Skip all front matter: Jump to Page 1 ▼

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

RAND
CORPORATION

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.

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

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

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

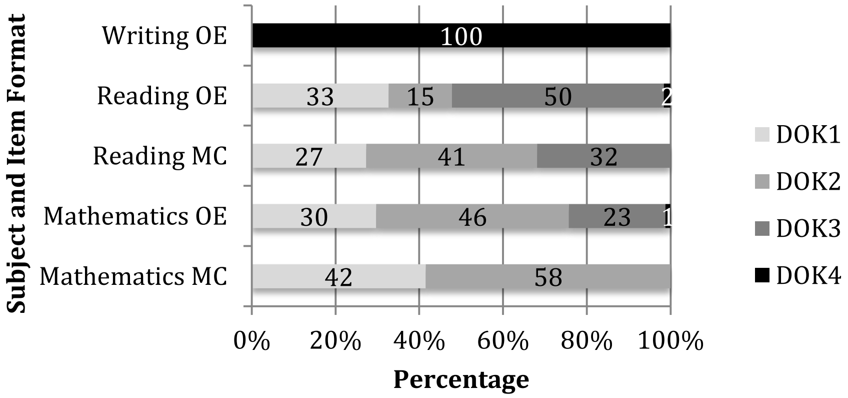
The Six Benchmark Tests Had Greater Cognitive Demand Than the State Tests

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.