N	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2
AP	193	67%	33%	130	5%	87%	8%	0%	63	8%	35%	51%	6%	94%	24%
IB	157	0%	100%	0	—	—	—	—	157	18%	47%	33%	2%	82%	35%
NAEP	190	68%	32%	129	77%	23%	0%	0%	61	44%	43%	13%	0%	34%	4%
PISA	89	24%	76%	21	52%	48%	0%	0%	68	35%	44%	19%	2%	61%	16%
TIMSS	161	50%	50%	81	73%	26%	1%	0%	80	65%	34%	1%	0%	31%	1%
Total	790	46%	54%	361	49%	48%	3%	0%	429	31%	42%	25%	2%	61%	16%

Table 4.5. Percentage of Reading Test Items Rated at Each DOK Level, by Test and Item Format

Test	All Items			MC Items					OE Items					Above DOK 1	Above DOK 2
	N	MC	OE	N	DOK 1	DOK 2	DOK 3	DOK 4	N	DOK 1	DOK 2	DOK 3	DOK 4		
AP	107	100%	0%	107	11%	33%	56%	0%	0	—	—	—	—	89%	56%
NAEP	147	63%	37%	92	26%	59%	15%	0%	55	11%	22%	64%	3%	80%	34%
PISA	110	55%	45%	60	35%	38%	27%	0%	50	40%	12%	48%	0%	63%	36%
PIRLS	54	56%	44%	30	73%	20%	7%	0%	24	67%	8%	25%	0%	30%	15%
Total	418	69%	31%	289	27%	41%	32%	0%	129	33%	15%	50%	2%	71%	38%

NOTE: IB is not included in this table because IB ELA tests did not have any reading items.

Table 4.6. Percentage of Reading Test Items Rated at Each Rescaled PARCC Level, by Test and Item Format

Test	All Items			MC Items					OE Items					Above Level 1	Above Level 2
	N	MC	OE	N	Level 1	Level 2	Level 3	Level 4	N	Level 1	Level 2	Level 3	Level 4		
AP	107	100%	0%	107	0%	36%	64%	0%	0	—	—	—	—	100%	64%
NAEP	147	63%	37%	92	64%	27%	9%	0%	55	2%	25%	67%	6%	59%	33%
PISA	110	55%	45%	60	31%	47%	17%	5%	50	6%	40%	48%	6%	80%	36%
PIRLS	54	56%	44%	30	80%	20%	0%	0%	24	17%	46%	38%	0%	48%	17%
Total	418	69%	31%	289	35%	34%	30%	1%	129	6%	35%	54%	5%	74%	40%

NOTE: IB is not included in this table because IB ELA tests did not have any reading items.

Differences between two frameworks in the sources of complexity considered for each subject and the weights assigned to each complexity source might have contributed to these differences in the results.

For example, the DOK mathematics framework focuses on the mathematical practices and response mode dimensions of the PARCC mathematics framework, with the majority of the weight given to the mathematical practices dimension. In contrast, the PARCC mathematics framework gives greater weight to the mathematical content dimension and less weight to the mathematical practices dimension than does the DOK mathematics framework.

For ELA, the DOK framework focuses on the command of textual evidence and response mode dimensions of the PARCC ELA framework, with the majority of the weight given to the command of textual evidence. The PARCC ELA framework gives greater weight to text complexity and less weight to the command of textual evidence.

These differences may have interacted with the features of each test on these key sources of complexity (i.e., content, practices, and response mode for mathematics, and textual complexity, command of textual evidence, and response mode for ELA) and produced differences in the results between two frameworks that varied by test.

Only Two Benchmark Tests Met Both Criteria for High-Quality Measures of Deeper Learning

In Chapter One, we noted that Darling-Hammond et al. (2013) set two criteria for the assessment of higher-order cognitive skills in high-quality measures of deeper learning. One is that at least two-thirds of the items should be rated at DOK levels 2, 3, or 4 (referred to as Criterion I). The other is that at least one-third of the items should be rated at DOK levels 3 or 4 for a mathematics test and at least half of the items should be rated at or above DOK level 3 for ELA (Criterion II). We used these two criteria to examine whether these benchmark tests can be considered high-quality measures of deeper learning. We combined the results for reading and writing in this analysis.

IB mathematics and ELA tests met both criteria under at least one framework (see Table 4.7). AP ELA tests met both criteria according to both frameworks. AP mathematics tests met Criterion I but not Criterion II according to both frameworks. PISA mathematics and ELA tests met Criterion I under one framework. Neither PISA's mathematics nor ELA tests met Criterion II under either framework, though PISA's ELA tests were closer to the goal of Criterion II than its mathematics tests. The NAEP mathematics test did not meet any of the criteria according to either framework. The NAEP ELA test met Criterion I according to the DOK framework but not the PARCC framework, and it did not meet Criterion II under either frameworks. Neither TIMSS nor PIRLS met the two criteria for high-quality assessment of higher-order cognitive skills.

Table 4.7. Whether a Selected Test Met Two Criteria for High-Quality Measures of Higher-Order Cognitive Skills Based on Two Frameworks

Subject	Test	DOK		PARCC	
		Criterion I	Criterion II	Criterion I	Criterion II
Mathematics	AP	✔		✔	
	IB	✔		✔	✔
	NAEP				
	PISA	✔			
	TIMSS				
ELA	AP	✔	✔	✔	✔
	IB	✔	✔	✔	✔
	NAEP	✔			
	PISA			✔	
	PIRLS				

NOTE: Criterion I indicates that at least two-thirds of the test items are rated at level 2 or higher. Criterion II indicates that at least one-third of the mathematics items and half of the ELA items are rated at level 3 or higher.

Cognitive Demand Level Varied by Test Purpose and Characteristics of Target Students

The results also suggest that the percentage of cognitively demanding items on the six benchmark tests was associated with the purpose of the test and the characteristics of the targeted student population. Because the IB and AP tests assess students' readiness for postsecondary academic learning, they target academically advanced high school students. Commensurately, these measures had a greater percentage of cognitive challenging items than the other tests included in this study. If the items on these types of tests were instead focused on lower-level skills, the tests may not have a sufficient level of discriminating power to differentiate among students who should receive college-level credit and those who should not.

In comparison, PISA, NAEP, TIMSS, and PIRLS assess what students know and can do for students at all academic performance levels. They have a lower proportion of cognitively demanding items than the IB and AP tests and were more likely to include items covering a range of cognitive demand so as to accommodate the abilities of the targeted population. If these tests were to include a disproportionate number of cognitively rigorous items, the test may be too difficult for most students, and the resulting floor effects may render the results less useful. Given that the CCSS assessments will measure students at all performance levels, results from these four tests provide a better benchmark for future analysis of CCSS assessments than do results from the IB and AP tests.

5. Discussion and Implications

In this chapter, we summarize our results, discuss the implications for the Foundation's Deeper Learning Initiative, and describe the limitations of this study.

Summary of Results

The overall composition pattern of cognitive demand in the six benchmark tests by subject and item format is similar to what was observed on state achievement tests (Yuan and Le, 2012). The cognitive demand of ELA tests is greater than that of mathematics tests. Format is associated with the cognitive demand of items, with OE items being more cognitively demanding than MC items.

On average, these six benchmark tests had greater cognitive demand than did state achievement tests in mathematics and ELA (Yuan and Le, 2012). The average share of items rated at or above DOK or PARCC level 3 was about 15 percent for mathematics and 40 percent for ELA for the six benchmark tests, compared with 2 percent for mathematics and 20 percent for ELA for the 17 state achievement tests in the earlier study.

The six benchmark tests varied in their percentages of cognitively demanding items. IB and AP had greater percentages of cognitively demanding items than other benchmark tests on both subjects. TIMSS and PIRLS appeared to be less cognitively demanding than other benchmark tests. By and large, results were similar between the two frameworks in terms of the percentage of items rated at higher levels (3 and 4). There were some differences between the two frameworks in terms of the percentage of items rated at or above level 2, however. Multiple factors might have contributed to such differences, such as the sources of complexity considered, weights assigned to each source, and the features of each test in terms of key sources of complexity.

Only two benchmark tests met both criteria for high-quality measurements of higher-order cognitive skills recommended by Darling-Hammond et al. (2013). IB mathematics and IB ELA tests met both criteria according to at least one framework. AP ELA tests met both criteria according to both frameworks. AP mathematics tests met Criterion I but not Criterion II under both frameworks. PISA mathematics and ELA tests met Criterion I under one framework. Neither PISA's mathematics nor ELA tests met Criterion II under either framework. The NAEP mathematics test did not meet any of the criteria according to either framework. The NAEP ELA test met Criterion I according to the DOK framework but not the PARCC framework, and it did

not meet Criterion II under either frameworks. Neither TIMSS nor PIRLS met either of the two criteria for high-quality assessments of higher-order cognitive skills.

Implications of the Findings

This study has implications for the Foundation as it gauges progress toward the Deeper Learning Initiative's goal of increasing the emphasis placed on deeper learning.

First, although Herman and Linn's (2013) analysis of content specifications suggests that the forthcoming CCSS assessments have the potential to place greater emphasis on deeper learning than is currently observed in most large-scale assessments (see Chapter One for details of their study), our results show that it is difficult to create high-quality deeper learning assessments in practice, as evidenced by the fact that the majority of the tests examined here failed to reach Criterion II. For a number of reasons, the operational versions of the CCSS assessments may not conform to the intended cognitive demand. Herman and Linn's analysis was based on the full domain of potential assessment targets, and time constraints will preclude all targets from being assessed. To the extent that the assessment targets that are eliminated are classified as DOK level of 3 or higher, the percentage of deeper learning items included on the operational versions of the tests will be reduced.

This caveat notwithstanding, if the operational forms of the CCSS assessments mirrors the distribution of the assessment targets across the DOK levels identified by Herman and Linn (2013), then the next generation of assessments will meet both Criterion I and II for high-quality measures of deeper learning. And if the forthcoming CCSS tests meet both Criterion I and II, this will mean that their cognitive demand will be greater than the NAEP, PISA, TIMSS, and PIRLS tests and comparable to that of the AP and IB tests.

Whether the potential of the CCSS assessments will actually be realized remains unknown. As underscored by our study, high-quality deeper learning assessments are difficult to develop. Although each of the benchmark tests included in our study had well-defined frameworks to guide their development and underwent thorough pilot testing and item analysis, few met Criterion II, and the majority of test items were coded at lower levels of cognitive demand. In fact, none of the testing programs that have the same target population as the CCSS assessments and are intended to measure the academic performance of students at all achievement levels (i.e., NAEP, PISA, TIMSS, and PIRLS) met Criterion II under either framework used in our study. Thus, while the CCSS content specifications show promise, only an empirical analysis of the operational forms of the CCSS assessments will reveal whether the final versions of the exams faithfully represent the proposed test specifications.

It should be noted that an empirical analysis of the operational versions of the Smarter Balanced assessments may be complicated by the computer adaptive feature of those exams, such that the items presented change dynamically as a function of the students' performance. This can result in different percentages of deeper learning items being presented to students of varying achievement levels. Thus, whether the Smarter Balanced assessments meet Criterion I or Criterion II may depend on the specific mix of items encountered by the examinees.

A second implication of our study is that future research that evaluates whether the CCSS tests are high-quality assessments will need to consider the characteristics of the students who take these tests, as well as the tests' purpose. Many factors come into play when designing a test, and cognitive demand is not the ultimate driving factor for test design. The purpose of a test must be considered, and a test with many items that have a relatively low level of cognitive demand does not necessarily indicate poor quality. Furthermore, because one test cannot measure all the skills that are deemed important for students to demonstrate, a test that is intended to measure overall mastery of a subject will generally need to strike a balance between lower-level skills and higher-order skills. Determinations of whether the CCSS tests contain a sufficient number of cognitively demanding items will ultimately need to take these complexities into account.

Third, results from this study and our previous study on state achievement tests suggest that future evaluations of the Deeper Learning Initiative may encounter the same types of challenges, such that only a limited selection of deeper learning skills can be examined. Like most standardized achievement tests, the Smarter Balanced and PARCC assessments are likely to focus only on the cognitive component of the Deeper Learning Initiative and may not assess the intrapersonal and interpersonal competencies that are also part of the larger deeper learning construct. This suggests that the intrapersonal and interpersonal aspects of the Deeper Learning Initiative will need to be assessed through supplemental measures to the achievement tests, yet measures of intrapersonal and interpersonal skills are still in the incipient stages of development (NRC, 2012; Soland, Hamilton, and Stecher, 2013). There are few standardized measures of intrapersonal and interpersonal skills, and these measures have unknown validity and reliability. Given the current assessment landscape, the Foundation may not be able to assess the full range of deeper learning skills outlined in the Deeper Learning Initiative without making trade-offs with respect to psychometric properties, costs, and other considerations (Yuan and Le, 2012).

Finally, results from the benchmark tests and state achievement tests studies also indicate the need to develop analytic frameworks that would allow an analysis of the mastery of core conceptual content as integrated with critical thinking and problem solving in each subject area.

In this study, because we did not identify the core concepts considered foundational for each subject area, we only considered assessment frameworks that adopted a undimensional

formulation of cognitive demand with respect to critical-thinking and problem-solving skills. However, as noted in Yuan and Le (2012), there is increasing evidence to suggest that students learn best when they acquire foundational content knowledge that allows them to transfer their skills to new domains and problem types (Schneider and Stern, 2010). That is, there is interdependence between critical-thinking and problem-solving skills and fluency with the core concepts, practices, and the organizing principles that constitute a subject domain.

Although the CCSS framework provides foundational knowledge and concepts for ELA and mathematics, it does not delineate the skills and knowledge by grade level at the upper grades, so it is difficult to apply the CCSS to tests geared toward the high school population, which constituted the target for the majority of the tests in our sample. Developing an analytic framework that would allow an analysis of the mastery of core conceptual content as integrated with critical thinking and problem solving for each high school grade was beyond the scope of this study, but future studies examining the Foundation's Deeper Learning Initiative should consider using the CCSS or other frameworks that define foundational concepts and knowledge for each subject area when assessing the cognitive demand of a test item.

Study Limitations

There are several caveats worth noting when interpreting the results of this analysis.

First, as a simplifying assumption, we treated cognitive demand as a fixed characteristic of the test item. However, it is important to recognize that the cognitive demand of an item as experienced by the examinee is a function of the interface between the individual's personal attributes, the testing environment, and the skills and knowledge being elicited by the test item (Kyllonen and Lajoie, 2003).

Second, we relied on released test items to examine the cognitive demand of the six benchmark tests. The degree to which these items are representative of the entire sample pool from which they are drawn varies across tests. Although we analyzed the complete test forms of AP and IB tests in a particular year, it is unknown whether the level of cognitive demand varies by year. Only partial forms of PISA, PIRLS and TIMSS were available for analysis; however, these released test items were representative of their item pools, according to the developers of these tests. Additionally, NAEP released items may not be representative of the entire item pool with respect to the level of cognitive demand. Collectively, differences in the representativeness of released items might have introduced bias in the evaluation of the cognitive demand of these tests; however, the direction of this potential bias is unknown.

Finally, for testing programs that included both MC and OE items, we generally did not have access to the score points for each individual item, the total score points for each test, or the amount of testing time allocated to the MC or OE items. Thus, we could not evaluate cognitive demand based on the proportion of the total test scores accounted for by items deemed as indicative of deeper learning, nor could we account for the emphasis on each type of item, based on testing time. As a result, items that were worth proportionately more in terms of score points or took proportionately more time to answer were treated the same as items that were worth fewer points and took less time to answer. It is unknown whether our results would change, and to what extent, if we were able to account for these factors, so the results of our analysis should be interpreted carefully. With this caveat in mind, we interpret our results as representing an approximate indication of the level of deeper learning that can be found on the selected benchmark tests.

Appendix A. Distributions of NAEP Items, by Grade and Specific Framework Dimension

This appendix provides additional background on the distributions of NAEP test items, by grade and framework dimension.

