PROJECT
MEMORANDUM

Estimating The Percentage of Students Who Were Exposed to Deeper Learning on the State Achievement Tests

KUN YUAN & VI-NHUAN LE

PM-3945-WFHF

February 2012

Prepared for the William and Flora Hewlett Foundation

RAND EDUCATION

PREFACE

The William and Flora Hewlett Foundation's Education Program initiated a new strategic initiative in 2010 that focuses on students' mastery of core academic content and their development of *deeper learning skills* (i.e., critical-thinking, problem-solving, collaboration, communication, and learn-how-to-learn skills). The Foundation would like to track the extent to which U.S. students are assessed in a way that emphasizes deeper learning skills during its 2010–2017 Deeper Learning Initiative. This report presents the results of a project to estimate the percentage of U.S. elementary and secondary students being assessed on deeper learning skills through statewide mathematics and English language arts achievement tests at the beginning of the Deeper Learning Initiative. This research has been conducted by RAND Education, a unit of the RAND Corporation.

This product is part of the RAND Corporation project memorandum series. Project memoranda are informal communications between members of RAND project teams and their sponsors. They have not been formally reviewed or edited. Project memoranda should not be cited, quoted, reproduced, or transmitted without RAND's permission.

CONTENTS

FIGURES

TABLES

SUMMARY

In 2010, the William and Flora Hewlett Foundation's Education Program initiated its strategic Deeper Learning Initiative that focuses on students' mastery of core academic content and their development of deeper learning skills (i.e., critical-thinking, problem-solving, collaboration, communication, and learn-how-to-learn skills). One of the goals of the Deeper Learning Initiative is to improve the proportion of U.S. elementary and secondary students nationwide being assessed on deeper learning skills to 15 percent by 2017. The Foundation asked RAND to conduct a study to examine the percentage of U.S. elementary and secondary students being assessed on deeper learning skills at the beginning of the Deeper Learning Initiative.

ABOUT THE STUDY

Selection of State Mathematics and English Language Arts Tests in 17 States

To estimate the percentage of U.S. elementary and secondary students assessed on deeper learning, we had to identify measures of student learning to be included in the analysis. Moreover, we needed access to information about test items and the number of test takers for each measure. We started by searching for tests for which these two types of information were publicly available.

We conducted a literature review and an online information search, consulted educational assessment experts, and considered a variety of tests to be included in the analysis, such as statewide achievement tests, Advanced Placement (AP) tests, International Baccalaureate (IB) exams, and benchmark tests. Among the tests we considered, information about both the test items and the number of test takers were available only for state achievement tests. Therefore, state achievement tests were the only type of student measures that we could include in this project.

Given the available project resources, it was not feasible to analyze the state achievement tests for all states, so we had to prioritize by focusing on a group of states whose achievement tests had higher probabilities of assessing deeper learning than those used in other states. We conducted a literature review on the design, format, and rigor of statewide achievement assessments. Prior literature suggested 17 states whose state achievement tests were more cognitively demanding and might have a higher probability of assessing deeper learning. Because statewide mathematics and English language arts tests are administered to students in grades 3–8 and in one high school grade level in most states, our analyses of the items focused on mathematics and English language arts tests at these grade levels in these 17 states.

Using Webb's Depth-of-Knowledge Framework to Analyze the Cognitive Processes of Deeper Learning Skills

The manner in which students are assessed on the state exams restricted our analysis to three types of deeper learning skills: the mastery of core academic content, critical-thinking and problem-solving skills, and written communication skills. To determine the extent to which each state test measures these deeper learning skills, we reviewed multiple frameworks that had been used to describe the cognitive processes of test items and learning tasks.

The frameworks we reviewed included Norman Webb's (2002a) four-level Depth-of-Knowledge (DOK) framework; Andrew Porter's (2002) five-level cognitive rigor 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 her colleagues' (2006) instructional quality assessment toolkit to measure the quality of instruction and the cognitive demand of student assignments.

Although these frameworks differed in their structure and purpose, they all focused on describing the cognitive rigor elicited by the task at hand. Therefore, we decided to assess whether a state test met the criteria for a deeper learning assessment based on the cognitive rigor of the test items. Among the five frameworks we reviewed, Webb's DOK framework is the most widely used to assess the cognitive rigor of state achievement tests and best suited the needs of this project. Therefore, we adopted Webb's DOK framework to analyze the cognitive rigor demanded of state tests.

Webb defined four levels of cognitive rigor, where level 1 represented recall, level 2 represented demonstration of skill/concept, level 3 represented strategic thinking, and level 4 represented extended thinking. We applied Webb's subject-specific descriptions for each of the DOK levels for mathematics, reading, and writing in our analysis. Our review of the DOK framework suggests that the cognitive demands associated with DOK level 4 most closely match the Deeper Learning Initiative's notion of deeper learning, so we use DOK level 4 as our indicator that a test item measures deeper learning.

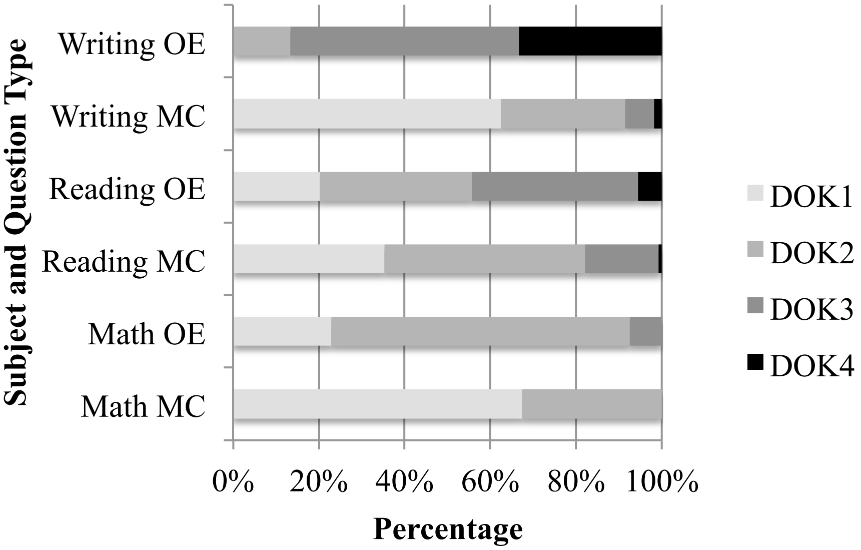
FINDINGS

The Overall Rigor of State Mathematics and English Language Arts Tests in 17 States Was Low, Especially for Mathematics

For each state test, we applied Webb's DOK framework to analyze the cognitive rigor of individual test items and summarized the percentage of items that met the criteria for each DOK level. Two researchers and two subject experts rated the cognitive rigor of more than 5,100 released state test items using Webb's DOK framework, with two raters per subject. The inter-rater reliability was high (above 0.90) for both subjects.

In general, the cognitive rigor of state mathematics and English language arts tests was low. Most items were at DOK level 1 or 2. Open-ended (OE) items had a greater likelihood of reaching DOK level 3 or 4 than did multiple-choice (MC) items. Figure S.1 shows the average percentage of test items at each DOK level by subject and item format.

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



MC and OE items had different likelihoods of being rated at DOK level 4, so we set two different criteria for a test to be considered as a deeper learning assessment that took into account the question format. Criterion A was more strict; it required that 5 percent of MC items were rated at DOK level 4 *and* at least one OE item was rated at DOK level 4. Criterion B was less strict; it required that 5 percent of MC items were rated at DOK level 4 *or* at least one OE item was rated at DOK level 4. We chose 5 percent as the cutoff level for MC items because it is the mean (and median) percentage of reading items that were rated at DOK level 4 on state reading tests across the 17 states..

We judged each test separately on the two criteria, giving us a range of results depending on how strictly deeper learning assessment was defined. None of the state mathematics tests we analyzed met the criteria for a deeper learning assessment using either criterion. Depending on the criterion we used, between 1 and 20 percent of the state reading tests and 28–31 percent of the state writing tests we analyzed qualified as deeper learning assessments.

Only 3–10 Percent of U.S. Elementary and Secondary Students Were Assessed on Deeper Learning Skills Through State Mathematics and English Language Arts Tests

Using our DOK coding results and 2009–2010 student enrollment data from the National Center for Educational Statistics, we estimated the percentage of U.S. elementary and secondary students assessed on deeper learning skills in mathematics, reading and writing, under the assumption that none of the tests in the other states not analyzed in this study measure deeper learning. We found that 0 percent of students in the U.S. were assessed on deeper learning in mathematics through state tests, 1–6 percent of students were assessed on deeper learning in reading through state tests, and 2–3 percent of students were assessed on deeper learning in writing through state tests. Overall, 3–10 percent of U.S. elementary and secondary students were assessed on deeper learning on at least one state assessment.

We also estimated the percentage of students assessed on deeper learning based on different cutoff scores for MC items. Results showed that when a cutoff percentage for MC items of 4 percent or higher was adopted, the final estimation of U.S. elementary and secondary students assessed on deeper learning through the state mathematics and English language arts tests stays approximately the same.

Interpreting The Results

There are several caveats worth noting when interpreting the results of this analysis. First, a lack of information about the test items and the number of test takers for other types of tests, such as AP, IB, and benchmark tests, prevented us from examining the extent to which these tests measure deeper learning skills. This constraint likely means that our findings underestimate the percentage of students assessed on deeper learning skills in our sample of states.

Second, the content and format of state achievement tests did not allow us to analyze collaboration, oral communication, or learn-how-to-learn skills. Although omitting these three deeper learning skills might have caused us to overestimate the percentage of state tests that meet the criteria for deeper learning assessments, doing so allowed us to conduct meaningful analysis of the extent to which the current state tests measure other important aspects of deeper learning.

Third, given the available project resources, we had to prioritize by focusing on 17 states' tests identified by prior studies as more rigorous than those used in the other two-thirds of U.S. states. We assumed that the results about the rigor of the 17 state tests published in prior reviews were accurate and that the tests' level of rigor had not changed substantially since those reviews were conducted. We also assumed that none of the tests used in the other two-thirds of states would meet the criteria for deeper learning assessments.

Fourth, the determination of whether a state test met the criteria for a deeper learning assessment might be biased because the full test form was not available in some states and the unreleased items might be different than the released items in the extent to which they measure deeper learning skills. However, the issue of partial test forms is unavoidable. There are a number of reasons states do not release full test forms and we could only work with the items they did release.