Table A.1. Distribution of NAEP Mathematics Items, by Grade and Content Area

Content Area	Grade 4	Grade 8	Grade 12
Number properties and operations	40%	20%	10%
Measurement	20%	15%	30%
Geometry	15%	20%	
Data analysis, statistics, and probability	10%	15%	25%
Algebra	15%	30%	35%

SOURCE: National Assessment Governing Board, 2010a.

Table A.2. Distribution of Literary and Information Passages in the NAEP Reading Test

Grade	Literary	Informational
4	50%	50%
8	45%	55%
12	30%	70%

SOURCE: National Assessment Governing Board, 2010b.

Table A.3. Distribution of NAEP Writing Items, by Grade and Writing Goals

Writing Goal	Grade 4	Grade 8	Grade 12
To persuade	30%	35%	40%
To explain	35%	35%	40%
To convey experience	35%	30%	20%

SOURCE: National Assessment Governing Board, 2010c.

Appendix B. Distributions of TIMSS Items, by Content and Cognitive Domain

This appendix provides additional background on the distributions of test items, by content and cognitive domain, according to the TIMSS 2011 mathematics assessment.

Table B.1. Expected Percentage of Test Items in Each Content Domain in the TIMSS 2011 Mathematics Assessment for Grade 4

Content Domain	Percentage
Number	50%
Geometric shapes and measures	35%
Data display	15%

SOURCE: Mullis, Martin, Ruddock, et al., 2009.

Table B.2. Expected Percentage of Test Items in Each Content Domain in the TIMSS 2011 Mathematics Assessment for Grade 8

Content Domain	Percentage
Number	30%
Algebra	30%
Geometry	20%
Data and chance	20%

SOURCE: Mullis, Martin, Ruddock, et al., 2009.

Table B.3. Expected Percentage of Test Items in Each Cognitive Domain in the TIMSS 2011 Mathematics Assessment, by Grade

Cognitive Domain	Percentage	
	Grade 4	Grade 8
Knowing	40%	35%
Applying	40%	40%
Reasoning	20%	25%

SOURCE: Mullis, Martin, Ruddock, et al., 2009.

Appendix C. Exemplary Test Items at Each DOK Level

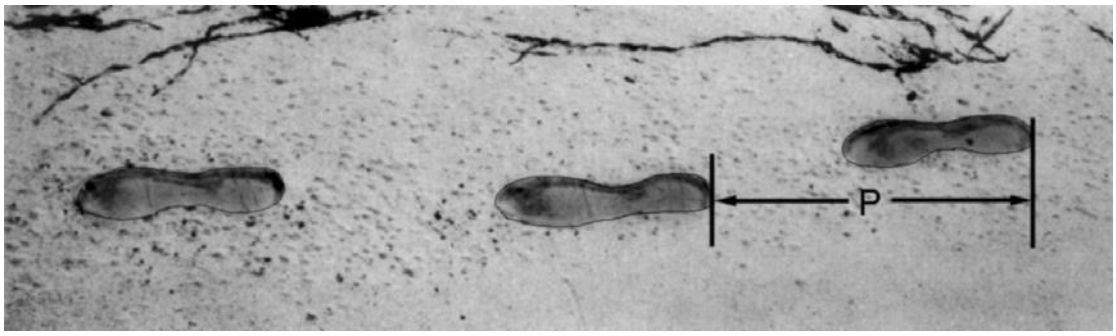
Mathematics DOK Level 1

Add: 20,000 + 790,000 =

A. 792,000
B. 810,000
C. 811,000
D. 990,000

SOURCE: NAEP 2011, Grade 4, Block M9, Item 5 (NCES, 2013).

Mathematics DOK Level 2



The picture shows the footprints of a man walking. The pacelentgh *P* is the distance between the rear of two consecutive footprints.

For mean, the formula, $\frac{n}{P} = 140$, gives an approximate relationship between *n* and *P,* where
n = number of steps per minutes, and
P = pacelength in meters

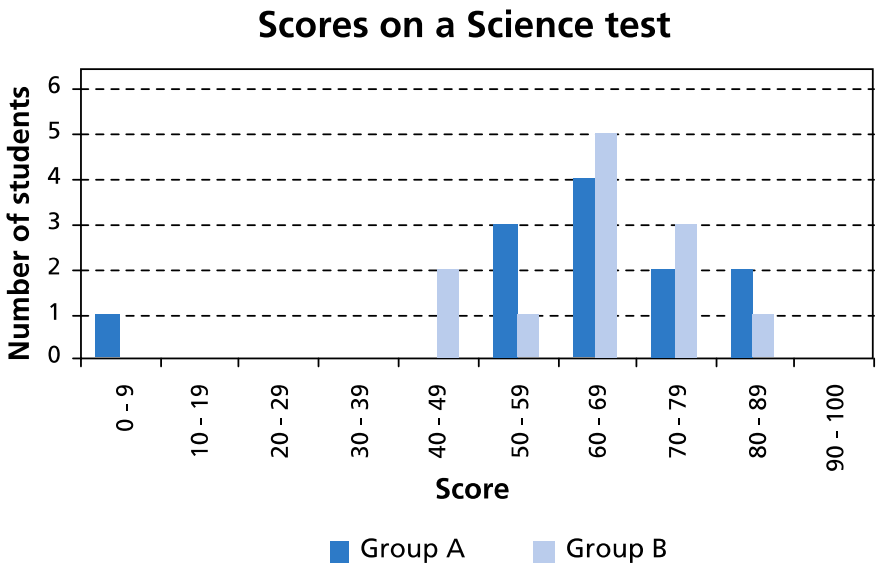
If the formula applies to Heiko's walking and Heiko takes 70 steps per minute, what is Heiko's pacelength? Show your work.

SOURCE: PISA Mathematics Item 2.1, released in 2009.

Mathematics DOK Level 3

The diagram below shows the results on a Science test for two groups, labeled as Group A and Group B.

The mean score for Group A is 62.0 and the mean for Group B is 64.5. Students pass this test when their score is 50 or above.



Looking at the diagram, the teacher claims that Group B did better than Group A in this test.

The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessarily have done better.

Give one mathematical argument, using the graph, that the students in Group A could use.

SOURCE: PISA Mathematics Item 20.1, released in 2009.

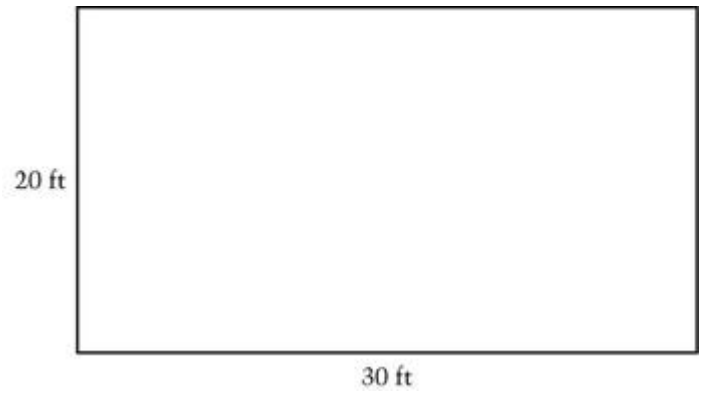
Mathematics DOK Level 4

The Morrisons are going to build a new one-story house. The floor of the house will be rectangular with a length of 30 feet and a width of 20 feet. The house will have a living room, a kitchen, two bedrooms, and a bathroom. In part (a) below create a floor plan that shows these five rooms by dividing the rectangle into rooms. Your floor plan should meet the following conditions.

- Each one of the five rooms must share at least one side with the rectangle in part (a); that is, each room must have at least one outside wall.
- The floor area of the bathroom should be 50 square feet.
- Each of the other four rooms (not the bathroom) should have a length of at least 10 feet and a width of at least 10 feet.

Be sure to label each room by name (living room, kitchen, bedroom, etc.) and include its length and width, in feet. (Do not draw any hallways on your floor plan.)

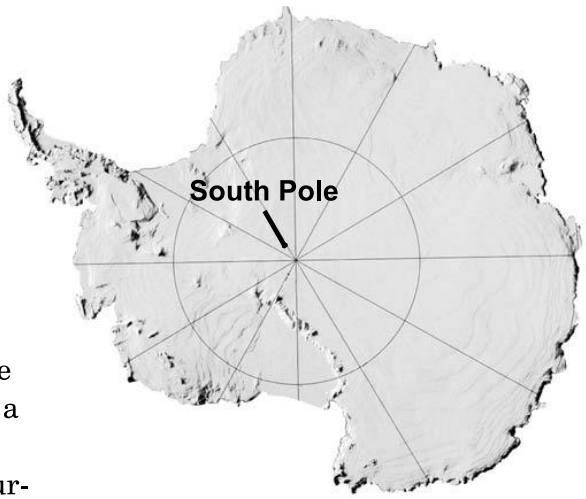
(a) Draw your floor plan on the figure below. Remember to label your rooms by name and include the length and width, in feet, for each room.



SOURCE: NAEP 2009, Grade 8, Block M10, Item 16 (NCES, 2013).

Antarctica: Land of Ice

Introducing Antarctica



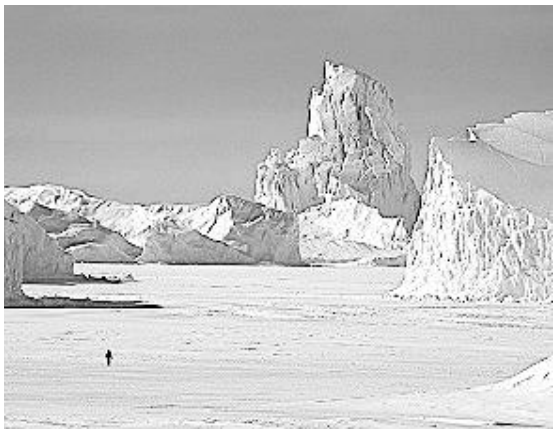
A Map of Antarctica

What is Antarctica?

Antarctica is a continent that is right at the south of the planet. (If you try to find it on a globe, you will see that it is at the bottom.)

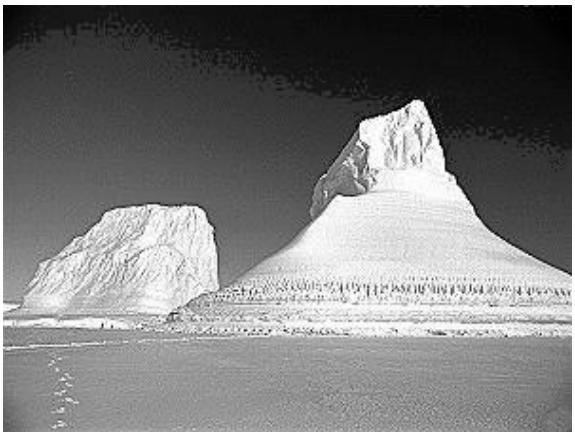
It takes up one-tenth of the Earth's surface and is covered with a blanket of ice that can be as thick as 1,500 metres or more. The South Pole is right in the middle of Antarctica.

Antarctica is the coldest continent, as well as the driest, the highest and the windiest. Very few people live there all year round. Scientists stay there for short periods, living in specially built research stations.

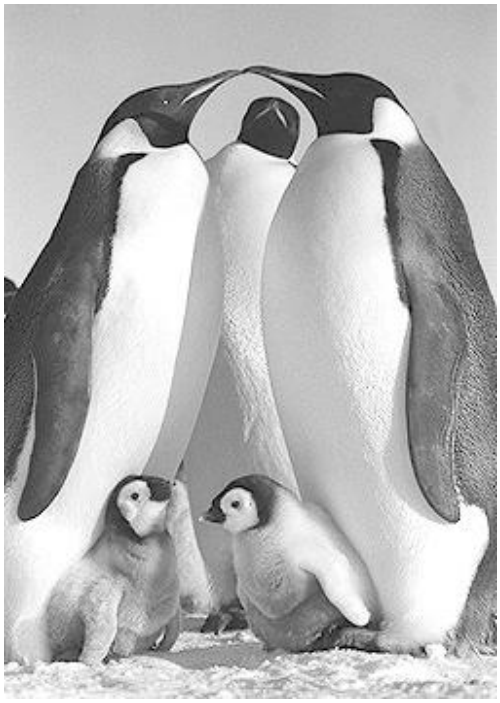


Summer in Antarctica is between October and March. During this time there is non-stop daylight. In winter, April to September, the opposite happens and Antarctica is plunged into six months of constant darkness.

In Antarctica, it is colder than you can possibly imagine, even in the summer! The South Pole is the coldest part of Antarctica. The average temperature for January, the middle of the summer, is minus 28 degrees Celsius (written as -28°C). Minus means colder than the freezing point, which is 0°C.



In the winter, April to September, the average temperature at the South Pole can be as cold as -89°C. When it is that cold, a mug of boiling water thrown in the air would freeze before it hit the ice. Sometimes the scientists have to use fridges to keep their samples warm!



Penguins in Antarctica

There are more penguins in the Antarctic than any other bird.

They cannot fly but use their short wings as swimming flippers. They are superb swimmers. On land, they waddle upright or move in short hops.

Penguins have many feathers that overlap each other. These, together with woolly down feathers and a thick layer of fat, keep out the cold air, wind and water. For extra warmth, penguins huddle together in groups.

A Letter from Antarctica

Sara Wheeler is one of the scientists working in Antarctica. By reading her letter to her nephew Daniel, you can learn more about her Antarctic experience.

Antarctica

Friday, 9 December

Dear Daniel,

Here is the letter I promised to write to you from Antarctica, and a photograph. Imagine how excited I am to be here at last, following in the footsteps of so many famous explorers. It is very different from the world I am used to.

There is nothing fresh down here—and no supermarkets—so we have to eat a lot of dried, tinned or frozen food (it doesn't have to be put in the freezer—you can just leave it outside). We cook on small gas stoves, which take much longer than cookers at home. Yesterday I made noodles with tomato paste and vegetables out of a tin, followed by dried strawberries that tasted like cardboard.

I miss fresh apples and oranges—I wish you could send me some!

Love from Sara

Reading DOK Level 1

2. Antarctica is the coldest place on Earth. What other records does it hold?

A. driest and cloudiest
B. wettest and windiest
C. windiest and driest
D. cloudiest and highest

Reading DOK Level 2

8. What are two things you learn about food in Antarctica from Sara's letter?
1_____
_____
2._____
_____

Reading DOK Level 3

9. Think about whether you would like to visit Antarctica. Use what you have read in both *Introducing Antarctica* and A *Letter from Antarctica* to explain why you would or would not like to visit.

SOURCE: PIRLS 2006, Passage: Antarctica, Items 2, 8, and 9 (Mullis, Martin, Kennedy, et al., 2007).

R081: Graffiti

I'm simmering with anger as the school wall is cleaned and repainted for the fourth time to get rid of graffiti. Creativity is admirable but people should find ways to express themselves that do not inflict extra costs upon society.

Why do you spoil the reputation of young people by painting graffiti where it's forbidden? Professional artists do not hang their paintings in the streets, do they? Instead they seek funding and gain fame through legal exhibitions.

In my opinion buildings, fences and park benches are works of art in themselves. It's really pathetic to spoil this architecture with graffiti and what's more, the method destroys the ozone layer. Really, I can't understand why these criminal artists bother as their "artistic works" are just removed from sight over and over again.

Helga

There is no accounting for taste. Society is full of communication and advertising. Company logos, shop names. Large intrusive posters on the streets. Are they acceptable? Yes, mostly. Is graffiti acceptable? Some people say yes, some no.

Who pays the price for graffiti? Who is ultimately paying the price for advertisements? Correct. The consumer.

Have the people who put up billboards asked your permission? No. Should graffiti painters do so then? Isn't it all just a question of communication – your own name, the names of gangs and large works of art in the street?

Think about the striped and chequered clothes that appeared in the stores a few years ago. And ski wear. The patterns and colours were stolen directly from the flowery concrete walls. It's quite amusing that these patterns and colours are accepted and admired but that graffiti in the same style is considered dreadful.

Times are hard for art.

Sophia

The two letters on the opposite page come from the Internet and are about graffiti. Graffiti is illegal painting and writing on walls and elsewhere. Refer to the letters to answer the questions below.

Which of the two letter writers do you agree with? Explain your answer by using your own words to refer to what is said in one or both of the letters.

SOURCE: PISA 2006 Reading Item R081Q06A (OECD, 2006).

Writing DOK Level 4

Some of your friends perform community service. For example, some tutor elementary school children and others clean up litter. They think helping the community is very important. But other friends of yours think community service takes too much time away from what they need or want to do. Your principal is deciding whether to require all students to perform community service. Write a letter to your principal in which you take a position on whether students should be required to perform community service. Support your position with examples.

SOURCE: NAEP 2011 Writing Test, Grade 8, Block W16, Item 1 (NCES, 2011).
NOTE: All writing items included in this study were rated at DOK level 4.

Appendix D. Exemplary Test Items at Each PARCC Level, by Subject and Dimension

Mathematics Content Dimension—Low Level

Which of the following numbers is twenty-three and eight-thousandths?

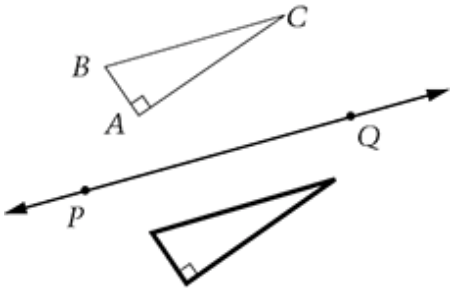
A. 230.8
B. 23.8
C. 23.08
D. 23.008
E. 23.0008

SOURCE: NAEP 2011, Grade 8, Block M8, Item 1 (NCES, 2013).

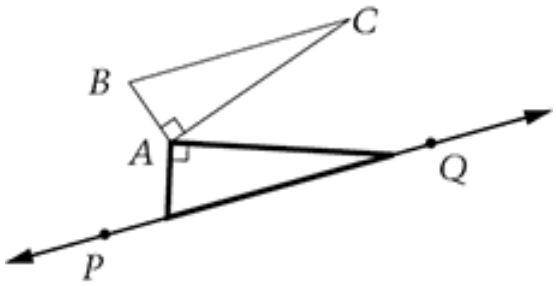
Mathematics Content Dimension—Moderate Level

Which of the following figures shows the reflection of triangle ABC over line PQ?

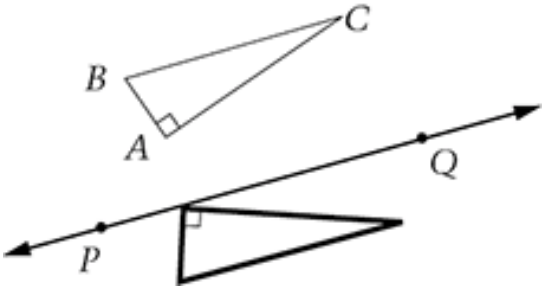
A.



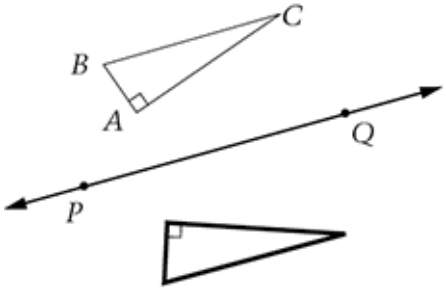
B.



C.



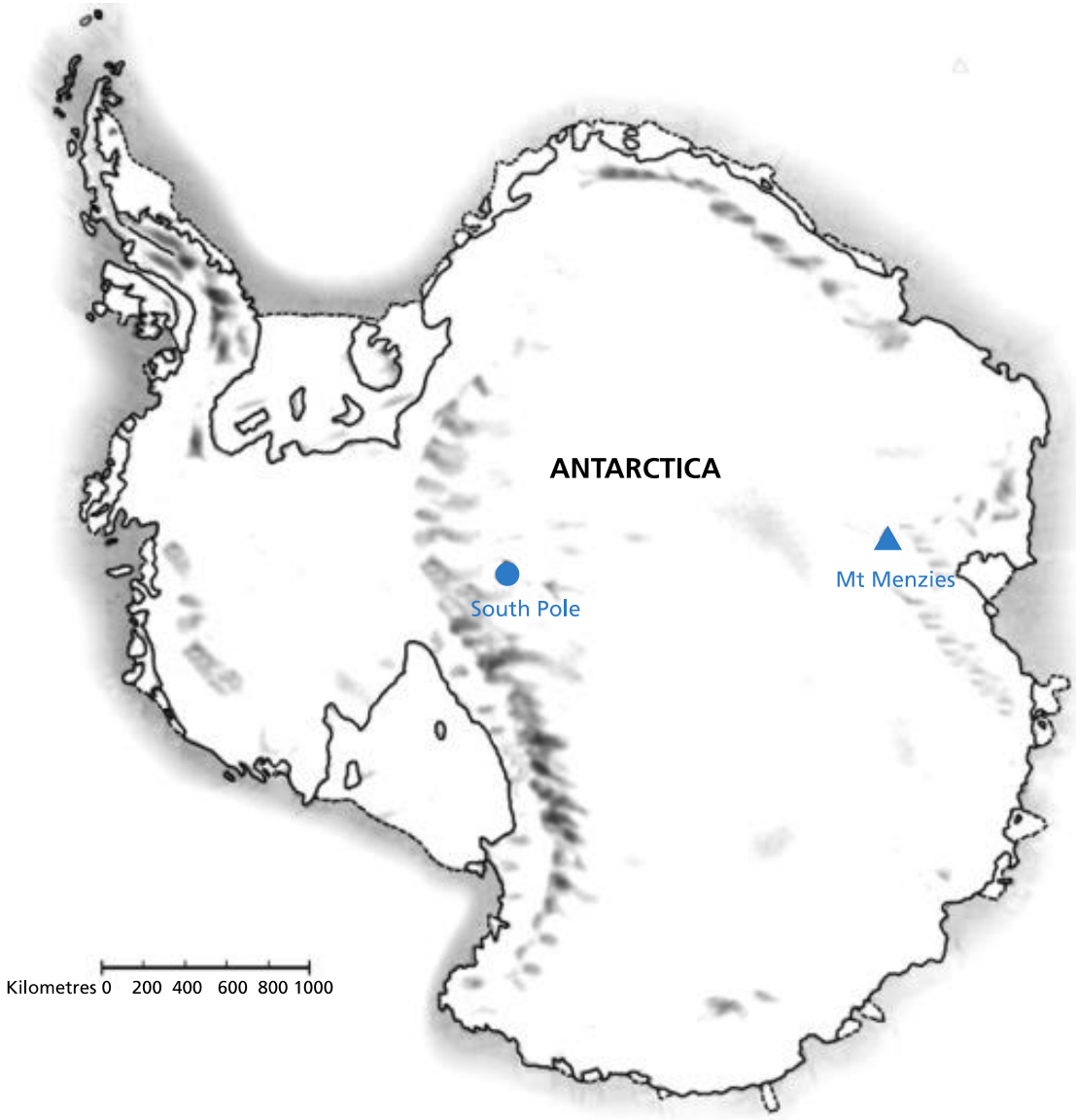
D.



SOURCE: NAEP 2011, Grade 8, Block M8, Item 2 (NCES, 2013).

Mathematics Content Dimension—High Level

Below is a map of Antarctica.



Estimate the area of Antarctica using the map scale.

Show your working out and explain how you made your estimate. (You can draw over the map if it helps you with your estimation)

SOURCE: PISA Mathematics Unit 5, released in 2009.

Mathematics Practices Dimension—Low Level

If $15 + 3x = 42$, then $x =$

A. 9
B. 11
C. 12
D. 14
E. 19

SOURCE: NAEP 2007, Grade 8, Block M9, Item 4 (NCES, 2013).

Mathematics Practices Dimension—Moderate Level

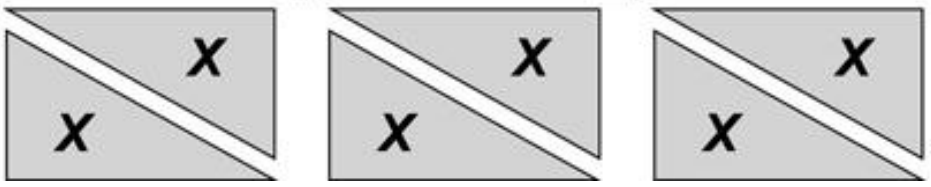
You will need two pieces labeled *X* to answer this question.

2. Use the pieces to make a shape that has these properties.

- It has four sides.
- No pieces overlap.
- No two sides are parallel.

In the space below, trace the shape.

Draw the line to show where the two pieces meet.

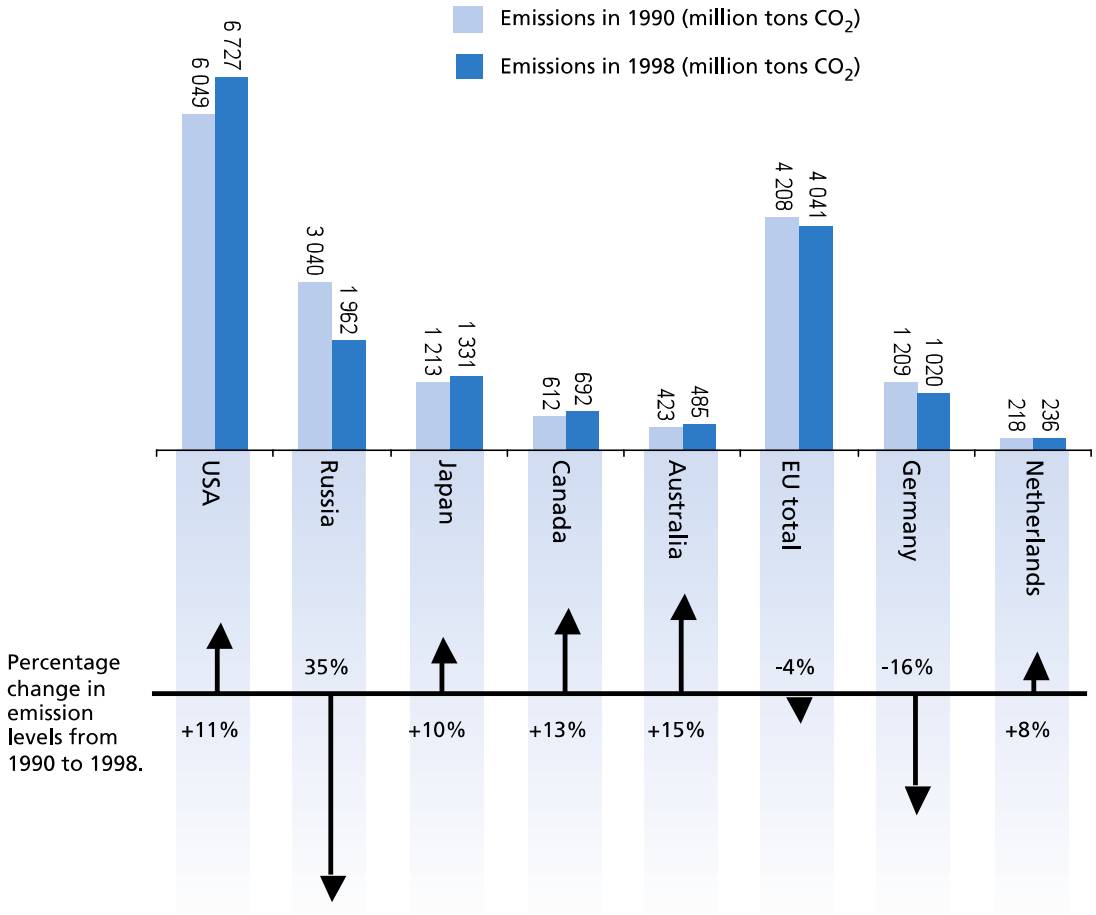


SOURCE: NAEP 2009, Grade 8, Block M5, Item 2 (NCES, 2013).

Mathematics Practices Dimension—High Level

Many scientists fear that the increasing level of CO_2 gas in our atmosphere is causing climate change.

The diagram below shows the CO_2 emission levels in 1990 (the light bars) for several countries (or regions), the emission levels in 1998 (the dark bars), and the percentage change in emission levels between 1990 and 1998 (the arrows with percentages).



Mandy and Niels discussed which country (or region) had the largest increase of CO2 emissions. Each came up with a different conclusion based on the diagram. Give two possible 'correct' answers to this question, and explain how you can obtain each of these answers.

SOURCE: PISA Mathematics Unit 44, released in 2009.

Mathematics Material Dimension—Low Level

If $15 + 3x = 42$, then $x =$

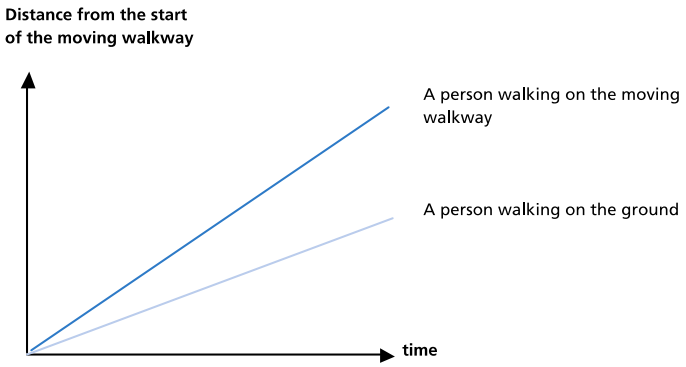
A. 9
B. 11
C. 12
D. 14
E. 19

SOURCE: NAEP 2007, Grade 8, Block M9, Item 4 (NCES, 2013).

Mathematics Material Dimension—Moderate Level

On the right is a photograph of moving walkways.

The following Distance-Time graph shows a comparison between "walking on the moving walkway" and "walking on the ground next to the moving walkway."

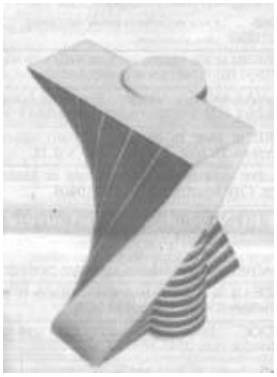


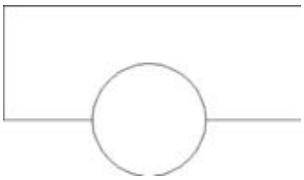
Assuming that, in the above graph, the walking pace is about the same for both persons, add a line to the graph that would represent the distance versus time for a person who is standing still on the moving walkway.

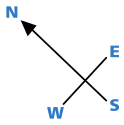
SOURCE: PISA Mathematics Unit 49, released in 2009.

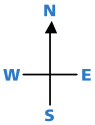
Mathematics Material Dimension—High Level

In modern architecture, buildings often have unusual shapes. The picture below shows a computer model of a 'twisted building' and a plan of the ground floor. The compass points show the orientation of the building.

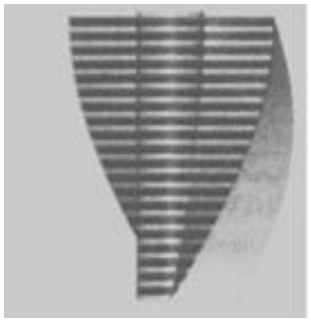








The ground floor of the building contains the main entrance and has room for shops. Above the ground floor there are 20 storys containing apartments. The plan of each story is similar to the plan of the ground floor, but each has a slightly different orientation from the story below. The cylinder contains the elevator shaft and a landing on each floor. The following pictures are sideviews of the twisted building.



Sideview 1

From which direction has Sideview 1 been drawn?

A. From the North.
B. From the West.
C. From the East.
D. From the South.

SOURCE: PISA Mathematics Unit 45, released in 2009.

Mathematics Response Mode Dimension—Low Level

Which of the following is always an odd integer?

A. The product of two odd integers
B. The product of two consecutive integers
C. The sum of three even integers
D. The sum of two odd integers
E. The sum of three consecutive integers

SOURCE: NAEP 2009, Grade 8, Block M5, Item 16 (NCES, 2013).

Mathematics Response Mode Dimension—Moderate Level

Write the next two numbers in the number pattern.