Fifth, we assessed whether a state test met the criteria for a deeper learning assessment based on the percentage or number of test items rated at the highest DOK level. We also considered using the portion of the total test score that is accounted for by DOK level 4 items to represent the cognitive rigor of a state test. However, we could not use this measure because some states did not provide the number of score points for released items or the total score of a state test.

Sixth, the choice of the cutoff percentage of MC items rated at DOK level 4 is admittedly arbitrary. Our analysis of different cutoff scores showed that raising or lowering the cutoff by one or two percent did not substantially change the estimate of the percentage of U.S. elementary and secondary students assessed on deeper learning through state tests.

ACKNOWLEDGMENTS

This report could not have been completed without assistance from the following individuals. We would like to thank Dr. Joan Herman for her insightful comments on the analysis plan and choice of analytical framework. We thank Susanna Berko and Justin Boyle for their help in analyzing the rigor of released state test items. We are also grateful to the sponsor of this project, the William and Flora Hewlett Foundation, and the support and feedback we received from the program staff.

Several RAND colleagues also contributed to the completion of this report. We thank Cathleen Stasz, Kate Giglio, and Lauren Skrabala for their thoughtful reviews and comments. We would also like to thank Sarah Hauer and Sharon Koga for their administrative assistance during the completion of the report.

ABBREVIATIONS

AP	Advanced Placement
CAE	Council for Aid to Education
CWRA	College and Work Readiness Assessment
DOK	Depth-of-Knowledge
IB	International Baccalaureate
MC	multiple-choice
NECAP	New England Common Assessment Program
OE	open-ended

1. INTRODUCTION

THE DEEPER LEARNING INITIATIVE

In 2010, the William and Flora Hewlett Foundation's Education Program implemented the 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 a set of 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 includes the ability to organize people, knowledge, and resources toward 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 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 work and learning.

As part of its efforts to promote deeper learning, the Foundation also launched a series of research activities. One aspect of the Foundation's efforts to increase students' exposure to deeper learning is the development of seven model school networks that embody the deeper learning approach (see Yuan and Le, 2010, for more details on these model school networks). Although the networks vary with respect to organizational structure, training and supports, and student populations served, the networks share commonalities with respect to their design principles for promoting deeper learning. Namely all the networks emphasize small learning communities, personalized learning opportunities, connections to the real world, and college and work readiness as core principles of deeper learning. The networks also indicate they often used "authentic assessments," such as student portfolios, performance during workplace internships, and cross-curricular team-based projects, to evaluate student learning.

Another aspect of these research activities included estimating the percentage of elementary and secondary students nationwide being assessed on deeper learning at the outset of the Deeper Learning Initiative. Although the Foundation currently lacks an estimate of the percentage of students being tested on deeper learning, it believes that the percentage is likely to be very low. Thus, one of its goals is to increase the nationwide percentage of students who are assessed on deeper learning skills to at least 15 percent by 2017.

PURPOSE OF THIS STUDY AND STRUCTURE OF THIS REPORT

The purpose of this project was to estimate the percentage of students being tested on deeper learning skills at the beginning of the Deeper Learning Initiative. These estimates are intended to serve as baseline measures to gauge the progress of the initiative toward its 15-percent benchmark in five years.

We conducted this project in four steps. Because the types of measures that could feasibly be included in our project restricted the types of deeper learning skills we could analyze, we started this project by searching for tests that we could analyze for this study. Chapter Two provides detailed descriptions of the tests that were included in our analysis, focusing on the process, criteria, and rationale for the selection of these measures.

After finalizing the choice of tests to be included in the analysis, we identified a framework to analyze the extent to which state tests measure deeper learning. In Chapter Three, we describe the types of deeper learning skills we analyzed and how we chose the framework.

With the tests to be included and the analytical framework chosen, we applied the analytical framework to the selected tests to assess the extent to which they measured deeper learning skills. Chapter Four presents our results about the degree to which selected tests measures deeper learning skills and how we assessed whether they qualified as deeper learning assessments.

Our final step was to estimate the percentage of U.S. elementary and secondary students assessed on deeper learning skills based on whether selected tests qualified as deeper learning assessments and on the number of students assessed by each test at the beginning of the Deeper Learning Initiative. The Chapter Five presents our estimation results and discusses the caveats and limitations of this project.

The report also includes two appendixes that provide sample test items correlated with each level in our analytical framework and present detailed findings by state, grade level, subject, and classification, respectively.

2. Identifying Tests for Analysis

Educators use a variety of assessment tools to measure student learning, such as homework, formative assessments, interim or benchmark tests, and state achievement tests. These tests vary substantially in many ways—for example, in their purposes, formats, and frequency and breadth of administration. Moreover, the use of these tests varies by school, district, and state. Not all measures used to assess student learning are analyzable in a project such as this, so our first step was to identify tests for analysis. In this chapter, we describe the process and criteria for choosing the tests to be included in our study.

Ideally, the measures of student learning outcomes that we analyzed would reflect the core principles set forth by the model school networks. However, as we conducted our search for the measures that would be available for an analysis of deeper learning, it became apparent that the types of assessments that were favored by the networks would not be analyzable on a large-scale basis. For example, previous studies that have attempted to examine student portfolios statewide have reported prohibitive costs (Koretz et al., 1994) and unreliability in scoring (Reckase, 1995). Because project resources did not allow us to analyze the types of "authentic assessments" that would likely capture all aspects of deeper learning on a large-scale basis, our analysis is restricted to the types of skills and knowledge that were likely to be assessed by the tests administered on a large-scale.

Specifically, to estimate the percentage of students assessed on deeper learning skills, we needed to have access to information about test items and the number of test takers. This information allowed us to determine the extent to which a test measured deeper learning skills and how many students were tested on these skills. We set out to identify student assessments for which both types of information were publicly or readily available.

Criteria for Selection

We conducted a literature review and online information search to identify tests to be included in the analysis. We considered a variety of tests that are commonly used to assess student learning in any given school year, such as statewide achievement tests, exams used by special assessment consortiums (such as the New York Performance Assessment Consortium), district-level assessments, and other tests currently used in K–

12 schools that might measure deeper learning skills, such as the Advanced Program (AP) tests, International Baccalaureate (IB) exams, and the College and Work Readiness Assessment (CWRA) developed by the Council for Aid to Education (CAE).[1]

We examined the availability of information about test items and test takers for each candidate test to determine the feasibility of including it in this study. Our analysis showed that information both about test items and the number of test takers is readily available only for state tests. Other types of candidate tests we considered lacked one or both types of information and thus could not be included (see Table 2.1).

For example, we considered AP tests and IB exams, for which the test content was readily available. However, information about the number of test takers in a given year is not publically available. Although each school district may collect student-level information regarding which student took the AP or IB exams, it was beyond the scope of this project to collect such information from all school districts nationwide.

Benchmark tests (also referred to as interim assessments) are used by many school districts to monitor student performance on a monthly or quarterly basis to predict students' performance on the state achievement tests (Goertz, Olah, and Riggan, 2009). The format and rigor of test items is usually similar to that of the state achievement tests, so students' performance on the benchmark tests provides useful information about students' possible test scores on the state tests. School districts make local decisions regarding whether to administer benchmark tests to monitor student performance, and the specific grades, years, and subject areas in which these tests are administered can vary by district or even by school. This variability made it impossible for us to identify school districts nationwide in which benchmark tests were used at the beginning of the Deeper Learning Initiative. Moreover, for proprietary reasons, commercial test developers do not provide public access to benchmark tests items. Without access to both the benchmark test items and the number of students who took these tests, we could not include benchmark assessments in the analysis.

Performance assessments were another type of test that we considered. Such tests measure student learning through constructed-response tasks (Stecher, 2010), while state

[1] We did not consider national or international tests (such as the National Assessment of Educational Progress and the Program for International Student Assessment) because these tests are administered every three or four years, not in one particular school year. Including these tests in this type of analysis might cause the estimate of U.S. elementary and secondary students assessed on deeper learning skills to fluctuate according to the year in which these tests are administered.

achievement tests and benchmark tests assess student learning mainly through multiple-choice items. Schools or districts that promote the use of project-based learning may use performance assessment to measure student learning (Yuan and Le, 2010). Schools affiliated with certain assessment consortiums (such as the New York Performance Standards Consortium) also use performance assessment to monitor student learning. However, the design and use of performance assessments is also a local decision within each classroom or at each school, and the resources needed to identify the population of students assessed using performance assessments was beyond the scope of this study.

The CWRA is an online tool developed by CAE to assess critical-thinking, analytical reasoning, problem-solving, and written communication skills (Stecher, 2010). It presents students with a realistic problem that requires them to analyze and synthesize information in multiple documents and write short essays to defend their arguments. It has been administered school-wide in a few states and is considered by the administering schools to be a good measure of deeper learning skills (Yuan and Le, 2010). Sample items are available from CAE's website, but the number of test takers is not. Although CAE does collect information about the number of CWRA test takers, this information is protected under its confidentiality agreement with participating schools. Thus, we could not include the CWRA in our analysis.

Table 2.1. Comparisons of Candidate Tests Considered for This Project

Test	Test Taker Population	Are the test items publically available?	Is the number of test takers in a given year publically available?
State tests	All students in tested grades in a state, usually students in grades 3–8 and one high school grade level	Yes	Yes
AP tests	Any high school student interested in taking AP exams	Yes	No
IB exams	High school students who are enrolled in IB programs	Yes	No
Benchmark tests	Depending on districts' or schools' choice	No	No
Performance assessment	Depending on districts' or schools' choice	No	No
CWRA	High school freshmen and/or seniors in schools interested in the CWRA	Yes	No

To summarize, due to the lack of information about test items and/or test takers for AP or IB exams, benchmark and performance assessments, and the CWRA, the study had to use statewide achievement exams to analyze the percentage of U.S. elementary and secondary students tested on deeper learning skills. The content of the tests and the number of students taking them were readily available for state tests. Specifically, many state departments of education release some or all of the items on the state achievement tests by grade and subject to the public through their Web sites. In addition, the number of tested students at each grade level in a given year is available from the Common Core of Data at the National Center for Education Statistics.

SAMPLE OF STATES INCLUDED IN THE ANALYSIS

Given the available project resources, it was impossible to examine whether the achievement tests in all 50 states and the District of Columbia assessed deeper learning skills. Through discussions with the Foundation, we decided to focus on a group of states whose state achievement tests might have a higher probability of assessing deeper learning skills than those of other states.