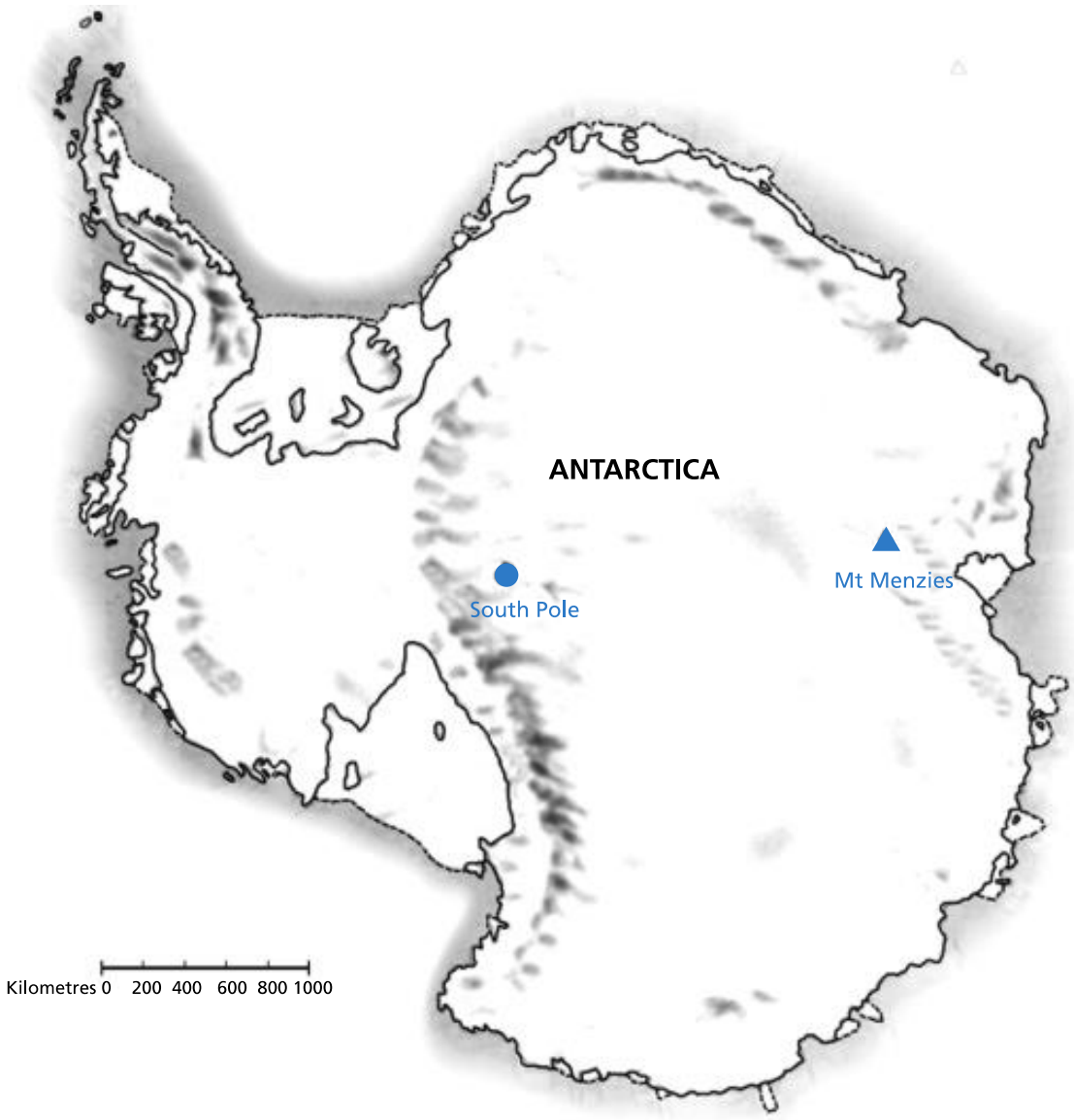
1 6 4 9 7 12 10 ____ ____

Write the rule that you used to find the two numbers you wrote.

SOURCE: NAEP 2009, Grade 8, Block M5, Item 11 (NCES, 2013).

Mathematics Response Mode Dimension—High Level

Below is a map of Antarctica.



Estimate the area of Antarctica using the map scale. Show your working out and explain how you made your estimate. (You can draw over the map if it helps you with your estimation)

SOURCE: PISA Mathematics Unit 5, released in 2009.

Mathematics Processing Demand Dimension—Low Level

In a car park, 762 cars were parked in 6 equal rows. How many cars were in each row?

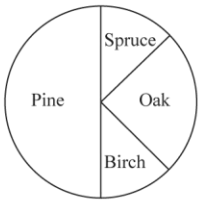
SOURCE: TIMSS 2007, Grade 4, Block M01, Item M031286 (TIMSS, 2007).

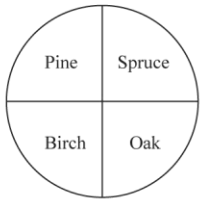
Mathematics Processing Demand Dimension—Moderate Level

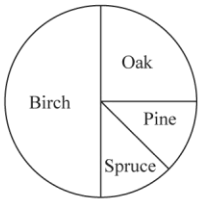
Type of Tree	Number of Trees
Pine	200
Spruce	100
Oak	50
Birch	50

The table above shows the numbers of four types of trees growing in a park. Which of the following pie charts correctly displays the information shown in the table?

Ⓐ 

Ⓑ 

Ⓒ 

Ⓓ 

SOURCE: TIMSS 2007, Grade 4, Block M01, Item M031045 (TIMSS, 2007).

Mathematics Processing Demand Dimension—High Level

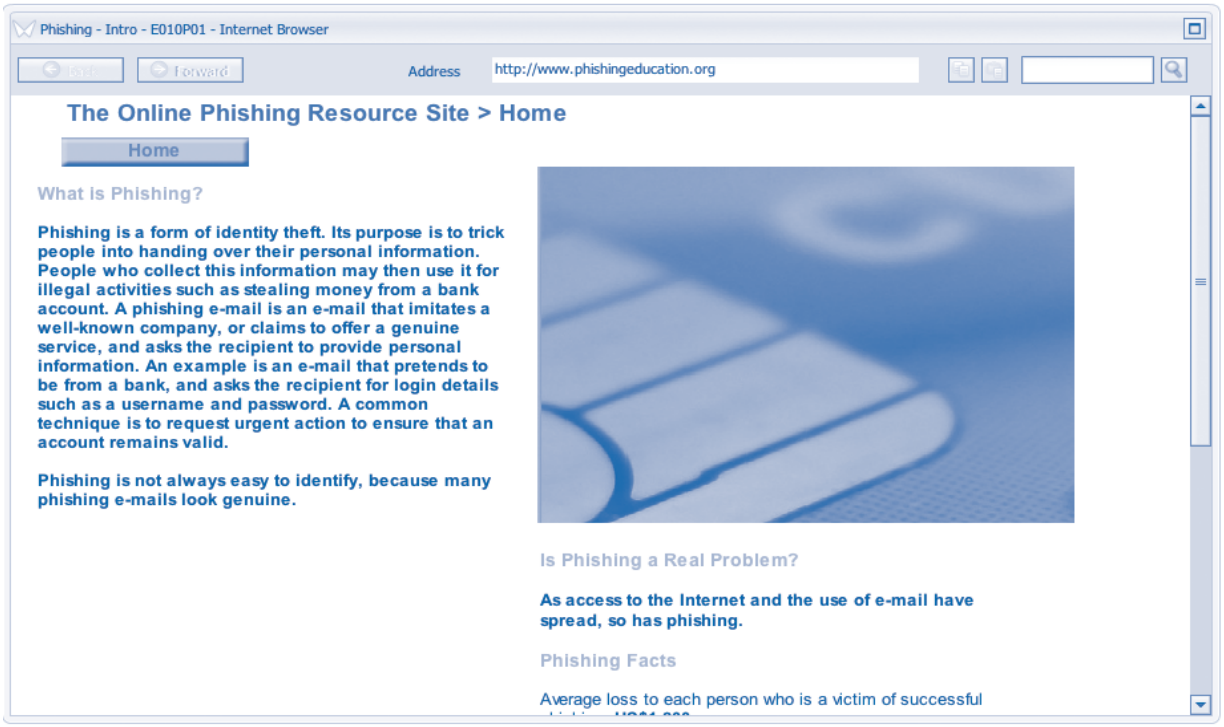
In a Sprinting event, the 'reaction time' is the time interval between the starter's gun firing and the athlete leaving the starting block. The 'final time' includes both this reaction time, and the running time. The following table gives the reaction time and the final time of 8 runners in a 100 metre sprint race.

Lane	Reaction time (sec)	Final time (sec)
1	0.147	10.09
2	0.136	9.99
3	0.197	9.87
4	0.180	Did not finish the race
5	0.210	10.17
6	0.216	10.04
7	0.174	10.08
8	0.193	10.13

To date, no humans have been able to react to a starter's gun in less than 0.110 second. If the recorded reaction time for a runner is less than 0.110 second, then a false start is considered to have occurred because the runner must have left before hearing the gun. If the Bronze medallist had a faster reaction time, would he have had a chance to win the Silver medal? Give an explanation to support your answer.

SOURCE: PISA Mathematics Unit 35, released in 2009

ELA Text Complexity Dimension—Low Level



PHISHING: Task 1 *IE010O021*

1. You are at the Home page of the Online Phishing Resource Site. According to the information on this page, which one of the following is a feature of a phishing e-mail?

A. It asks for personal information
B. It contains unwanted advertising
C. It offers a genuine service
D. It comes from a well-known company

SOURCE: PISA 2009 Electronic Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A2).

ELA Text Complexity Dimension—Moderate Level

Dazzled by so many and such marvellous inventions, the people of Macondo did not know where their amazement began. They stayed up all night looking at the pale electric bulbs fed by the plant that Aureliano Triste had brought back when the train made its second trip, and it took time and effort for them to grow accustomed to its obsessive toom-toom. They became indignant over the living images that the prosperous merchant Don Bruno Crespi projected in the theatre with the lion-head ticket windows, for a character who had died and was buried in one film, and for whose misfortune tears of affliction had been shed, would reappear alive and transformed into an Arab in the next one. The audience, who paid two centavos apiece to share the difficulties of the actors, would not tolerate that outlandish fraud and they broke up the seats. The mayor, at the urging of Don Bruno Crespi, explained by means of a proclamation that the cinema was a machine of illusions that did not merit the emotional outburst of the audience. With that discouraging explanation many felt that they had been the victims of some new and showy gypsy business and they decided not to return to the movies, considering that they already had too many troubles of their own to weep over the acted-out misfortunes of imaginary beings.

1. What feature of the movies caused the people of Macondo to become angry?

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 1 (OECD, 2009, Annex A1).

ELA Text Complexity Dimension—High Level

Democracy in Athens
Part A
Thucydides was a historian and military man who lived in the fifth century BC, during the Classical Greek period. He was born in Athens. During the Peloponnesian War (431 BC to 404 BC) between Athens and Sparta he was in command of a fleet whose mission was to protect the city of Amphipolis in Thrace. He failed to reach the city in time. It fell into the hands of Brasidas, the Spartan general, which forced Thucydides into a twenty-year exile. This granted him the opportunity of collecting detailed information from the two warring factions and the possibility of doing research for his work *History of the Peloponnesian War.*

Thucydides is regarded as one of the great historians of Ancient times. He focuses on natural causes and the behaviour of each individual rather than on fate or the intervention of divinities to explain the evolution of History. In his work, facts are not presented as mere anecdotes; rather, they are explained in an attempt to find out the reasons that led the main characters to act as they did. Thucydides' emphasis on the behaviour of individuals is why he sometimes introduces fictitious speeches: these help him explain the motivations of the historical characters.

Part B
Thucydides attributes to Pericles (fifth century BC), the Athenian ruler, the following speech in honour of the soldiers who fell in the first year of the Peloponnesian War.

Our system of government does not copy the laws of neighbouring states; we are rather a pattern to others than imitators ourselves. Our system is called democracy, since its administration depends on the many instead of the few. Our laws afford equal rights to all in their private affairs, whereas the prestige in public life depends on merit rather than on social class.

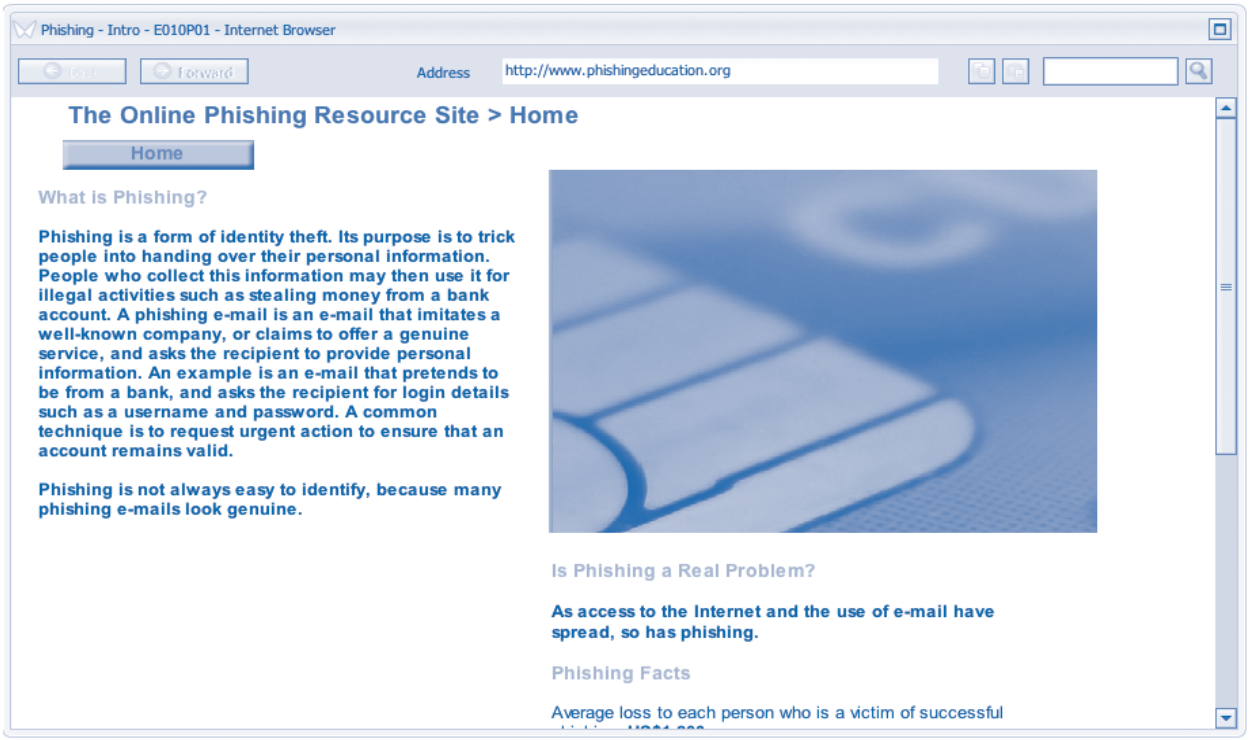
Social class does not prevent a person from holding any public position either (…). And, at the same time that we do not interfere in private affairs, we do not break the law as regards public matters. We give our obedience to those whom we put in positions of authority, and we obey the laws themselves, especially those which are for the protection of the oppressed, and those unwritten laws which it is an acknowledged shame to break.

Furthermore, we provide plenty of means for the pleasure of the mind. The games and sacrifices we celebrate all the year round, and the elegance of our private places of residence, form a daily source of pleasure that helps to banish any worry; while the many inhabitants of the city draw the produce of the world into Athens, so that to the Athenian the fruits of other countries are as familiar as those of his own.

3. One purpose of the speech in Part B was to honor soldiers who fell in the first year of the Peloponnesian War. What was ANOTHER purpose of this speech?

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 7 (OECD, 2009, Annex 1A).

ELA Command of Textual Evidence Dimension—Low Level



PHISHING: Task 1 /E010O021

How many phishing e-mails are sent around the world in an average month?