We conducted a literature review on the design, format, and rigor of statewide achievement assessments and identified 17 states to be included in the analysis: California, Colorado, Connecticut, Delaware, Kentucky, Massachusetts, Maryland, Missouri, New Jersey, New York, Ohio, Texas, Washington, and four states that use the New England Common Assessment Program (NECAP)—Maine, New Hampshire, Rhode Island, and Vermont. We chose these 17 states based on previous literature review results showing that their assessments are more cognitively demanding than those of other states (Great Lakes West Comprehensive Center, 2009; Darling-Hammond and Adamson, 2010; Stecher, 2010) and might have a higher probability of assessing deeper learning skills. Because statewide mathematics and English language arts tests are administered to students in grades 3–8 and in one high school grade level in most states, our analyses focused on these achievement tests in these two subjects at the tested grade levels in 17 states.

LIMITATIONS OF DATA ON STATE ACHIEVEMENT TESTS

It is important to note that the information provided about statewide achievement exams varies substantially across states, and in ways that make it difficult to assess the cognitive rigor of the tests. First, states varied in the form of the tests they released. Most states (82 percent of the 17 states) released partial test forms instead of complete test booklets. For these states, we proceeded on the assumption that the percentage of items that measure deeper learning on the partial form was the same as the percentage of deeper learning items in the full test booklet. To the extent that the released items are not representative of the complete form, our final estimate may be inaccurate.

Second, states varied in terms of the subject and grade levels for which they released sample test items. For mathematics, the majority of states (82 percent of the 17 states) released tests for each grade at the elementary and middle school levels and for one or two grades at the high school level. The remaining states provided tests at selected grade levels only. For English language arts tests, one-third of the 17 states administered a reading test that may or may not contain a writing portion. The remaining three-quarters of the 17 states we analyzed administered a reading test to each tested grade level and a writing test to selected or all tested grade levels.

Third, states varied in the number of released test items by grade and subject. Among the 17 states that we examined, the average number of released items per grade, per subject, and per state was 30 for mathematics (*S.D.* = 26, Min = 2, Max = 96), 26 for reading (*S.D.* = 29, Min = 2, Max = 114), and seven for writing (*S.D.* = 10, Min = 1, Max = 40).

Finally, states varied in terms of the year in which those released test items were actually administered. For instance, Massachusetts released test items used in the 2010–2011 academic year, while released mathematics test items from Kentucky dated back to 1998–1999. Overall, 16 states released test items that were used within three years prior to the 2010–2011 school year, the start of the Deeper Learning Initiative.

These features of released state test items made it difficult to obtain an accurate baseline estimate of the percentage of students assessed on deeper learning skills at the start of the Deeper Learning Initiative.

3. Choosing a Framework to Analyze the Cognitive Rigor of State Tests

After identifying tests to be analyzed in this study, our next step was to choose a framework for assessing the extent to which selected state tests measured deeper learning skills. In this chapter, we describe which types of deeper learning skills state tests allowed us to analyze and how we chose the framework to assess the extent to which these types of deeper learning skills were assessed by state tests.

Analysis of Three Types of Deeper Learning Skills

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. This set of "deeper learning" skills defines the target constructs that an ideal deeper learning assessment should measure.

However, as noted in Chapters One and Two, the types of tests that could feasibly be included in our study meant that we could measure only a limited set of deeper learning skills. The first deeper learning skill that we omitted was working collaboratively. Given that the state exams are intended to assess students' knowledge, independent of the input from other students, it was not possible to assess student collaboration in the context of state achievement tests. Similarly, we omitted learning how to learn from our targeted constructs. Learning how to learn is an aspect of metacognition and is usually measured through think-aloud cognitive interviews or questionnaires (Zimmerman and Martinez-Pons, 1990; Le et al., 2005). However, concerns about costs and social desirability make such methods impractical for state testing programs. Thus, learning how to learn could not be included as a deeper learning skill in our analysis. Finally, we were able to measure only limited aspects of effective communication. Effective communication requires not only written skills but also oral skills. None of the state tests in our analysis tested oral communication skills, so the effective communication construct focused only on written communication skills.

These limitations of the state tests meant that only three deeper learning skills were likely to be assessed on these tests: (1) mastery of core academic content, (2) critical-thinking and problem-solving skills, and (3) written communication skills.

SELECTION OF FRAMEWORK TO ANALYZE THE COGNITIVE RIGOR OF STATE TESTS

After restricting the deeper learning construct to these three specific skills, we reviewed multiple frameworks of educational objectives, cognitive processes, and learning standards to identify one framework that we could use to determine whether tests assessed these skills. In total, we considered five frameworks that could be used to describe mental processes that reflect deeper learning skills: Norman Webb's (2002a) four-level Depth-of-Knowledge (DOK) framework; Andrew Porter's (2002) five-level cognitive rigor 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 her colleagues' (2006) instructional quality assessment toolkit to measure the quality of instruction and the cognitive demand of student assignments.

Although these frameworks differed in their structure and purpose, they all focused on describing the cognitive rigor or cognitive complexity elicited by the task at hand. Typical descriptions of lower-level cognitive processes included "recalling/memorizing" or "performing routine procedures." In contrast, typical descriptions of higher-level cognitive processes included "analyzing and synthesizing information from multiple sources" or "applying concepts to novel contexts or problems." In reviewing the frameworks, it became apparent that the types of mental processes that reflected a mastery of core academic content, critical-thinking and problem-solving skills, and effective written communication skills were not necessarily distinguishable from one another. However, a common feature of these deeper learning skills is that their demonstration would require tasks that demanded a high degree of cognitive complexity or rigor. Thus, we decided to assess whether a state test met the criteria for a deeper learning assessment based on the cognitive rigor of its test items.

To select a framework we first considered its purpose and ease of use. The standards developed by Newmann et al. (1998) and the instructional assessment toolkit developed by Matsumura and her colleagues (2006) are more commonly used to examine the cognitive demand of classroom instruction, assignments, and student work, as opposed to the cognitive demand of test items (Mitchell, et al., 2005; Matsumura et al., 2008). Therefore, we eliminated these frameworks from further consideration.

Of the three remaining frameworks used to examine the cognitive rigor of tests (i.e., Webb's DOK framework, Bloom's Taxonomy of Educational Objectives, and Porter's cognitive rigor model), Webb's framework is the most widely used to examine the cognitive demand of state tests (Rothman, 2003; Webb, 1999, 2002a, 2002b, 2007). According to Webb's framework, the cognitive rigor of a test item is evaluated based on both the complexity of the required cognitive tasks and the content to be analyzed. This approach is similar to Porter's approach, but we ultimately favored Webb's model because we found it more relevant and easier to implement[2].

We also considered Bloom's revised Taxonomy of Educational Objectives (i.e., remember, understand, apply, analyze, evaluate, and create) as a potential framework. However, Bloom's "understand" category is ambiguous because it can represent either lower- or higher-level cognitive processes (Hess et al., 2009). Furthermore, when mapped onto Webb's DOK framework, Bloom's categories cut across the four levels, such that an item falling into the "understand" category could be rated at DOK level 1 or DOK level 4, depending on the cognitive demands of the task (Hess et al., 2009). Because of this ambiguity, we rejected Bloom's framework.

WEBB'S DEPTH-OF-KNOWLEDGE FRAMEWORK

We adopted Webb's DOK framework to analyze the cognitive rigor of state test items and applied the subject-specific descriptions for each of the DOK levels for mathematics, reading, and writing in our analysis[3]. Webb defined four levels of cognitive rigor, where level 1 represented recall, level 2 represented demonstration of skill/concept, level 3 represented strategic thinking, and level 4 represented extended thinking (Webb, 2002b). He also provided subject-specific descriptions for each of the DOK levels, as follows:

- Mathematics
    - DOK1: Recall of a fact, term, concept, or procedure.

[2] Specifically, Porter's model combined the cognitive rigor descriptions for reading and writing into a single language arts dimension, whereas Webb's framework provided separate descriptions for reading and writing. Because many states administer separate tests for reading and writing, the Webb framework was more directly applicable to the state tests we were analyzing.

[3] Webb's framework has been widely used to examine the cognitive rigor of state tests in prior studies, the main purpose of which was to examine the alignment between state standards and achievement tests. The studies did not provide detailed results about the state tests' cognitive demand levels. Thus, we cannot compare the results of previous studies with the results of our current study.

  - DOK2: Use information, conceptual knowledge, and procedures in two or more steps.
  - DOK3: Requires reasoning, 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 non-routine 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 with 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 (Webb, 2002b).

The next chapter discusses how we applied the framework to assess the selected state tests.

4. Examining the Rigor of State Achievement Tests in Targeted States

After we identified the states and tests to include in this project and the analysis framework, the next step was to apply Webb's DOK framework to evaluate the cognitive rigor of selected state tests. In this chapter, we give an overview of our rating process, present the coding results, and describe the criteria used to determine if a test could be classified as a deeper learning assessment.

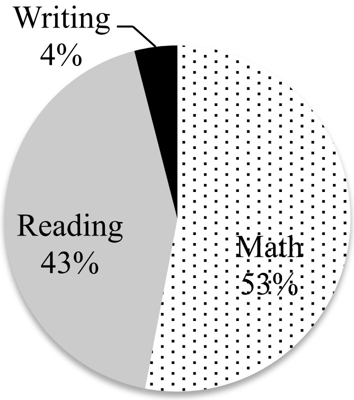
Applying Webb's DOK Framework to Our Analysis

Two subject-matter experts (one in mathematics and one in English language arts) and two members of the study team rated the cognitive rigor of the state test items using Webb's DOK framework. We trained all raters using a set of released tests (NECAP grade 3 mathematics and reading tests and a grade 5 writing test). After training, raters practiced coding using a second set of released tests (NECAP grade 4 mathematics and reading tests and a grade 8 writing test). After each of the training and practice sessions, all raters reconvened to discuss their ratings and resolved any discrepancies. Two raters for each subject then coded the NECAP grade 11 mathematics and reading tests and a grade 11 writing test to check inter-rater reliability. The weighted kappa coefficient for the English language arts raters was high, at 0.92. The two mathematics raters did not reach desirable inter-rater reliability in the first round. They reconvened and discussed the discrepancies again before conducting another round of calibration using the NECAP grade 8 mathematics test. After this second calibration, the weighted kappa coefficient for the two mathematics raters was 0.93. Given that the inter-rater reliability was high for both mathematics and English language arts, the subject-matter expects then independently analyzed the remaining test forms. (Appendix A provides a sample of items rated at each DOK level in each subject and rationales for the rating.)