A 1,200.
B Over 6 billion.
C About 25,000.
D 55,000.

SOURCE: PISA 2009 Electronic Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A2).

ELA Command of Textual Evidence Dimension—Moderate Level

Dazzled by so many and such marvellous inventions, the people of Macondo did not know where their amazement began. They stayed up all night looking at the pale electric bulbs fed by the plant that Aureliano Triste had brought back when the train made its second trip, and it took time and effort for them to grow accustomed to its obsessive toom-toom. They became indignant over the living images that the prosperous merchant Don Bruno Crespi projected in the theatre with the lion-head ticket windows, for a character who had died and was buried in one film, and for whose misfortune tears of affliction had been shed, would reappear alive and transformed into an Arab in the next one. The audience, who paid two centavos apiece to share the difficulties of the actors, would not tolerate that outlandish fraud and they broke up the seats. The mayor, at the urging of Don Bruno Crespi, explained by means of a proclamation that the cinema was a machine of illusions that did not merit the emotional outburst of the audience. With that discouraging explanation many felt that they had been the victims of some new and showy gypsy business and they decided not to return to the movies, considering that they already had too many troubles of their own to weep over the acted-out misfortunes of imaginary beings.

3. At the end of the passage, why did the people of Macondo decide not to return to the movies?
A. They wanted amusement and distraction, but found that the movies were realistic and depressing.
B. They could not afford the ticket prices.
C. They wanted to save their emotions for real-life occasions.
D. They were seeking emotional involvement, but found the movies boring, unconvincing and of poor quality.

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 1 (OECD, 2009, Annex A1).

ELA Command of Textual Evidence Dimension—High Level

Student Opinions

There are so many people out there dying from hunger and disease, yet we are more concerned about future advancements. We are leaving these people behind as we attempt to forget and move on. Billions of dollars are poured into space research by large companies each year. If the money spent on space exploration was used to benefit the needy and not the greedy, the suffering of millions of people could be alleviated.

Ana

The challenge of exploring space is a source of inspiration for many people. For thousands of years we have been dreaming of the heavens, longing to reach out and touch the stars, longing to communicate with something we only imagine could exist, longing to know... Are we alone? Space exploration is a metaphor for learning, and learning is what drives our world. While realists continue to remind us of our current problems, dreamers stretch our minds. It is the dreamers' visions, hopes and desires that will lead us into the future.

Beatrice

We ruin rain forests because there is oil under them, put mines in sacred ground for the sake of uranium. Would we also ruin another planet for the sake of an answer to problems of our own making? Of course! Space exploration strengthens the dangerous belief that human problems can be solved by our ever-increasing domination of the environment. Human beings will continue to feel at liberty to abuse natural resources like rivers and rain forests if we know there is always another planet around the corner waiting to be exploited. We have done enough damage on Earth. We should leave outer space alone.

Dieter

The earth's resources are quickly dying out. The earth's population is increasing at a dramatic rate. Life cannot be sustained if we continue to live in such a way. Pollution has caused a hole in the ozone layer. Fertile lands are running out and soon our food resources will diminish. Already there are cases of famine and disease caused by over-population. Space is a vast empty region which we can use to our benefit. By supporting exploration into space, one day we may find a planet that we can live on. At the moment this seems unimaginable, but the notion of space travel was once thought of as impossible. Discontinuing space exploration in favour of solving immediate problems is a very narrowminded and short-term view. We must learn to think not only for this generation but for the generations to come.

Felix

To ignore what the exploration of space has to offer would be a great loss to all mankind. The possibilities of gaining a greater understanding of the universe and its beginnings are too valuable to waste. The study of other celestial bodies has already increased our understanding of our environmental problems and the possible direction Earth could be heading in if we don't learn to manage our activities. There are also indirect benefits of research into space travel. The creation of laser technology and other medical treatments can be attributed to space research. Substances such as teflon have come out of mankind's quest to travel into space. Thus new technologies created for space research can have immediate benefits for everyone.

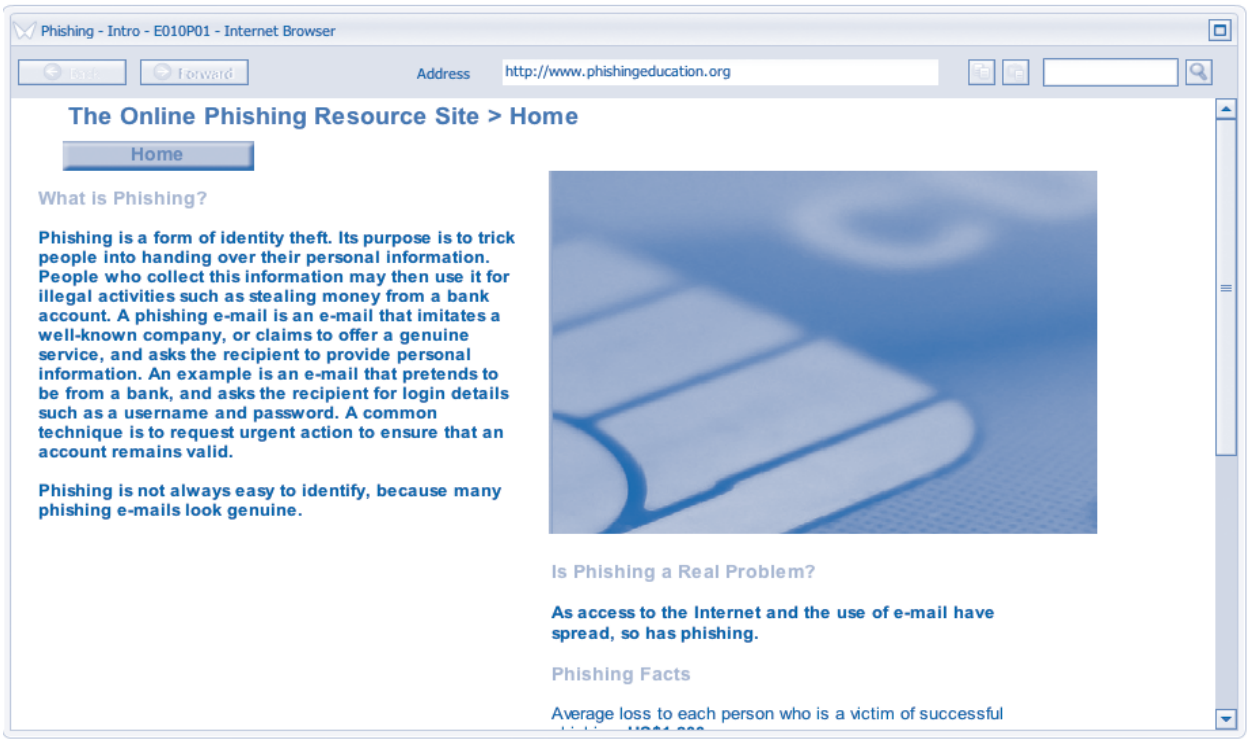
Kate

Which of the following questions do the students seem to be responding to?

A. What is the major problem facing the world today?
B. Are you in favour of space exploration?
C. Do you believe in life beyond our planet?
D. What recent advances have there been in space research?

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A1).

ELA Response Mode Dimension—Low Level



PHISHING: Task 1 /E010O021/

How many phishing e-mails are sent around the world in an average month?

A 1,200.
B Over 6 billion.
C About 25,000.
D 55,000.

SOURCE: PISA 2009 Electronic Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A2).

ELA Response Mode Dimension—Moderate Level

Student Opinions

There are so many people out there dying from hunger and disease, yet we are more concerned about future advancements. We are leaving these people behind as we attempt to forget and move on. Billions of dollars are poured into space research by large companies each year. If the money spent on space exploration was used to benefit the needy and not the greedy, the suffering of millions of people could be alleviated.

Ana

The challenge of exploring space is a source of inspiration for many people. For thousands of years we have been dreaming of the heavens, longing to reach out and touch the stars, longing to communicate with something we only imagine could exist, longing to know... Are we alone? Space exploration is a metaphor for learning, and learning is what drives our world. While realists continue to remind us of our current problems, dreamers stretch our minds. It is the dreamers' visions, hopes and desires that will lead us into the future.

Beatrice

We ruin rain forests because there is oil under them, put mines in sacred ground for the sake of uranium. Would we also ruin another planet for the sake of an answer to problems of our own making? Of course! Space exploration strengthens the dangerous belief that human problems can be solved by our ever-increasing domination of the environment. Human beings will continue to feel at liberty to abuse natural resources like rivers and rain forests if we know there is always another planet around the corner waiting to be exploited. We have done enough damage on Earth. We should leave outer space alone.

Dieter

The earth's resources are quickly dying out. The earth's population is increasing at a dramatic rate. Life cannot be sustained if we continue to live in such a way. Pollution has caused a hole in the ozone layer. Fertile lands are running out and soon our food resources will diminish. Already there are cases of famine and disease caused by over-population. Space is a vast empty region which we can use to our benefit. By supporting exploration into space, one day we may find a planet that we can live on. At the moment this seems unimaginable, but the notion of space travel was once thought of as impossible. Discontinuing space exploration in favour of solving immediate problems is a very narrowminded and short-term view. We must learn to think not only for this generation but for the generations to come.

Felix

To ignore what the exploration of space has to offer would be a great loss to all mankind. The possibilities of gaining a greater understanding of the universe and its beginnings are too valuable to waste. The study of other celestial bodies has already increased our understanding of our environmental problems and the possible direction Earth could be heading in if we don't learn to manage our activities. There are also indirect benefits of research into space travel. The creation of laser technology and other medical treatments can be attributed to space research. Substances such as teflon have come out of mankind's quest to travel into space. Thus new technologies created for space research can have immediate benefits for everyone.

Kate

6. Thinking about the main ideas presented by the five students, which student do you agree with most strongly?

Student's name: .....

Using your own words, explain your choice by referring to your own opinion and the main ideas presented by the student.

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A1).

ELA Response Mode Dimension—High Level

Which one of the writers most directly contradicts Felix's argument?

A. Dieter.
B. Ana.
C. Kate.
D. Beatrice.

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A1).

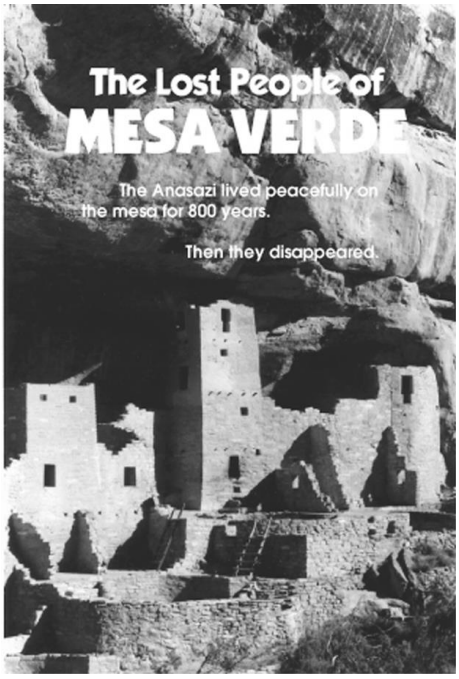
ELA Processing Demands Dimension—Low Level

Dazzled by so many and such marvellous inventions, the people of Macondo did not know where their amazement began. They stayed up all night looking at the pale electric bulbs fed by the plant that Aureliano Triste had brought back when the train made its second trip, and it took time and effort for them to grow accustomed to its obsessive toom-toom. They became indignant over the living images that the prosperous merchant Don Bruno Crespi projected in the theatre with the lion-head ticket windows, for a character who had died and was buried in one film, and for whose misfortune tears of affliction had been shed, would reappear alive and transformed into an Arab in the next one. The audience, who paid two centavos apiece to share the difficulties of the actors, would not tolerate that outlandish fraud and they broke up the seats. The mayor, at the urging of Don Bruno Crespi, explained by means of a proclamation that the cinema was a machine of illusions that did not merit the emotional outburst of the audience. With that discouraging explanation many felt that they had been the victims of some new and showy gypsy business and they decided not to return to the movies, considering that they already had too many troubles of their own to weep over the acted-out misfortunes of imaginary beings.

1. What feature of the movies caused the people of Macondo to become angry?

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 1 (OECD, 2009, Annex A1).

ELA Processing Demands Dimension—Moderate Level



By Elsa Marston The Image Bank

In the dry land of southwestern Colorado a beautiful plateau rises. It has so many trees that early Spanish explorers called it Mesa Verde, which means "green table." For about eight hundred years Native Americans called the Anasazi lived on this mesa. And then they left. Ever since the cliff houses were first discovered a hundred years ago, scientists and historians have wondered why.

Anasazi is a Navajo word meaning "the ancient ones." When they first settled there, around 500 A. D., the Anasazi lived in alcoves in the walls of the high canyons. Later they moved to the level land on top, where they built houses of stone and mud mortar. As time passed, they constructed more elaborate houses, like apartment buildings, with several families living close together.

The Anasazi made beautiful pottery, turquoise jewelry, fine sashes of woven hair, and baskets woven tightly enough to hold water. They lived by hunting and by growing corn and squash. Their way of life went on peacefully for several hundred years.

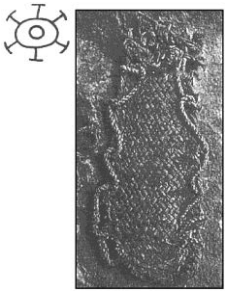
Then around 1200 A.D. something strange happened, for which the reasons are not quite clear. Most of the people moved from the level plateau back down into alcoves in the cliffs. The move must have made their

lives difficult because they had to climb back up to the plateau to do the farming. But it seems the Anasazi planned to stay in the canyon walls, for they soon filled the alcoves with amazing cliff dwellings. "Cliff Palace," the most famous of these, had more than two hundred rooms.

For all the hard work that went into building these new homes, the Anasazi did not live in them long. By 1300 A.D. the cliff dwellings were empty. Mesa Verde was deserted and remained a ghost country for almost six hundred years. Were the people driven out of their homes by enemies? No sign of attack or fighting, or even the presence of other tribes, has been found.

Archaeologists who have studied the place now believe there are other reasons. Mesa Verde, the beautiful green table, was no longer a good place to live. For one thing, in the second half of the thirteenth century there were long periods of cold, and very little rain fell—or else it came at the wrong time of year. Scientists know this from examining the wood used in the cliff dwellings. The growth rings in trees show good and bad growing seasons. But the people had survived drought and bad weather before, so there must have been another reason.

The sturdy baskets, woven sandals, and beautiful pottery left behind by the Anasazi may be 1,000 years old.



Bureau of Land Management -Anasazi Heritage Center Collections

As the population grew, more land on the mesa top had to be farmed in order to feed the people. That meant that trees had to be cut to clear the land and also to use for houses and fuel. Without the forests, the rain began to wash away the mesa top.

How do we know about erosion problems that happened about eight hundred years ago? The Anasazi built many low dams across the smaller valleys on the mesa to slow down rain runoff. Even so, good soil washed away, and the people could no longer raise enough food. As the forests dwindled, the animals, already over-hunted, left the mesa for mountainous areas with more trees.

And as the mesa "wore out," so did the people. It appears that the Anasazi were not healthy. Scientists can learn a lot about ancient people's health by studying the bones and teeth found in burials. The mesa dwellers had arthritis, and their teeth were worn down by the grit in corn meal, a main part of their diet.

As food became scarce, people grew weaker. Not many lived

3. If you had lived with the Anasazi at Mesa Verde, would you have preferred living on the top of the mesa or in the cliff houses built into the alcoves? Explain your preference by using information from the article.

SOURCE: NAEP 2007, Sample Reading Questions, Grade 8 (NCES, 2007).

ELA Processing Demands Dimension—High Level

Student Opinions

There are so many people out there dying from hunger and disease, yet we are more concerned about future advancements. We are leaving these people behind as we attempt to forget and move on. Billions of dollars are poured into space research by large companies each year. If the money spent on space exploration was used to benefit the needy and not the greedy, the suffering of millions of people could be alleviated.