Determining the Cognitive Rigor of State Mathematics and English Language Arts Test Items

In total, the research team examined more than 5,100 state test items from 201 tests. The percentage of mathematics, reading, and writing test items analyzed was 53, 43, and 4 percent, respectively (see Figure 1). State tests for all three subjects used multiple-choice (MC) items as their main format. The proportion of MC items analyzed was 78 percent for mathematics, 86 percent for reading, and 85 percent for writing, on average. The remaining items were categorized as open-ended (OE) items.

Figure 4.1. Percentage of Test Items Analyzed, by Subject

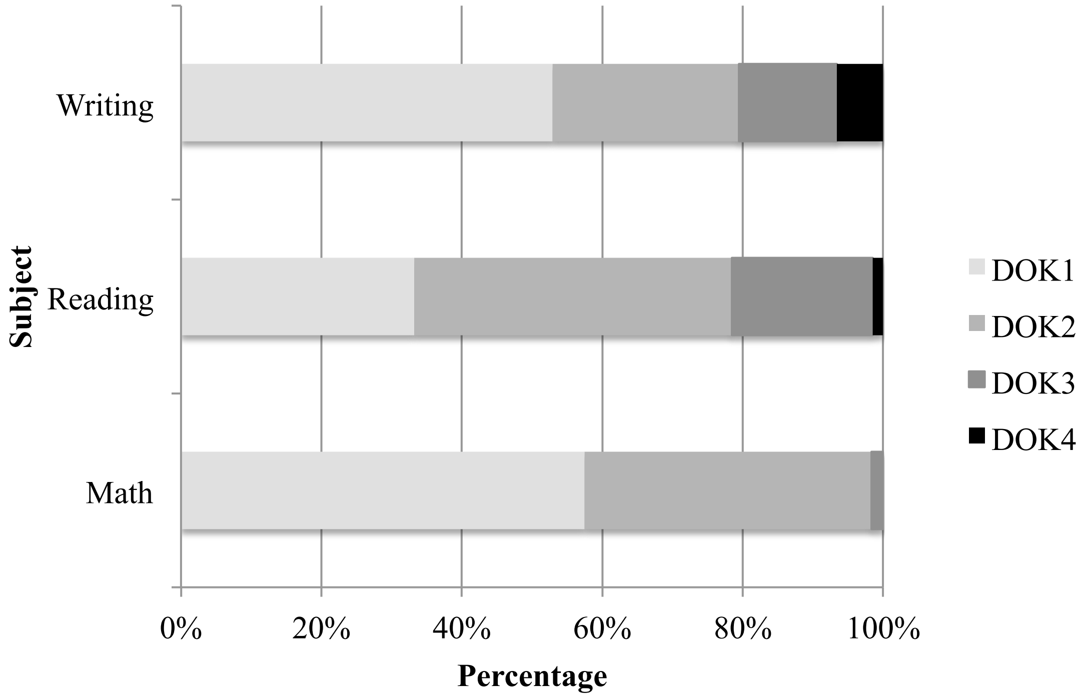


For mathematics, the cognitive rigor of all MC items was rated at or below a DOK level of 2. For most states, more than half of the MC test items were at DOK level 1. The rigor levels of OE items were at or below DOK level 3, with most OE items rated at DOK level 2. In summary, mathematics items tended to be rated at the lower cognitive levels, regardless of question format (see Figure 4.2).

For reading, all four DOK levels were represented in the MC items; however, about 80 percent were rated at or below DOK level 2. The OE items were also distributed across the four DOK levels. For most states, the majority of OE items were rated at DOK level 2 or 3. Overall, reading items tended to be rated at a higher level of cognitive rigor than mathematics items.

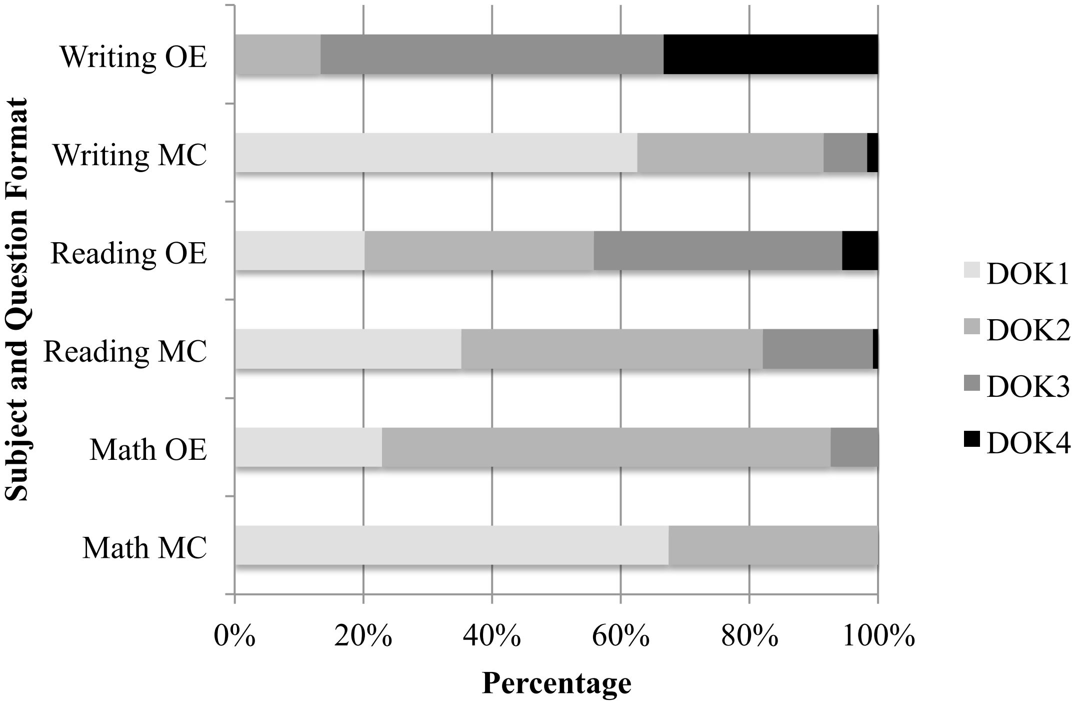
For writing, the total number of items analyzed per state was smaller than for mathematics and reading. This is mainly because most of the released writing items were essay questions. A few states also used MC items to measure writing skills, particularly editing skills. The total number of writing items for these states was larger than for other states. The distribution of the cognitive rigor of writing test items shows a similar pattern to that of the reading test items. That is, although the rigor level of the MC items ranged from DOK level 1 to 4, most of the items were rated at the lower ends of the scale (i.e., at DOK level 1 and 2).

Figure 4.2. Percentage of Test Items Rated at Each DOK Level, by Subject



The general pattern across the three subjects is that the MC items were generally rated at the lower DOK levels (1 and 2), while OE items were generally rated at higher DOK levels (3 and 4) (see Figure 4.3). Although the DOK ratings represent the rigor level of the cognitive tasks required to complete an item, the results may also be associated with the question format. For instance, MC items typically do not require students to undertake extended activities to answer a question, although the item can be quite challenging and sophisticated. For a reading or writing OE item, the format provides an opportunity for students to produce extensive compositions. Such an item should be rated at DOK level 4, according to Webb's framework. However, although the OE format may afford students the opportunity to produce extensive work, it is difficult to determine whether students actually produced work that reflected deeper learning based solely on an examination of the question prompt. Although we analyzed scoring rubrics for OE items when they were available, only six states provided such materials. Without the scoring rubrics, there is a risk of overrating the DOK levels of the OE items.

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



Tables 4.1–4.3 show the number of released state test items analyzed and the percentage of test items at each DOK level by subject, state, and question format (MC or OE items).

Table 4.1. Percentage of Released State Mathematics Test Items at Each DOK Level, by State and Question Format

	All items			MC items					OE items				
State/Test	N	% MC	% OE	N	% DOK1	% DOK2	% DOK3	% DOK4	N	% DOK1	% DOK2	% DOK3	% DOK4
California	767	100%	0%	767	74%	26%	0%	0%	0	/	/	/	/
Colorado	37	5%	95%	2	50%	50%	0%	0%	35	20%	74%	6%	0%
Connecticut	10	0%	100%	0	/	/	/	/	10	10%	30%	60%	0%
Delaware	50	32%	68%	16	63%	38%	0%	0%	34	3%	68%	29%	0%
Kentucky	36	0%	100%	0	/	/	/	/	36	8%	86%	6%	0%
Maryland	140	73%	27%	102	47%	53%	0%	0%	38	37%	61%	3%	0%
Massachusetts	207	61%	39%	126	66%	34%	0%	0%	81	26%	73%	1%	0%
Missouri	188	72%	28%	136	54%	46%	0%	0%	52	23%	71%	6%	0%
NECAP	136	52%	48%	71	68%	32%	0%	0%	65	49%	43%	8%	0%
New Jersey	129	56%	44%	72	57%	43%	0%	0%	57	19%	77%	4%	0%
New York	419	63%	37%	265	60%	40%	0%	0%	154	20%	76%	4%	0%
Ohio	144	82%	18%	118	58%	42%	0%	0%	26	4%	81%	15%	0%
Texas	438	99%	1%	432	75%	25%	0%	0%	6	100%	0%	0%	0%
Washington	49	51%	49%	25	56%	44%	0%	0%	24	8%	79%	13%	0%

NOTE*:* NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

Table 4.2. Percentage of Released State Reading Test Items at Each DOK Level, by State and Question Format

	All items			MC items					OE items				
State/Test	N	% MC	% OE	N	% DOK1	% DOK2	% DOK3	% DOK4	N	% DOK1	% DOK2	% DOK3	% DOK4
California	780	89%	11%	695	44%	41%	14%	1%	85	66%	31%	4%	0%
Colorado	14	0%	100%	0	/	/	/	/	14	14%	50%	36%	0%
Connecticut	6	100%	0%	6	83%	17%	0%	0%	0	/	/	/	/
Delaware	33	64%	36%	21	43%	43%	14%	0%	12	0%	42%	42%	17%
Kentucky	55	80%	20%	44	45%	34%	20%	0%	11	0%	27%	73%	0%
Maryland	216	88%	13%	189	20%	67%	13%	1%	27	0%	19%	74%	7%
Massachusetts	142	87%	13%	124	35%	48%	16%	1%	18	0%	17%	78%	6%
Missouri	42	36%	64%	15	33%	53%	13%	0%	27	0%	44%	56%	0%
NECAP	89	83%	17%	74	69%	20%	11%	0%	15	7%	0%	93%	0%
New Jersey	38	84%	16%	32	31%	50%	19%	0%	6	0%	33%	33%	33%
New York	221	84%	16%	186	40%	47%	12%	1%	35	0%	57%	29%	14%
Ohio	136	82%	18%	112	19%	61%	21%	0%	24	4%	71%	25%	0%
Texas	355	98%	2%	348	22%	47%	30%	1%	7	0%	0%	43%	57%
Washington	75	73%	27%	55	22%	65%	13%	0%	20	5%	35%	55%	5%

NOTE*:* NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

Table 4.3. Percentage of Released State Writing Test Items at Each DOK Level, by State and Question Format

	All items			MC items					OE items				
State/Test	N	% MC	% OE	N	% DOK1	% DOK2	% DOK3	% DOK4	N	% DOK1	% DOK2	% DOK3	% DOK4
California	2	0%	100%	0	/	/	/	/	2	0%	0%	0%	100%
Colorado	6	0%	100%	0	/	/	/	/	6	0%	50%	50%	0%
Delaware	3	0%	100%	0	/	/	/	/	3	0%	0%	100%	0%
Kentucky	6	0%	100%	0	/	/	/	/	6	0%	0%	0%	100%
Missouri	3	33%	67%	1	0%	0%	100%	0%	2	0%	0%	50%	50%
NECAP	29	100%	0%	29	59%	10%	21%	10%	0	/	/	/	/
New Jersey	6	0%	100%	0	/	/	/	/	6	0%	0%	100%	0%
Ohio	31	87%	13%	27	67%	22%	11%	0%	4	0%	25%	75%	0%
Texas	109	99%	1%	108	63%	36%	1%	0%	1	0%	0%	0%	100%

NOTE*:* States that did not administer a writing test are not included in this table. NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

SETTING THE CRITERIA FOR DEEPER LEARNING ASSESSMENTS

After we examined the cognitive rigor of the state test items, the next step was to establish the criteria to determine whether a state test qualifies as a deeper learning assessment. Our review of the DOK framework suggests that the cognitive demands associated with DOK level 4 most closely match the Deeper Learning Initiative's notion of deeper learning. Although the complexity of DOK level 3 items may appear to be close to the required rigor level of deeper learning, it does not always reach the level of rigor required by the descriptions of deeper learning. For instance, a reading item that asks students to determine the author's main intent for a reading passage can be rated at DOK level 3, according to Webb's framework. However, the reading passages included in state tests were generally straightforward, making the task of inferring the author's intent undemanding. Thus, we took a more conservative approach in deciding that level 4 should be used to denote deeper learning in reading tests.

Analogous situations were observed with the mathematics and writing items at DOK level 3. For example, a mathematics OE item that requires students to explain their thinking is categorized at DOK level 3, according to Webb's framework. However, the mathematics thinking process required to solve the mathematics problem might be relatively straightforward. For a writing item, some of the DOK level 3 items can be completed with a one-paragraph writing task. Therefore, we also used DOK level 4 as an indicator of deeper learning for these tests.

Because question format can influence the cognitive demand of an item, we adopted different criteria for the MC and OE items. For MC items, the percentage of MC items in a test rated at DOK level 4 had to reach a certain cutoff level for the test to be considered a deeper learning assessment. For OE items, a test had to have at least one OE item rated at DOK level 4 to be considered a deeper learning assessment.

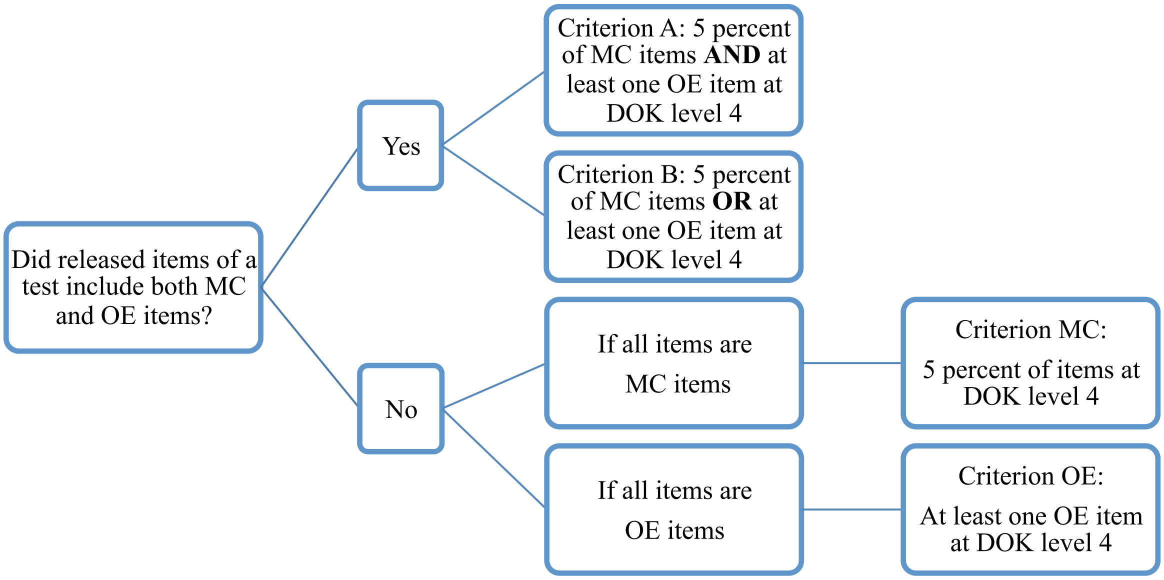
We applied the relevant criteria to the tests across all three subjects. For tests with only one type of test item, only the criterion that corresponded to that type of test items was considered. For instance, a state mathematics test with only MC items was judged using only the criterion for MC items. For a writing test with only OE items, the criterion for OE items was used to assess whether the test was a deeper learning assessment.

There are two alternative ways to use the criteria for MC and OE items to determine whether a test with both MC and OE items, as a whole, represents a deeper learning assessment. The first approach requires that a test meet both criteria to qualify as a deeper learning assessment (referred to as Criterion A). The second approach requires a test meet only one of the two criteria to be considered a deeper learning assessment (referred to as Criterion B). Criterion A is more stringent than Criterion B, so it might disadvantage the states in our study that released both MC and OE items. For example, suppose there are two states whose reading tests had the same percentage of MC items and the same number of OE items rated at DOK level 4. However, State X released both MC and OE items while State Y released only OE items. The chance that the state reading test in State X will be classified as a deeper learning assessment under Criterion A will be lower than that for the state reading test in State Y under Criterion B. To account for this type of discrepancy, we used each criteria separately to assess whether a test with both MC and OE items could be classified as a deeper learning assessment. This produced two estimates, and we report the results as a range.

To classify a state test with MC items as being indicative of a deeper learning assessment, we needed to establish a cut-score for the percentage of MC items rated at DOK level 4. Tests for which the percentage of MC items rated at DOK level 4 exceeded the cut-score would be classified as deeper learning assessments. Although there have been many studies on the alignment between state assessment standards and state tests, in which cognitive rigor is one dimension of alignment, there is no research evidence that can be drawn upon to define this cut-score (Webb, 2007).

We specified the cut-score based on the DOK level coding results for MC items. Admittedly, this approach to defining a threshold for classifying state tests is arbitrary. However, a threshold is needed in order to calculate the percentage of students assessed on deeper learning. Based on the DOK level coding results, we used 5 percent as the threshold for MC items to classify a state test as a deeper learning assessment. We chose this cutoff score mainly based on DOK level coding results for reading tests, because no mathematics tests MC items were rated at DOK level 4 and only four writing tests used MC items to measure writing skills. Furthermore, 5 percent is the mean and median of the distribution of the percentage of reading items rated at DOK level 4 in a state reading test. We also considered alternative cutoff scores, calculating the percentage of students assessed on deeper learning based on different cutoff scores (discussed in greater detail in Chapter Five).

Figure 4.4. Procedures to Determine Whether a State Test Qualified as a Deeper Learning Assessment


Did released items of a test include both MC and OE items?
Yes
No
Criterion A: 5 percent of MC items AND at least one OE item at DOK level 4
Criterion B: 5 percent of MC items OR at least one OE item at DOK level 4
If all items are MC items
If all items are OE items
Criterion MC: 5 percent of items at DOK level 4
Criterion OE: At least one OE item at DOK level 4

We assessed whether a state test for a certain grade level met the criteria for deeper learning assessment by state, grade, and subject. Figure 4.4 shows the steps used to make this determination. Among the 201 state tests we analyzed, two-thirds had both MC and OE items. Both Criterion A and B applied to these tests. Because Criterion A is more stringent than Criterion B, the percentage of tests that qualified as deeper learning assessments would be smaller under Criterion A than under Criterion B. We treated the percentage of tests that qualified as deeper learning assessments under Criterion A and Criterion B as the lower and upper bounds, respectively, of an interval for the estimated percentage of analyzed state tests that qualified as deeper learning assessments.

The proportion of state mathematics tests that met the criteria for deeper learning assessments was 0 percent, irrespective of whether Criterion A or Criterion B was used. The proportion of reading tests that qualified as deeper learning assessments was 1 percent under Criterion A and 20 percent under Criterion B. For writing tests, these numbers were 28 percent and 31 percent, respectively. Tables B.1–B.3 in Appendix B present the total number of items analyzed, the percentage of items at each DOK level, and whether a test met the criteria for deeper learning assessments by subject, state, and grade.

5. Estimating the Percentage of U.S. Elementary and Secondary Students Assessed on Deeper Learning Through State Achievement Tests

After we finished analyzing whether selected state tests qualified as deeper learning assessments, our final step was to estimate the percentage of students assessed on deeper learning skills through state achievement tests. In this chapter, we present our estimation results and discuss the caveats and limitations of this study.