Ana

The challenge of exploring space is a source of inspiration for many people. For thousands of years we have been dreaming of the heavens, longing to reach out and touch the stars, longing to communicate with something we only imagine could exist, longing to know... Are we alone? Space exploration is a metaphor for learning, and learning is what drives our world. While realists continue to remind us of our current problems, dreamers stretch our minds. It is the dreamers' visions, hopes and desires that will lead us into the future.

Beatrice

We ruin rain forests because there is oil under them, put mines in sacred ground for the sake of uranium. Would we also ruin another planet for the sake of an answer to problems of our own making? Of course! Space exploration strengthens the dangerous belief that human problems can be solved by our ever-increasing domination of the environment. Human beings will continue to feel at liberty to abuse natural resources like rivers and rain forests if we know there is always another planet around the corner waiting to be exploited. We have done enough damage on Earth. We should leave outer space alone.

Dieter

The earth's resources are quickly dying out. The earth's population is increasing at a dramatic rate. Life cannot be sustained if we continue to live in such a way. Pollution has caused a hole in the ozone layer. Fertile lands are running out and soon our food resources will diminish. Already there are cases of famine and disease caused by over-population. Space is a vast empty region which we can use to our benefit. By supporting exploration into space, one day we may find a planet that we can live on. At the moment this seems unimaginable, but the notion of space travel was once thought of as impossible. Discontinuing space exploration in favour of solving immediate problems is a very narrowminded and short-term view. We must learn to think not only for this generation but for the generations to come.

Felix

To ignore what the exploration of space has to offer would be a great loss to all mankind. The possibilities of gaining a greater understanding of the universe and its beginnings are too valuable to waste. The study of other celestial bodies has already increased our understanding of our environmental problems and the possible direction Earth could be heading in if we don't learn to manage our activities. There are also indirect benefits of research into space travel. The creation of laser technology and other medical treatments can be attributed to space research. Substances such as teflon have come out of mankind's quest to travel into space. Thus new technologies created for space research can have immediate benefits for everyone.

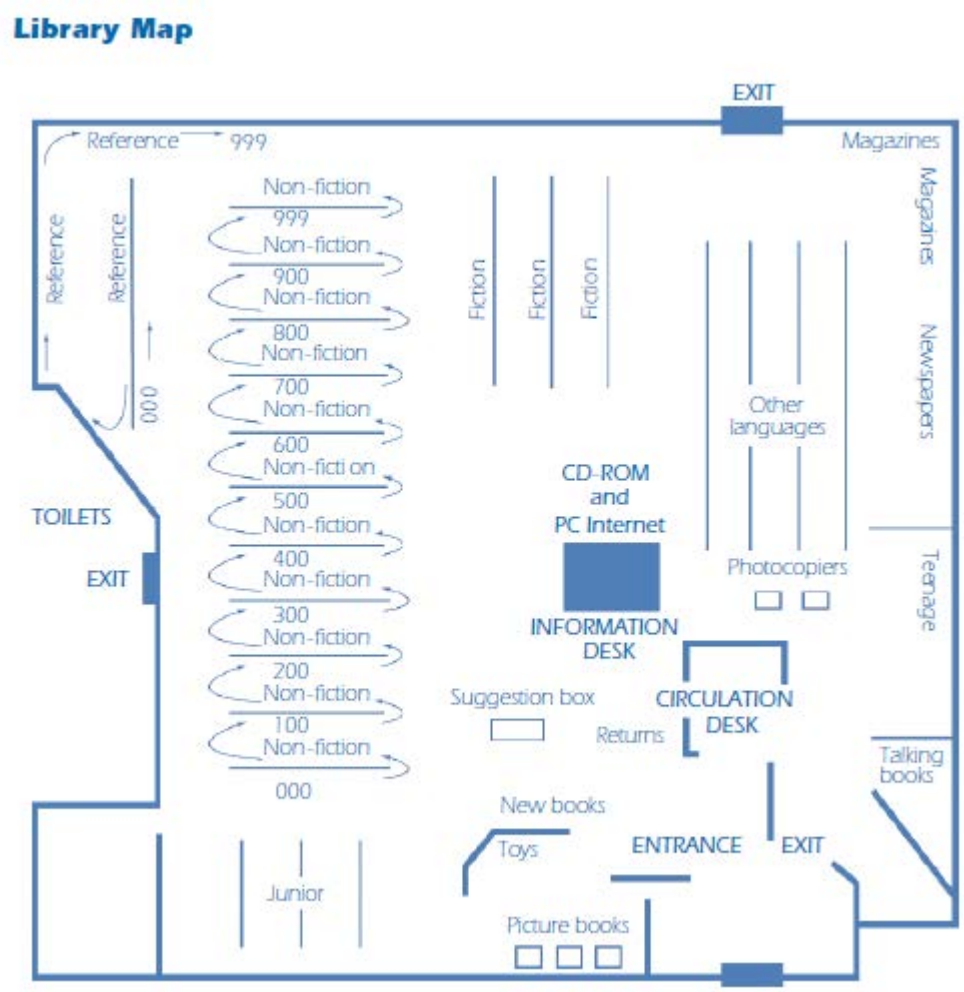
Kate

Some statements are matters of opinion, based on the ideas and values of the writer. Some statements are matters of fact, which may be tested objectively and are either correct or incorrect. Draw a circle around "matter of opinion" or "matter of fact" next to each of the quotations from the students' writing listed below.

Quotation from students' writing	Matter of opinion or Matter of fact?
"Billions of dollars are poured into space research by large companies each year." (Ana)	Matter of opinion / Matter of fact
"Space exploration strengthens the dangerous belief that human problems can be solved by our ever-increasing domination of the environment." (Dieter)	Matter of opinion / Matter of fact
"Discontinuing space exploration in favour of solving immediate problems is a very narrow-minded and short-term view." (Felix)	Matter of opinion / Matter of fact

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 3 (OECD, 2009, Annex A1).

ELA Stimulus Material Dimension—Low Level



For school you need to read a novel in French. On the map draw a circle around the section where you would be most likely to find a suitable book to borrow.

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 2 (OECD, 2009, Annex A1).

ELA Stimulus Material Dimension—Moderate Level

"Tall buildings" is an article from a Norwegian magazine published in 2006.

Figure 1. **Tall Buildings of the World**

Figure 1 shows the number of buildings of at least 30 storeys that have been built, or are under construction This includes buildings that have been proposed since January 2001.

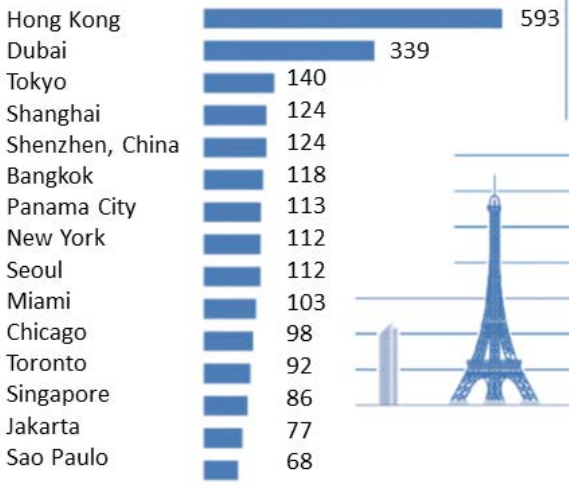
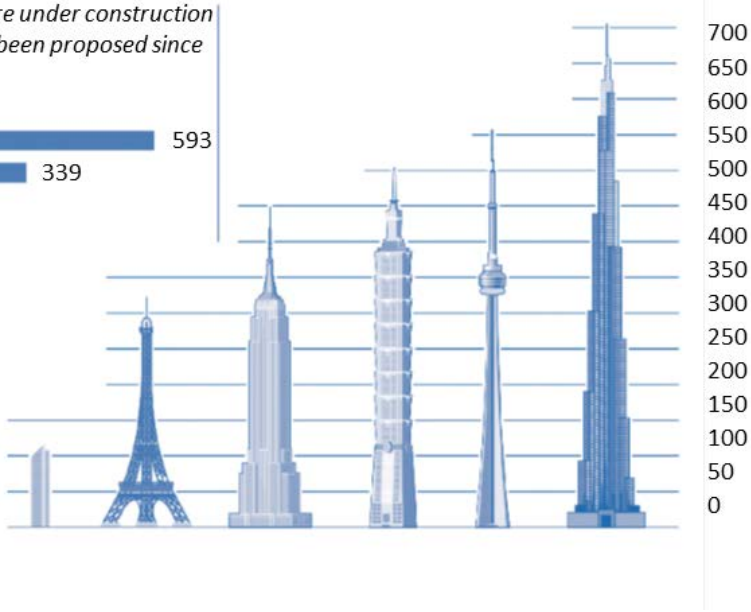


Figure 2: ***Some of the World's Tallest Buildings***

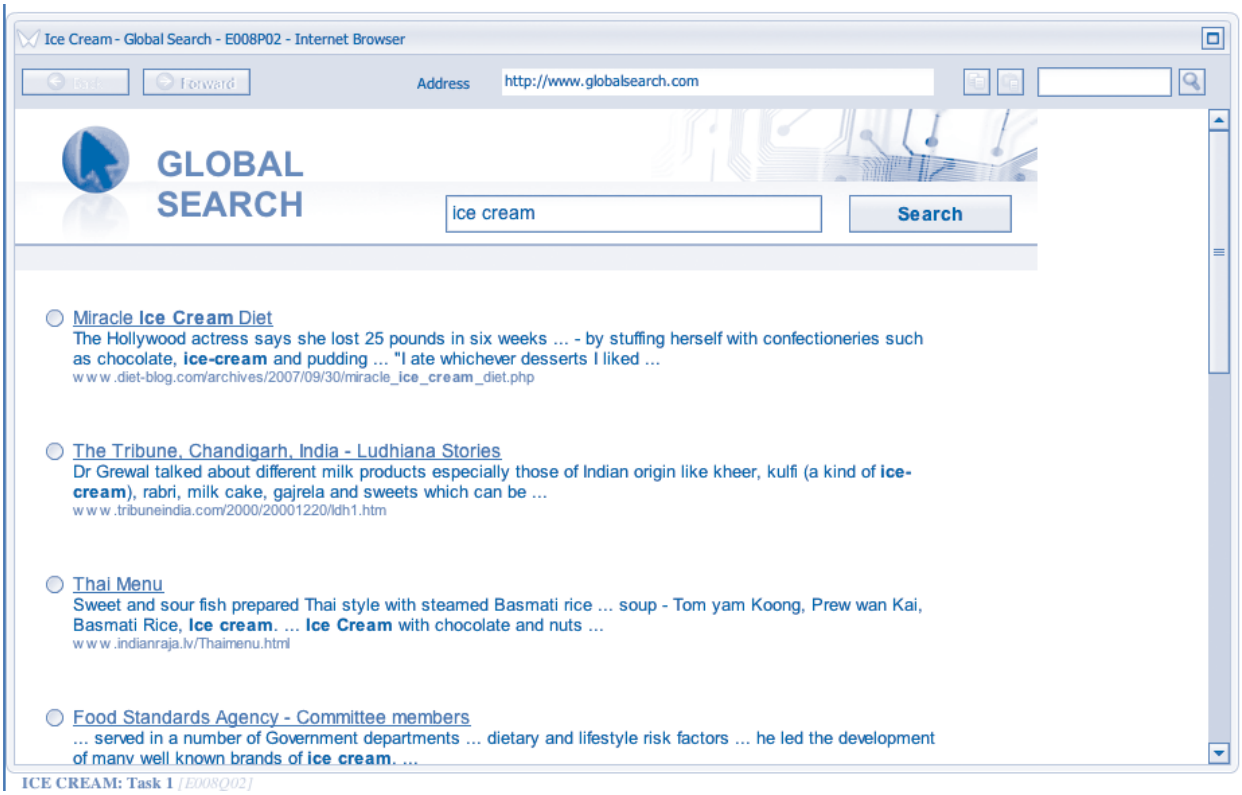
The Burj Tower in Dubai is expected to be the tallest building in the world, at 700 metres, when it is finished in 2008.



1. When the magazine article was published, which of the buildings in Figure 2 was the tallest completed building?

SOURCE: PISA 2009 Print Reading Sample Tasks, Unit 6 (OECD, 2009, Annex A1).

ELA Stimulus Material Dimension—High Level



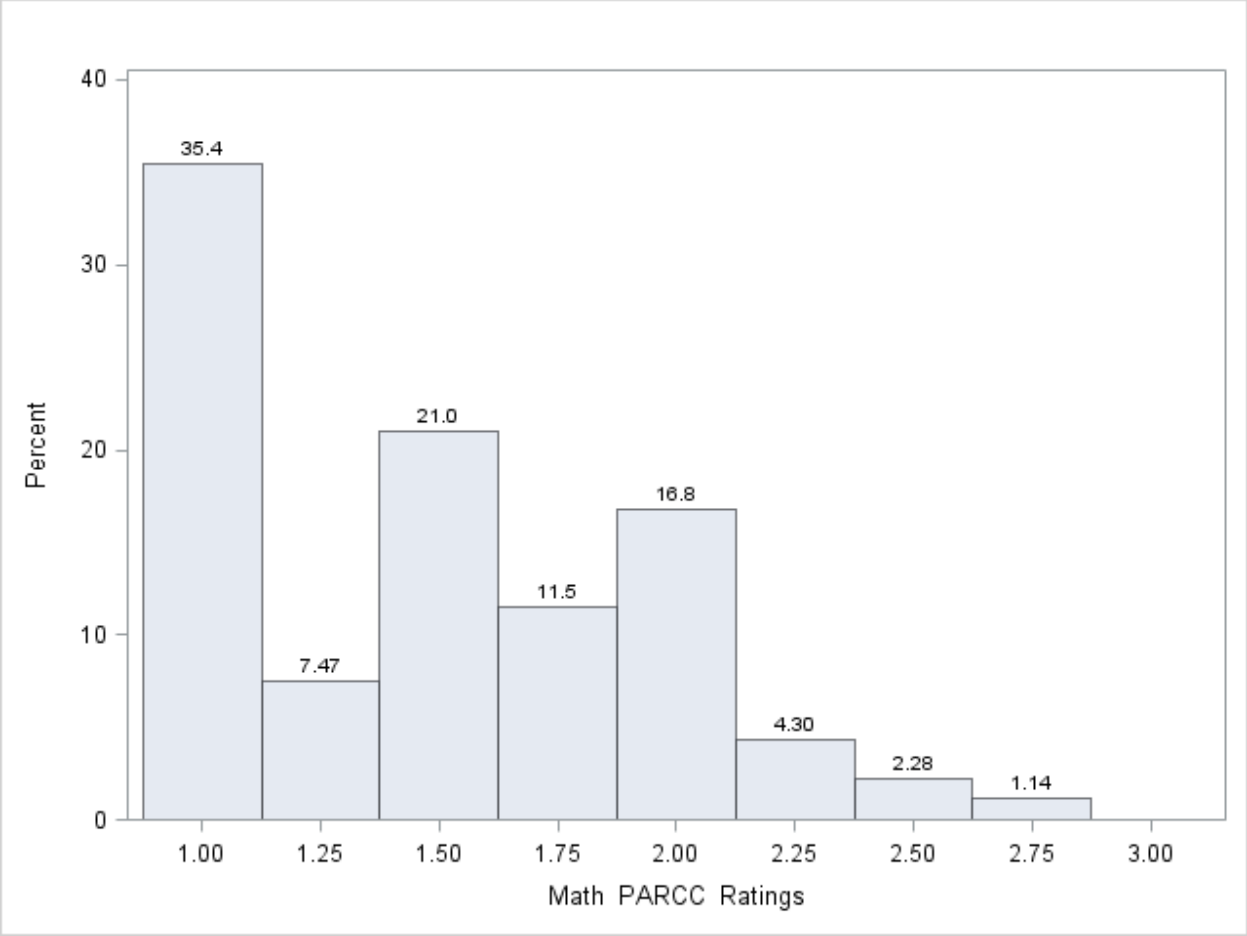
1. This page shows search results for ice cream and similar foods from around the world. Which search results in most likely to provide a history of ice cream? Click the button next to the link.

SOURCE: PISA 2009 Electronic Reading Sample Tasks, Unit 2 (OECD, 2009, Annex A2).

Appendix E. Distribution of Modified PARCC Ratings

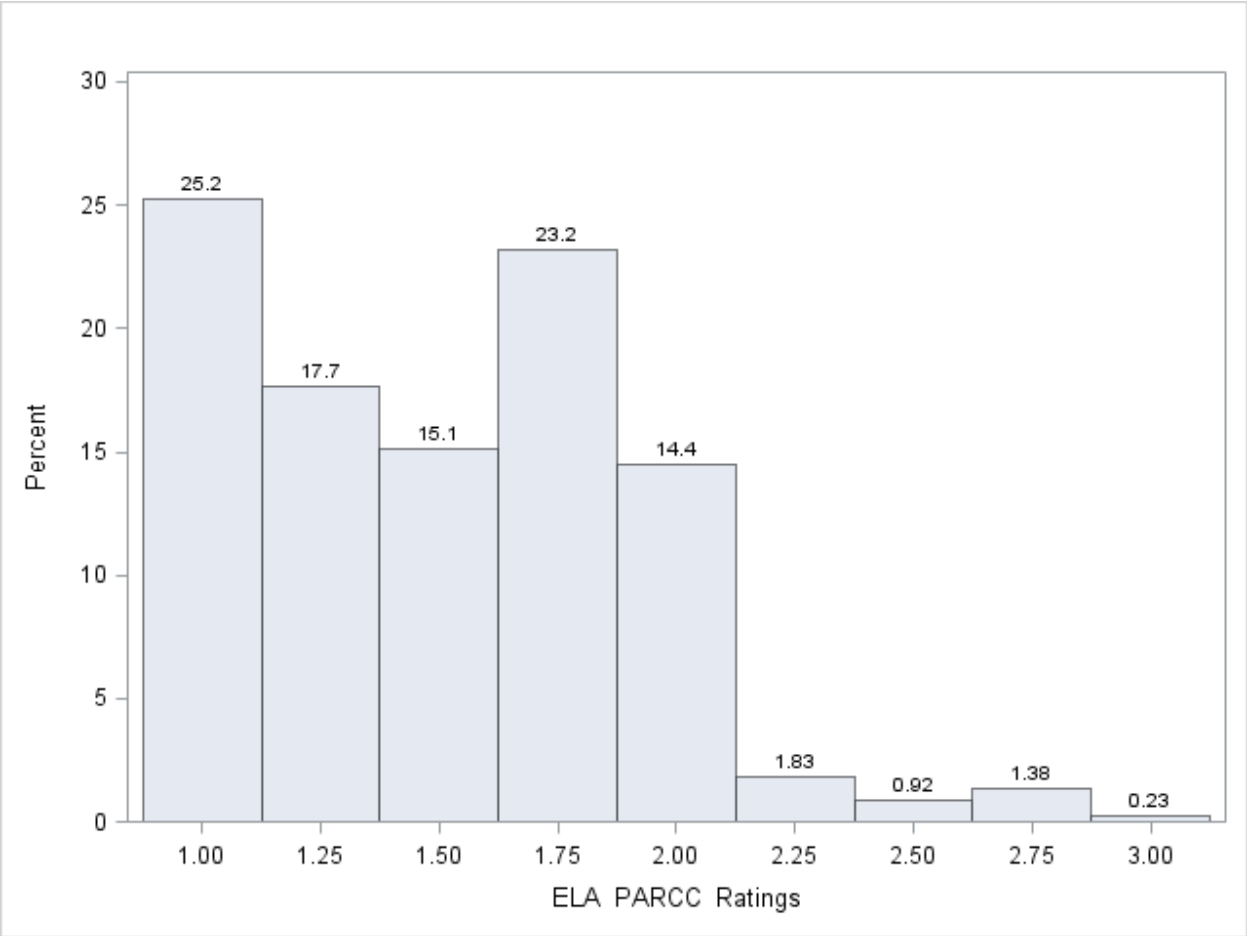
This appendix provides the distribution for the modified PARCC ratings used in the analysis.

Figure E.1. Distribution of Modified PARCC Ratings for Mathematics



We used the same cut scores on the ratings obtained under the proposed PARCC scoring system to create four-point ratings (see Figure E.1). We then compared the agreement between the four-point ratings under the proposed PARCC scoring system and the four-point ratings under our modified scoring system. The weighted kappa was very high, at 0.80.

Figure E.2. Distribution of Modified PARCC Ratings for ELA



As with mathematics, we created four-point ratings under the proposed PARCC scoring system, then compared those four-point ratings to the four-point ratings under our modified scoring system (see Figure E.2). The weighted kappa was 0.74.

Appendix F. Results for the Original PARCC Dimensions and Levels

This appendix presents the results for each content dimension of the PARCC framework. Figures F.1–F.5 show the percentage of test items rated at each level on the original PARCC dimensions. Tables F.1–F.15 show the percentage of released items at each level, by test and item format.

Figure F.1. Percentage of Test Items Rated at Each Level on PARCC Dimensions for Mathematics

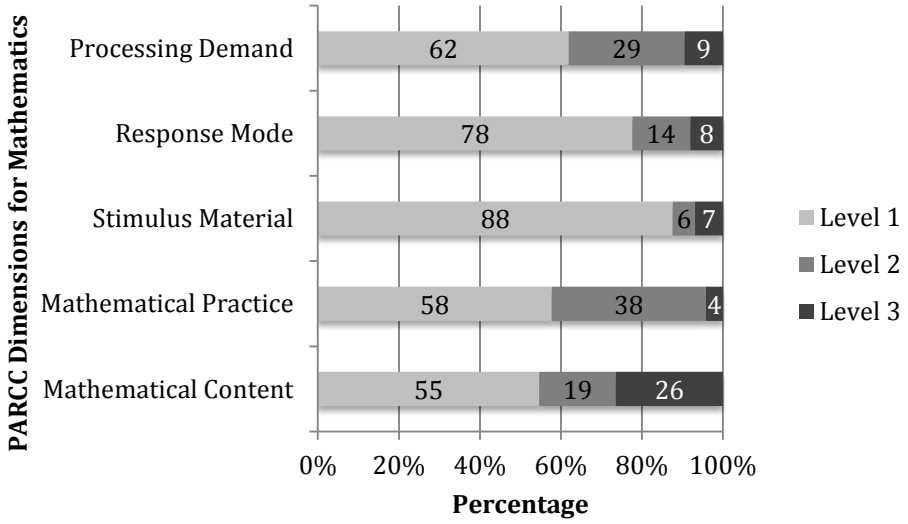


Figure F.2. Percentage of Test Items Rated at Each Level on PARCC Dimensions for Mathematics, by Dimension and Item Format

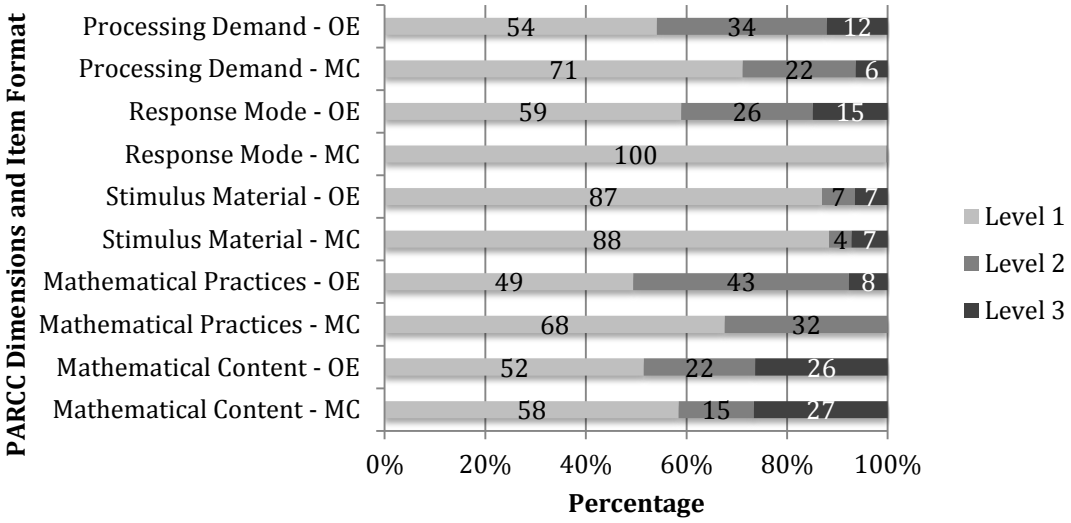


Figure F.3. Percentage of Reading Test Items Rated at Each Level on PARCC Dimensions for ELA

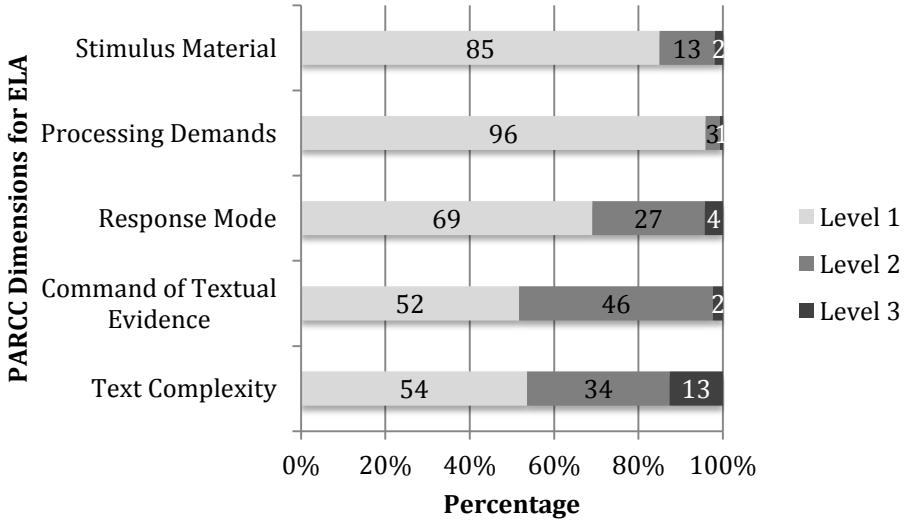


Figure F.4. Percentage of Reading Test Items Rated at Each Level on PARCC Dimensions for ELA, by Dimension and Item Format

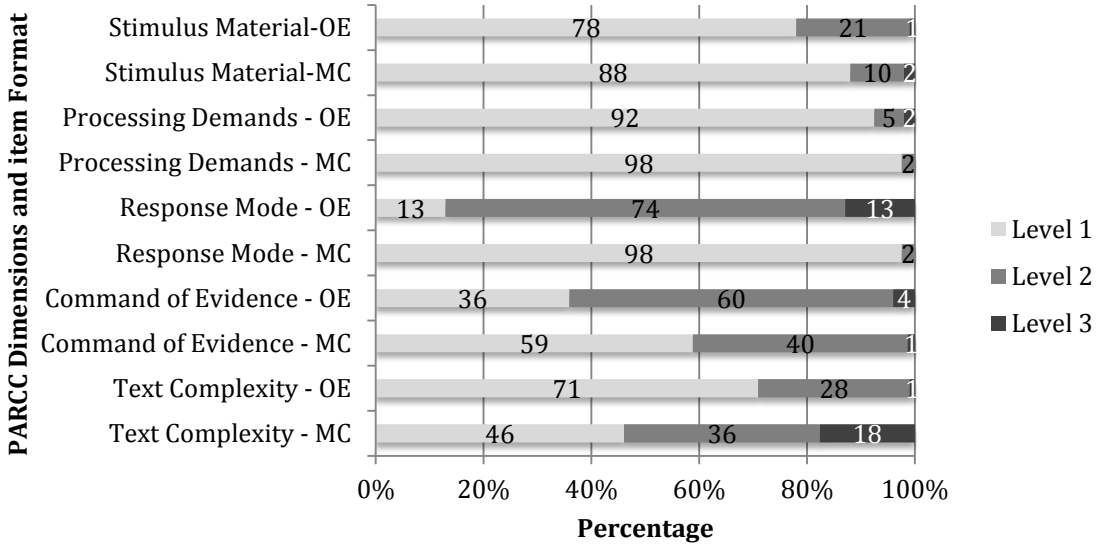


Figure F.5. Percentage of Writing Test Items Rated at Each Level on PARCC Dimensions for ELA

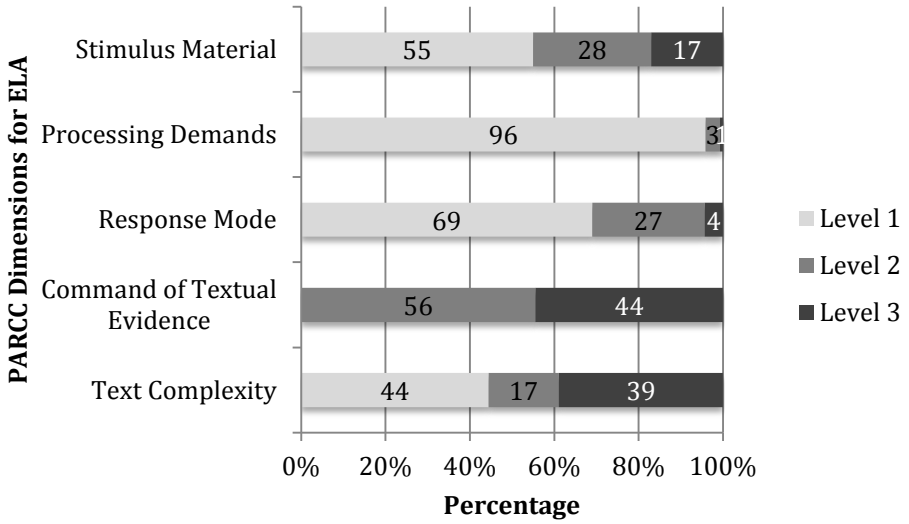


Table F.1. Percentage of Released Mathematics Test Items at Each Level for the PARCC Content Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	193	67%	33%	130	5%	26%	68%	63	17%	16%	67%
IB	157	0%	100%	0	—	—	—	157	22%	38%	40%
NAEP	190	68%	32%	129	91%	7%	2%	61	89%	11%	0%
PISA	89	24%	76%	21	62%	33%	5%	68	69%	22%	9%
TIMSS	161	50%	50%	81	90%	5%	5%	80	94%	4%	2%
Total	790	46%	54%	361	58%	15%	27%	429	52%	22%	26%

Table F.2. Percentage of Released Mathematics Test Items at Each Level for the PARCC Practice Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	193	67%	33%	130	44%	56%	0%	63	40%	48%	12%
IB	157	0%	100%	0	—	—	—	157	36%	52%	12%
NAEP	190	68%	32%	129	82%	18%	0%	61	56%	38%	6%
PISA	89	24%	76%	21	76%	24%	0%	68	60%	35%	5%
TIMSS	161	50%	50%	81	80%	20%	0%	80	69%	31%	0%
Total	790	46%	54%	361	68%	32%	0%	429	49%	43%	8%

Table F.3. Percentage of Released Mathematics Test Items at Each Level for the PARCC Material Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	193	67%	33%	130	98%	2%	0%	63	100%	0%	0%
IB	157	0%	100%	0	—	–	—	157	100%	0%	0%
NAEP	190	68%	32%	129	86%	6%	7%	61	80%	3%	16%
PISA	89	24%	76%	21	57%	19%	24%	68	60%	22%	18%
TIMSS	161	50%	50%	81	84%	2%	14%	80	79%	14%	7%
Total	790	46%	54%	361	88%	4%	7%	429	87%	6%	7%

Table F.4. Percentage of Released Mathematics Test Items at Each Level for the PARCC Response Mode Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	193	67%	33%	130	100%	0%	0%	63	28%	32%	40%
IB	157	0%	100%	0	—	—	—	157	47%	38%	15%
NAEP	190	68%	32%	129	100%	0%	0%	61	66%	28%	6%
PISA	89	24%	76%	21	100%	0%	0%	68	71%	13%	16%
TIMSS	161	50%	50%	81	100%	0%	0%	80	91%	9%	0%
Total	790	46%	54%	361	100%	0%	0%	429	59%	26%	15%

Table F.5. Percentage of Released Mathematics Test Items at Each Level for the PARCC Processing Demand Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	193	67%	33%	130	64%	22%	14%	63	16%	43%	41%
IB	157	0%	100%	0	—	—	—	157	71%	24%	5%
NAEP	190	68%	32%	129	78%	21%	1%	61	59%	39%	2%
PISA	89	24%	76%	21	19%	67%	14%	68	13%	64%	22%
TIMSS	161	50%	50%	81	86%	12%	1%	80	81%	16%	3%
Total	790	46%	54%	361	71%	22%	7%	429	54%	34%	12%

Table F.6. Percentage of Released Reading Test Items at Each Level for the PARCC Text Complexity Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	107	100%	0%	107	0%	54%	46%	0	—	—	—
NAEP	147	63%	37%	92	75%	25%	0%	55	66%	34%	0%
PISA	110	55%	45%	60	57%	40%	3%	50	62%	34%	4%
PIRLS	54	56%	44%	30	100%	0%	0%	24	100%	0%	0%
Total	418	69%	31%	289	46%	36%	18%	129	71%	28%	1%

NOTE: IB is not included in the table because IB ELA tests did not have any reading items.