ESTIMATION RESULTS

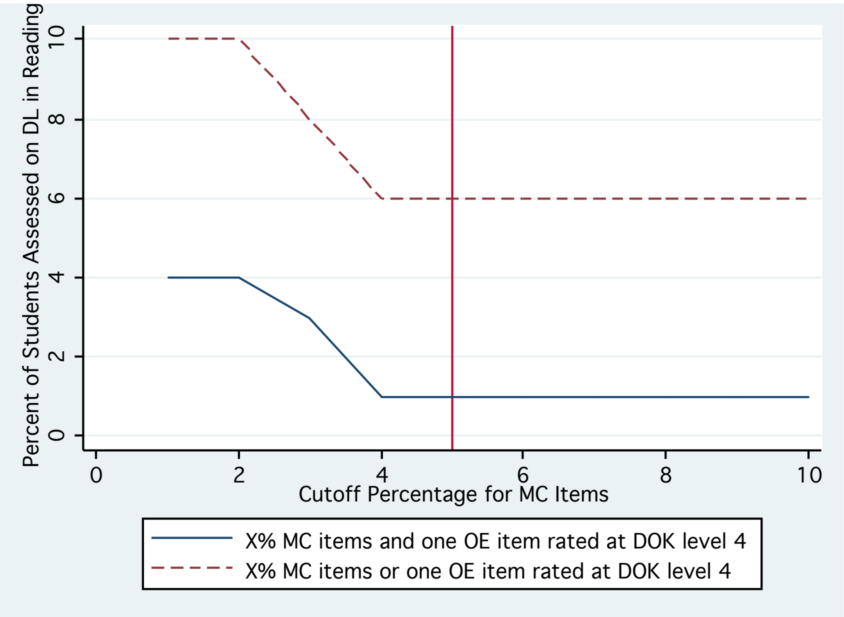
We downloaded a data file containing the number of students enrolled at each grade level in the 2009–2010 school year, by state, for all 50 states and the District of Columbia from the Common Core of Data at the National Center for Education Statistics. We merged this data file with the file containing the classification results of state tests as deeper learning assessments to calculate the percentage of U.S. elementary and secondary students assessed on deeper learning skills through the state achievement tests.

We conducted these analyses under the assumption that none of the tests from the remaining states would meet the criteria for deeper learning assessments. We treated the percentage of students assessed on deeper learning skills under Criterion A and Criterion B as the lower and upper bounds of an interval for the final estimated percentage of U.S. elementary and secondary students assessed on deeper learning skills through state achievement tests. In addition, we also experimented with different cutoff scores for the percentage of MC items rated at DOK level 4 to see how stable our final estimation would be.

Our estimation results show that the percentage of students assessed on deeper learning skills in mathematics nationwide was zero, regardless of which criterion was used. For writing, 2–3 percent of students were assessed on deeper learning skills through state tests. The proportion of students assessed on deeper learning in reading ranged from 1 to 6 percent (see Figure 5.1). Figure 5.1 also shows the range for the percentage of students assessed on deeper learning skills in reading when different cutoff scores were adopted for the percentage of MC items rated at DOK level 4. When a cutoff percentage for MC items of 4 percent or higher was adopted, the range for the percentage of students assessed on deeper learning skills in reading stayed the same.

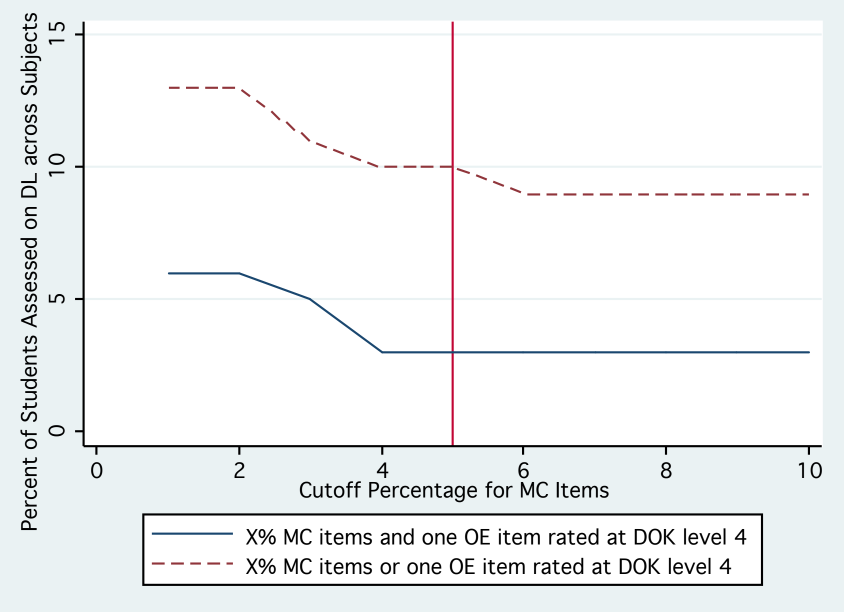
After we combined the results across the three subjects, the estimated proportion of students assessed on deeper learning skills through state mathematics and English language arts achievement tests ranged from 3 to 10 percent (see Figure 5.2). When a cutoff percentage for MC items of 4 percent or higher was adopted, the lower bound of the interval for the percentage of students assessed on deeper learning skills through state mathematics and English language arts tests stayed at 3 percent, while the upper bound dropped slightly from 10 percent to 9 percent when the cutoff percentage reached 6 percent.

Figure 5.1. Estimated Percentage of Students Assessed on Deeper Learning Skills in Reading Nationwide with Different Cutoff Percentages for MC Items



NOTE: The reference line is the 5-percent cutoff that we adopted in this study.

Figure 5.2. Estimated Percentage of Students Assessed on Deeper Learning Skills in Mathematics and English Language Arts Nationwide with Different Cutoff Percentages for MC Items



NOTE: The reference line is the 5-percent cutoff that we adopted in this study.

Interpreting The Results

There are several caveats worth noting when interpreting the results of this analysis. First, due to the lack of information about test items and the number of test takers for some tests, such as the AP and IB exams, we had to use state achievement tests to determine the extent to which students are assessed on deeper learning skills. This constraint likely underestimates the percentage of students assessed on deeper learning skills in these states.

Second, the manner in which state achievement tests are administered did not allow us to analyze collaboration, oral communication, and learn-how-to-learn skills in this project. If we had included these three aspects of deeper learning skills in our operational definition, no state test would have met the criteria for a deeper learning assessment. Although omitting these three aspects from the operational definition of deeper learning might have overestimated the percentage of state tests that met the criteria for deeper learning assessments, it allowed us to conduct meaningful analysis of the extent to which current state tests measure other important aspects of deeper learning, including the mastery of core academic standards, critical-thinking and problem-solving skills, and written communication skills.

Third, due to limited project resources, our estimate of the percentage of U.S. elementary and secondary students assessed on deeper learning skills through state mathematics and English language arts achievement tests was based on a judgment of cognitive rigor of the state tests used in 17 states only. State tests used in the 17 states in our sample were identified as more rigorous than state tests used in the other two-thirds of U.S. states by prior studies using different methods and criteria. We assumed that the results concerning the rigor of the 17 state tests published in prior reviews were accurate and that the rigor level of state tests has not changed substantially since those reviews were conducted. We also assumed that none of the state tests used in the other two-thirds of states met the criteria for deeper learning assessments.

Fourth, we examined the cognitive rigor of state mathematics and English language arts tests based on released state test items. However, many states released partial test forms instead of complete test forms. In these states, it is unknown whether released test items are representative of all the test items used in 2009–2010 in terms of the level of cognitive rigor. Thus, a lack of access to the full test forms might have introduced bias in this portion of the evaluation. However, the issue of partial test forms is unavoidable. There are a number of reasons states do not release full test forms and we could only work with the items they did release.

Fifth, we assessed whether a state test met the criteria for a deeper learning assessment based on the percentage or number of test items that were rated at DOK level 4, which is only one possible measure of the cognitive rigor of a state test. We also considered using the portion of the total test score that is accounted for by DOK level 4 items to represent the cognitive rigor of a state test. However, because not all 17 states provided the number of score points for released items and the total scores of their state tests, we could not assess cognitive rigor based on the portion of the total test scores accounted for by DOK level 4 items.

Sixth, the choice of the cutoff percentage of MC items rated at DOK level 4 is admittedly arbitrary. To examine how sensitive the final estimation results would be to the changes in the cutoff score, we conducted the analysis with different cutoff scores. The results showed that the final estimation of the percentage of students assessed on deeper learning skills in mathematics and English language arts was relatively stable once the cutoff was 4percent or above. Our choice of 5 percent as the cutoff score is within the range of cutoff scores that led to a relatively stable estimation of the percentage of U.S.

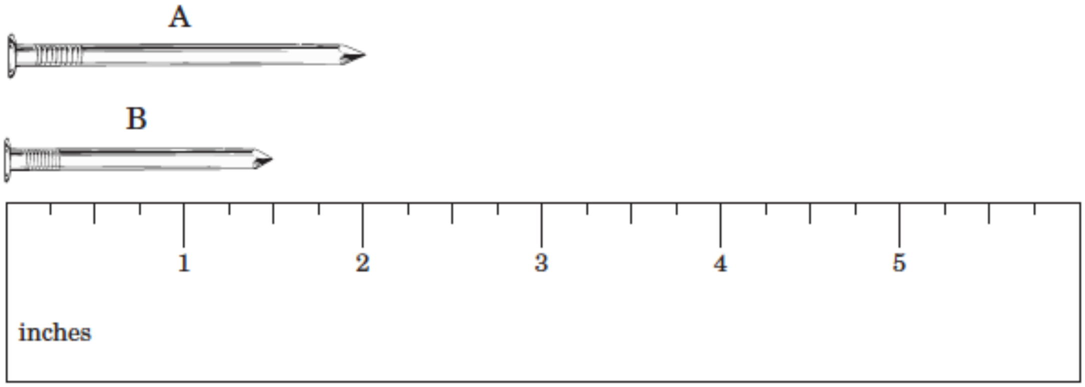
elementary and secondary students assessed on deeper learning through state tests. However, had more states and more complete state test forms been included in the analysis, the range for the cutoff scores might differ, as would the final estimation results.

To summarize, we estimated the percentage of U.S. elementary and secondary students assessed on three types of deeper learning skills (i.e., mastery of core academic content, critical-thinking and problem-solving skills, and written communication skills) through state mathematics and English language arts achievement tests in the 2009–2010 school year to be between 3 and 10 percent. However, this estimation was conducted under a set of assumptions and constraints. Caution is warranted when interpreting the results.

APPENDIX A: EXEMPLARY TEST ITEMS AT EACH DOK LEVEL

Math DOK1 (Grade 4)

Look at the length of nails A and B.



How much longer is nail A than nail B?

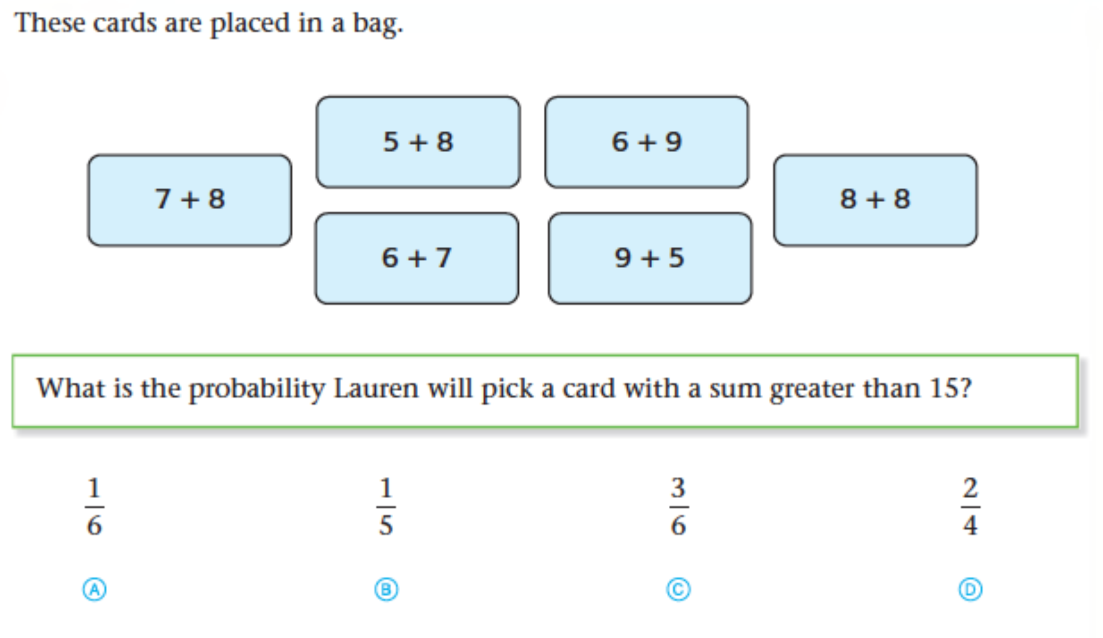
○ **A.** $\frac{1}{2}$ inch

○ **B.** $1\frac{1}{2}$ inches

○ **C.** $3\frac{1}{2}$ inches

SOURCE: State of Washington Office of Superintendent of Public Instruction (undated).
NOTE: This item asks students to measure the difference in the length of two nails. It is a one-step task and requires students to recognize the length of the difference on a ruler.

Math DOK2 (Grade 4)



These cards are placed in a bag.

7 + 8

5 + 8

6 + 7

6 + 9

9 + 5

8 + 8

What is the probability Lauren will pick a card with a sum greater than 15?

$\frac{1}{6}$ Ⓐ

$\frac{1}{5}$ Ⓑ

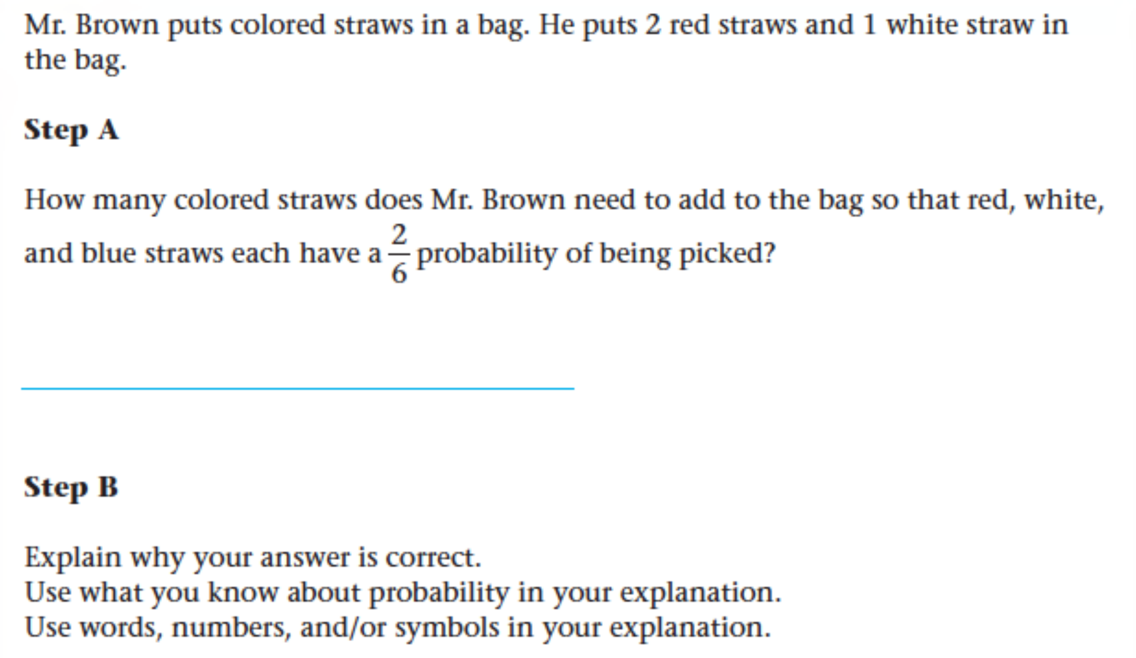
$\frac{3}{6}$ Ⓒ

$\frac{2}{4}$ Ⓓ

SOURCE: Maryland State Department of Education (2009).

NOTE: This task requires two steps: (1) calculating the total for each card and (2) calculating the probability of picking a card with a sum greater than 15. It involves more than one step and the application of the probability formula.

Math DOK 3 (Grade 4)



Mr. Brown puts colored straws in a bag. He puts 2 red straws and 1 white straw in the bag.

Step A

How many colored straws does Mr. Brown need to add to the bag so that red, white, and blue straws each have a $\frac{2}{6}$ probability of being picked?

____

Step B

Explain why your answer is correct.
Use what you know about probability in your explanation.
Use words, numbers, and/or symbols in your explanation.

SOURCE: Maryland State Department of Education (2009).
NOTE: This item requires students to explain their thinking process; the thinking process used to solve the probability question is abstract and requires multiple steps.

Math DOK4 (Grade 4, adapted from an existing test item)

Pick a Phone Plan

This month, Mrs. Smith's telephone bill included information about a new long-distance plan being offered. The plans are listed below.

Current Plan	Monthly service fee of $4.75 plus $0.08 a minute for each call anytime of the day.
New Flat Rate Plan	No monthly service fee, pay $0.20 a minute for each call anytime of the day.
New Variable Rate Plan	Monthly service fee of $2.50, pay $0.12 for each call between 8 AM and 5 PM on weekdays, pay $0.14 for each call after 5 PM on weekdays, and pay $0.16 anytime on weekends.

Mrs. Smith would like to speak to her grandchildren as much as possible but would also like to make the calls during times that fit her schedule. Each of her calls lasts for 10 minutes. She can reach her grandchildren at anytime of the day on the weekends, but only between 3 PM and 7 PM on weekdays. If she allots $30 a month to spend on her telephone bill, which plan should she choose and why?

Show all your work. Explain in words **which plan you chose and why.** Also tell **why** you took the steps you did to solve the problems and explain the **advantages and disadvantages** of the chosen plan.

SOURCE: Rockford Public Schools (undated).

NOTE: This item requires students to choose between three alternative plans, describe their thinking processes and their assumptions, and justify their choice. There can be multiple answers, depending on the assumptions and trade-offs.

Hamburger Me at the Car!

by Unknown

Watson sat on the edge of his bed and looked out his bedroom window. He glanced at his watch. It was 9:59 a.m. A moment later, at exactly 10:00, he saw Ross Bailey leave his house. Ross strolled down the sidewalk with an empty white sack slung over his shoulder.

Watson felt like a detective on his first case. He raced down the steps to follow Ross. Every Saturday at exactly 10:00 a.m., Ross Bailey left with that empty sack. He always returned later with his friend Buddy and a full sack. *Just where were they going, and what did they put in that sack?* Watson meant to find out!

The front door slammed behind Watson as he left to follow Ross down the street.

Watson raced to the end of the street and looked both ways, but Ross had disappeared.

For the rest of the week, Watson watched Ross on the school bus and in the classroom. He followed him home. But nothing happened.

"Getting the money!"

For a moment, Ross and Buddy looked confused. Then Ross began to laugh.

"You must have read my note!" he said.

Buddy laughed too. "But we haven't gotten the money yet," he said as he turned the sack upside down. Aluminum cans spilled to the ground.

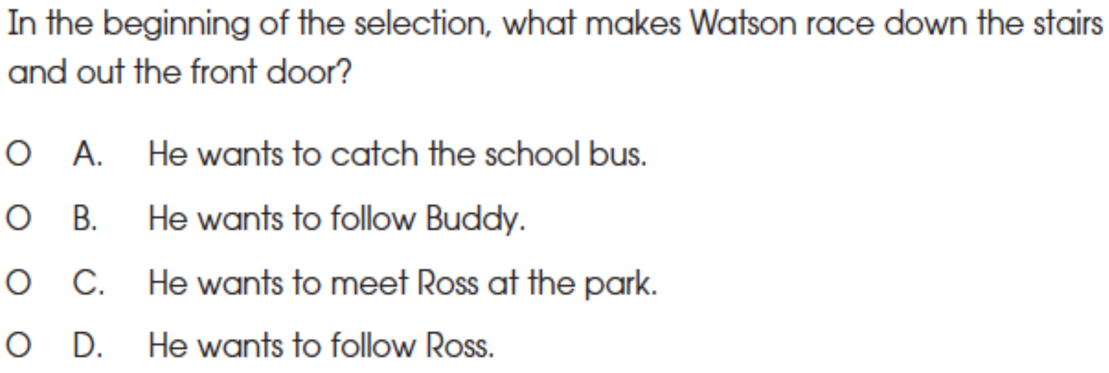
Watson looked at the cans, then at the boys.

"We get the money when we take the cans to the recycling center," Buddy explained. "You can help us if you want to."

Watson was quiet for a moment. "Thanks," he said finally, "but I still have some work to do."

Already, Watson was planning his next case. *Just who is dropping empty cans in the park? And why?* Watson meant to find out!

Reading DOK1 (Grade 4)



In the beginning of the selection, what makes Watson race down the stairs and out the front door?

○ A. He wants to catch the school bus.

○ B. He wants to follow Buddy.

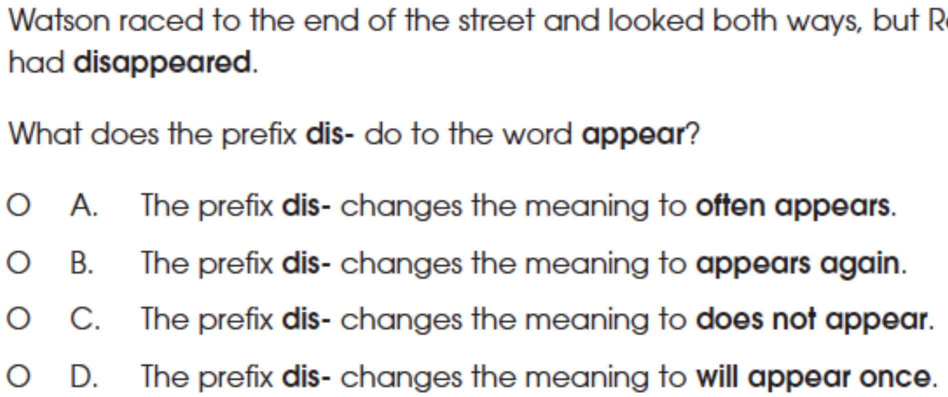
○ C. He wants to meet Ross at the park.