Table F.7. Percentage of Released Reading Test Items at Each Level for the PARCC Command of Textual Evidence Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	107	100%	0%	107	36%	64%	0%	0	—	—	—
NAEP	147	63%	37%	92	80%	20%	0%	55	16%	78%	6%
PISA	110	55%	45%	60	55%	38%	7%	50	44%	50%	6%
PIRLS	54	56%	44%	30	80%	20%	0%	24	62%	38%	0%
Total	418	69%	31%	289	59%	40%	1%	129	36%	60%	4%

NOTE: IB is not included in the table because IB ELA tests did not have any reading items.

Table F.8. Percentage of Released Reading Test Items at Each Level for the PARCC Response Mode Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	107	100%	0%	107	100%	0%	0%	0	—	—	—
NAEP	147	63%	37%	92	100%	0%	0%	55	13%	85%	2%
PISA	110	55%	45%	60	100%	0%	0%	50	16%	84%	0%
PIRLS	54	56%	44%	30	100%	0%	0%	24	17%	83%	0%
Total	418	69%	31%	289	100%	0%	0%	129	15%	85%	0%

NOTE: IB is not included in the table because IB ELA tests did not have any reading items.

Table F.9. Percentage of Released Reading Test Items at Each Level for the PARCC Processing Demand Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	107	100%	0%	107	100%	0%	0%	0	—	—	—
NAEP	147	63%	37%	92	99%	1%	0%	55	98%	2%	0%
PISA	110	55%	45%	60	90%	10%	0%	50	98%	2%	0%
PIRLS	54	56%	44%	30	100%	0%	0%	24	96%	4%	0%
Total	418	69%	31%	289	98%	2%	0%	129	98%	2%	0%

NOTE: IB is not included in the table because IB ELA tests did not have any reading items.

Table F.10. Percentage of Released Reading Test Items at Each Level for the PARCC Stimulus Material Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	107	100%	0%	107	100%	0%	0%	0	—	—	—
NAEP	147	63%	37%	92	90%	10%	0%	55	85%	15%	0%
PISA	110	55%	45%	60	70%	22%	8%	50	71%	27%	2%
PIRLS	54	56%	44%	30	77%	23%	0%	24	79%	21%	0%
Total	418	69%	31%	289	88%	10%	2%	129	78%	21%	1%

NOTE: IB is not included in the table because IB ELA tests did not have any reading items.

Table F.11. Percentage of Released Writing Test Items at Each Level for the PARCC Text Complexity Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	6	0%	100%	0	—	—	—	6	33%	17%	50%
IB	6	0%	100%	0	—	—	—	6	0%	33%	67%
NAEP	6	0%	100%	0	—	—	—	6	100%	0%	0%
Total	18	0%	100%	0	—	—	—	18	44%	17%	39%

NOTE: Tests that did not assess writing are not included in the table.

Table F.12. Percentage of Released Writing Test Items at Each Level for the PARCC Command of Textual Evidence Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	6	0%	100%	0	—	—	—	6	0%	83%	17%
IB	6	0%	100%	0	—	—	—	6	0%	0%	100%
NAEP	6	0%	100%	0	—	—	—	6	0%	83%	17%
Total	18	0%	100%	0	—	—	—	18	0%	56%	44%

NOTE: Tests that did not assess writing are not included in the table.

Table F.13. Percentage of Released Writing Test Items at Each Level for the PARCC Response Mode Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	6	0%	100%	0	—	—	—	6	0%	0%	100%
IB	6	0%	100%	0	—	—	—	6	0%	0%	100%
NAEP	6	0%	100%	0	—	—	—	6	0%	0%	100%
Total	18	0%	100%	0	—	—	—	18	0%	0%	100%

NOTE: Tests that did not assess writing are not included in the table.

Table F.14. Percentage of Released Writing Test Items at Each Level for the PARCC Processing Demand Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	6	0%	100%	0	—	—	—	6	0%	50%	50%
IB	6	0%	100%	0	—	—	—	6	67%	33%	0%
NAEP	6	0%	100%	0	—	—	—	6	100%	0%	0%
Total	18	0%	100%	0	—	—	—	18	56%	28%	16%

NOTE: Tests that did not assess writing are not included in the table.

Table F.15. Percentage of Released Writing Test Items at Each Level for the PARCC Stimulus Material Dimension, by Test and Item Format

Test	All Items			MC Items				OE Items			
	N	MC	OE	N	Level 1	Level 2	Level 3	N	Level 1	Level 2	Level 3
AP	6	0%	100%	0	—	—	—	6	83%	17%	50%
IB	6	0%	100%	0	—	—	—	6	0%	67%	33%
NAEP	6	0%	100%	0	—	—	—	6	83%	67%	0%
Total	18	0%	100%	0	—	—	—	18	56%	28%	16%

NOTE: Tests that did not assess writing are not included in the table.

References

Binkley, Marilyn, and Dana L. Kelly, *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments*, Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, NCES 2003-10, April 2003.

College Board, "May 2014 AP Exam Format," 2014. As of February 10, 2014: http://apcentral.collegeboard.com/apc/public/repository/AP_Exam_Format.pdf

Darling-Hammond, Linda, Joan Herman, James Pellegrino, Jamal Abedi, J. Lawrence Aber, Eva Baker, Randy Bennett, Edmund Gordon, Edward Haertel, Kenji Hakuta, Andrew Ho, Robert Lee Linn, P. David Pearson, James Popham, Lauren Resnick, Alan H. Schoenfeld, Richard Shavelson, Lorrie A. Shephard, Lee Shulman, and Claude M. Steele, *Criteria for High-Quality Assessment*. Stanford, Calif.: Stanford Center for Opportunity Policy in Education, 2013. As of February 10, 2014: http://edpolicy.stanford.edu/publications/pubs/847

Faxon-Mills, Susannah, Laura S. Hamilton, Mollie Rudnick, and Brian M. Stecher, *New Assessments, Better Instruction? Designing Assessment Systems to Promote Instruction Improvement*, Santa Monica, Calif.: RAND Corporation, RR-354-WFHF, 2013. As of February 10, 2014: http://www.rand.org/pubs/research_reports/RR354.html

Hamilton, Laura S., Brian M. Stecher, and Kun Yuan, *Standards-Based Reform in the United States: History, Research, and Future Directions,* Washington, D.C.: Center on Education Policy, 2009. As of February 10, 2014: http://www.rand.org/pubs/reprints/RP1384.html

Herman, Joan L., "The Effects of Testing on Instruction," in Susan H. Fuhrman and Richard F. Elmore, eds., *Redesigning Accountability Systems for Education*, New York: Teachers College Press, 2004, pp. 141–166.

Herman, Joan L., and Robert Linn, *On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia,* Los Angeles, Calif.: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Report No. 823, 2013. As of February 10, 2014: http://www.cse.ucla.edu/products/reports/R823.pdf

Hess, Karin K., Dennis Carlock, Ben Jones, and John R. Walkup, *What Exactly Do "Fewer, Clearer, and Higher Standards" Really Look Like in the Classroom? Using a Cognitive Rigor Matrix to Analyze Curriculum, Plan Lessons, and Implement Assessments*, Dover,

N.H.: National Center for the Improvement of Educational Assessment, 2009. As of February 10, 2014: http://www.nciea.org/beta-site/publication_PDFs/cognitiverigorpaper_KH11.pdf

International Baccalaureate, "IB Fast Facts," web page, undated(a). As of February 10, 2014: http://www.ibo.org/facts/fastfacts/index.cfm

———, "IB World School Statistics," web page, undated(b). As of February 10, 2014: http://www.ibo.org/facts/schoolstats/progsbycountry.cfm

Kelly, Dana L., Senior Research Scientist, National Center for Education Statistics, Assessment Division, International Assessment Branch, personal communication, September 21, 2012.

Koretz, Daniel M., and Laura S. Hamilton, "Testing for Accountability in K–12," in Robert L. Brennan, ed., *Educational Measurement,* 4th ed. Westport, Conn.: American Council on Education/Praeger, 2006.

Kyllonen, Patrick C., and Susanne P. Lajoie, "Reassessing Aptitude: Introduction to a Special Issue in Honor of Richard E. Snow," *Educational Psychologist*, Vol. 38, No. 2, 2003, pp. 79–83.

Martin, Michael, and Ina Mullis, Executive Directors, TIMSS and PIRLS International Study Center, personal communication, September 19, 2012.

Matsumura, Lindsay Clare, Sharon Cadman Slater, Mikyung Kim Wolf, Amy Crosson, Allison Levison, Maureen Peterson, Lauren Resnick, and Brian Junker, *Using the Instructional Quality Assessment Toolkit to Investigate the Quality of Reading Comprehension Assignments and Student Work,* Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, Report No. 669, 2006. As of February 10, 2014: https://www.cse.ucla.edu/products/reports/r669.pdf

Measured Progress and ETS Collaborative, *Smarter Balanced Assessment Consortium General Item Specifications*, Dover, N.H., April 16, 2012. As of February 10, 2014: http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf

Mullis, Ina V. S., Michael O. Martin, Ann M. Kennedy, and Pierre Foy, "Appendix D: Sample Passages, Questions, and Scoring Guides," *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries*, Boston, Mass.: TIMSS and PIRLS International Study Center, 2007. As of February 18, 2014: http://timssandpirls.bc.edu/pirls2006/intl_rpt.html

Mullis, Ina V. S., Michael O. Martin, Graham J. Ruddock, Christine Y. O'Sullivan, and Corinna Preuschoff, *TIMSS 2011 Assessment Frameworks*, Boston, Mass.: International Association

for the Evaluation of Educational Achievement, 2009. As of February 10, 2014: http://timssandpirls.bc.edu/timss2011/frameworks.html

National Assessment Governing Board, *Mathematics Framework for the 2011 National Assessment of Educational Progress*, Washington, D.C.: U.S. Department of Education, September 2010a. As of February 20, 2014: http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/math-2011-framework.pdf

———, *Reading Framework for the 2011 National Assessment of Educational Progress*, Washington, D.C.: U.S. Department of Education, September 2010b. As of February 10, 2014: http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading-2011-framework.pdf

———, *Writing Framework for the 2011 National Assessment of Educational Progress*, Washington, D.C.: U.S. Department of Education, September 2010c. As of February 10, 2014: http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/writing-2011.pdf

National Center for Education Statistics, 2007 NAEP sample reading questions, 2007. As of February 18, 2014: http://nces.ed.gov/nationsreportcard/pdf/demo_booklet/SQ-07-grade08_2.pdf

———, NAEP Questions Tool: Writing, online database, last updated 2011. As of February 18, 2014: http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=writing

———, NAEP Questions Tool: Mathematics, online database, last updated 2013. As of February 18, 2014: http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=mathematics

National Research Council, *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools*, Washington, D.C.: National Academies Press, 2002.

———, *Assessing 21st Century Skills: Summary of a Workshop*, Washington, D.C.: National Academies Press, 2011.

———, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, Washington, D.C.: National Academies Press, 2012.

NCES—*See* National Center for Education Statistics.

Newmann, Fred M., Gudelia Lopez, and Anthony S. Bryk, *The Quality of Intellectual Work in Chicago Schools: A Baseline Report*, Chicago, Ill.: Consortium on Chicago School Research, 1998.

Nohara, David, *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)*, Washington, D.C.: National Center for Education Statistics, U.S. Department of Education,, NCES 2001-07, June 2001.

NRC—*See* National Research Council.

Rothman, Robert A., *Imperfect Matches: The Alignment of Standards and Tests*, Washington, D.C.: National Research Council, 2003.

OECD—*See* Organization for Economic Co-operation and Development.

Organization for Economic Co-operation and Development, *PISA Released Items—Reading*, December 2006. As of February 18, 2014: http://www.oecd.org/pisa/38709396.pdf

———, *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*, Paris, France, 2009. As of February 10, 2014: http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/pisa2009assessmentframework-keycompetenciesinreadingmathematicsandscience.htm

PARCC—*See* Partnership for Assessment of Readiness for College and Careers.

Partnership for Assessment of Readiness for College and Careers, *Proposed Sources of Cognitive Complexity in PARCC Items and Tasks: ELA/Literacy*, Washington, D.C., 2012a.

———, *Proposed Sources of Cognitive Complexity in PARCC Items and Tasks: Mathematics*, Washington, D.C., 2012b.

Polikoff, Morgan S., Andrew C. Porter, and John Smithson, "How Well Aligned Are State Assessments of Student Achievement with State Content Standards? *American Educational Research Journal,* Vol. 48, No. 4, August 2011, pp. 965–995.

Porter, Andrew C., "Measuring the Content of Instruction: Uses in Research and Practice," *Educational Researcher,* Vol. 31, No. 7, October 2002, pp. 3–14.

Provasnik, Stephen, Patrick Gonzales, and David Miller, *U.S. Performance Across International Assessments of Student Achievement: Special Supplement to the Condition of Education 2009,* Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, NCES 2009-083, August 2009.

Schneider, Michael, and Elsbeth Stern, "The Cognitive Perspective on Learning: Ten Cornerstone Findings," in Hanna Dumont, David Istance, and Francisco Benavides, eds.,

The Nature of Learning: Using Research to Inspire Practice, Paris: Organization for Economic Co-operation and Development, 2010, pp. 69–90.

Stephens, M., and M. Coleman, *Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science,* working paper, Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, 2007. As of February 10, 2014: http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf

Soland, Jim, Laura S. Hamilton, and Brian M. Stecher, *Measuring 21st-Century Competencies: Guidance for Educators*, Santa Monica, Calif.: RAND Corporation, WR-1021, 2013. As of February 10, 2014: http://www.rand.org/pubs/working_papers/WR1021.html

Thompson, Sheila, Stephen Provasnik, David Kastberg, David Ferraro, Nita Lemanski, Stephen Roey, and Frank Jenkins, *Highlights from PIRLS 2011: Reading Achievement of U.S. Fourth-Grade Students in an International Context*, Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, NCES 2013-010, December 2012. As of February 10, 2014: http://nces.ed.gov/pubs2013/2013010rev.pdf

Webb, Norman L., *Alignment of Science and Mathematics Standards and Assessments in Four States,* Washington, D.C.: Council of Chief State School Officers, August 1999.

———, *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States*, Washington, D.C.: Council of Chief State School Officers, 2002a.

———, Depth-of-Knowledge Levels for Four Content Areas, Madison, Wis.: Wisconsin Center for Education Research, University of Wisconsin, March 28, 2002b. As of February 10, 2014: http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc

———, "Issues Related to Judging the Alignment of Curriculum Standards and Assessments," *Applied Measurement in Education*, Vol. 20, No. 1, 2007, pp. 7–25.

Yuan, Kun, and Vi-Nhuan Le, *Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items Through the State Achievement Tests,* Santa Monica, Calif.: RAND Corporation, WR-967-WFHF, 2012. As of February 10, 2014: http://www.rand.org/pubs/working_papers/WR967.html