○ D. He wants to follow Ross.

SOURCE: Ohio Department of Education (undated).

NOTE: This item asks students to support ideas by referencing details in the text.

Reading DOK2 (Grade 4) Ohio Grade 4 2005



Watson raced to the end of the street and looked both ways, but Ross had **disappeared**.

What does the prefix **dis-** do to the word **appear**?

○ A. The prefix **dis-** changes the meaning to **often appears**.

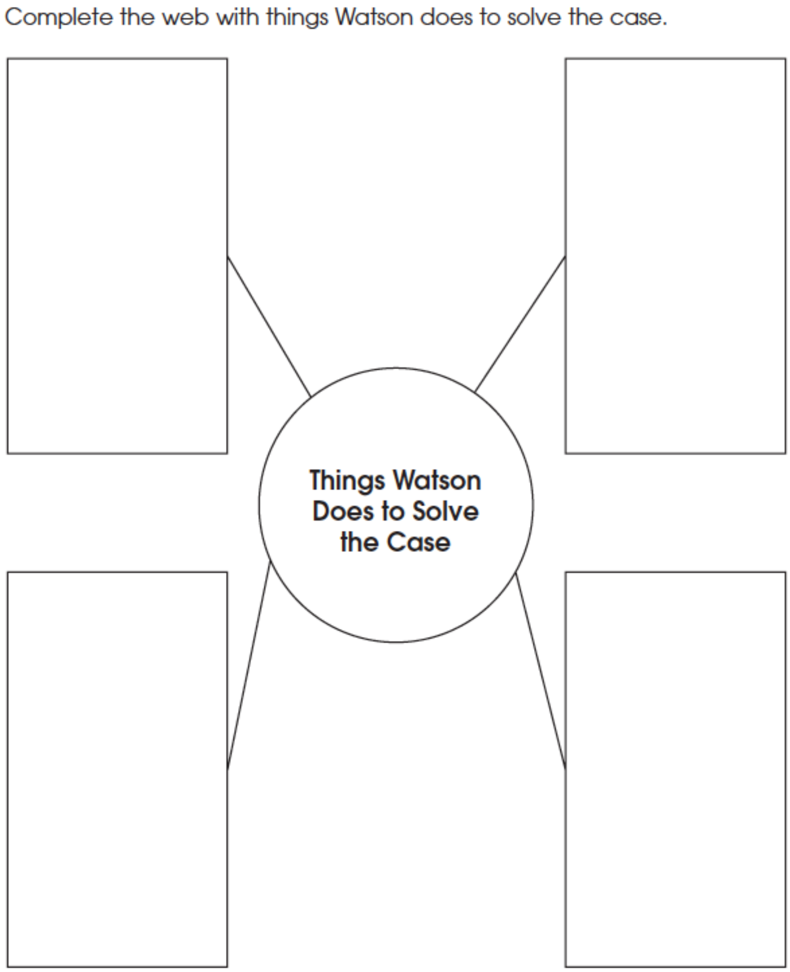
○ B. The prefix **dis-** changes the meaning to **appears again**.

○ C. The prefix **dis-** changes the meaning to **does not appear**.

○ D. The prefix **dis-** changes the meaning to **will appear once**.

SOURCE: Ohio Department of Education (undated).
NOTE: This item requires students to use contextual cues to identify the meaning of unfamiliar words.

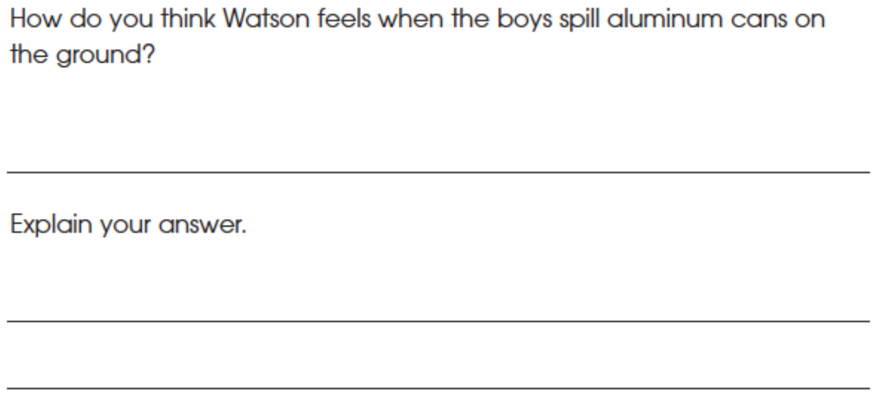
Reading DOK3 (Grade 4)



Complete the web with things Watson does to solve the case.

SOURCE: Ohio Department of Education (undated).
NOTE: This item asks students to summarize information from different parts of the passage to address a specific topic.

Reading DOK4 (Grade 4)



How do you think Watson feels when the boys spill aluminum cans on the ground?

____

Explain your answer.

____

____

SOURCE: Ohio Department of Education (undated).

NOTE: Students are asked to develop their own opinion and justify their answer. This requires analyzing and synthesizing information from all sections of the passage and supporting their answers with evidence from the text.

Writing DOK1 (Grade 4)

Row, Row, Row Your *Pumpkin*!

(1) Around the beginning of October, you probably start to notice that pumpkins are everywhere. (2) They're piled high outside markets. (3) People set them in their front yards. (4) Businesses are decorated with them and schools, too. (5) The festive orange fruits they are a familiar symbol of the fall season.

(6) Pumpkins come in different shapes and sizes. (7) Some are small enough to fit in your hand. (8) Others, however, is enormous. (9) These giant pumpkins can weigh more than 1,000 pounds. (10) What can you do with a pumpkin that large? (11) Believe it or not, some people choose to go pumpkin boating.

(12) Wayne Hackney, of New Milford, Connecticut, was the first person to make a boat out of a pumpkin. (13) He wears an orange tuxedo on special occasions. (14) In 1996 he attached a small motor to the back of a large, hollowed-out pumpkin. (15) He then climbed inside he made his way across a lake. (16) Their clever idea gained the attention of several reporters. (17) Before long, pumpkin-boat races were popping up in places all over the country.

What change, if any, should be made in sentence 5?

A Delete ***they***
B Change ***are*** to **were**
C Change ***season*** to **Season**
D Make no change

Source: Texas Department of Education (2009). Writing Grade 4.
Note: Students are asked to identify standard English grammatical structures.

Writing DOK2 (Grade 4)

What revision, if any, is needed in sentence 15?
A He then climbed inside and made
his way. Across a lake.
B He then climbed inside. And made
his way across a lake.
C He then climbed inside and made
his way across a lake.
D No revision is needed.

SOURCE: Texas Education Agency (2009).
NOTE: Students are asked to construct compound sentences.

Writing DOK3 (Grade 8)
Some people think that it is the responsibility of schools to teach students to save and manage money. Do you think there should be a class at school that teaches students how to save and manage money? Write a paragraph that develops one strong argument supporting your position.

SOUCE: Rhode Island Department of Elementary and Secondary Education (2010).
NOTE: Students are asked to support their stance with supporting details and examples from the passage.

Writing DOK4 (Grade 8)

A student wrote this fact sheet about writing in ancient Rome. As you read the fact sheet, think about what a person from ancient Rome would find familiar and/or different about writing today. Then write a response to the prompt that follows.

Writing in Ancient Rome

- Romans used sticks to write rough drafts on wax-covered boards and rubbed the words away afterward
- made pens by cutting the end of a bamboo or reed stem into a point and filling the point with ink
- *papyrus* (pf pí rfss): a paper-like material made from the papyrus plant and used for writing
- before books, Romans used scrolls, sheets of papyrus sewn together and rolled out to read
- writing tools affected the shape of Roman letters:
  - hammer and chisel made angular letters
  - reed or bamboo pen made flowing letters
- invented books from sheets of papyrus sewn together to replace scrolls
- first Romans to use writing were the upper classes; eventually, most Romans were taught to read and write
- three types of Roman handwriting:
  - squared letters for inscribing monuments and buildings
  - flowing letters for writing official documents
  - plain letters for writing first drafts
- government, business, and legal documents were written in ink on papyrus so they would be permanent
- no lowercase letters; writing tools were not useful for making detailed letters
- used inks made of combinations of
  - berries, plants, and minerals
  - soot, resin, wine, and octopus ink

What would a person from ancient Rome find familiar and/or different about writing today? Select information from the fact sheet and use your own knowledge to write a report.

SOURCE: Rhode Island Department of Elementary and Secondary Education (2010).

NOTE: Students are asked to compare and contrast and to synthesize ideas across both passages.

APPENDIX B. PERCENTAGE OF RELEASED STATE TEST ITEMS RATED AT EACH DOK LEVEL, BY SUBJECT, STATE, AND GRADE LEVEL

Table B.1. Percentage of Released State Mathematics Test Items Rated at Each DOK Level, by State and Grade Level, and Deeper Learning Assessment Classification Results

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
California	3	96	100%	0%	96	88%	13%	0%	0%	0	/	/	/	/			N[2]	
California	4	96	100%	0%	96	86%	14%	0%	0%	0	/	/	/	/			N	
California	5	96	100%	0%	96	83%	17%	0%	0%	0	/	/	/	/			N	
California	6	96	100%	0%	96	81%	19%	0%	0%	0	/	/	/	/			N	
California	7	96	100%	0%	96	77%	23%	0%	0%	0	/	/	/	/			N	
California	12	287	100%	0%	287	59%	41%	0%	0%	0	/	/	/	/			N	
Colorado	3	7	29%	71%	2	50%	50%	0%	0%	5	40%	60%	0%	0%	N	N		
Colorado	4	9	11%	89%	1	0%	100%	0%	0%	8	25%	75%	0%	0%	N	N		
Colorado	5	5	0%	100%	0	/	/	/	/	5	20%	80%	0%	0%				N
Colorado	6	7	0%	100%	0	/	/	/	/	7	43%	57%	0%	0%				N
Colorado	7	4	0%	100%	0	/	/	/	/	4	0%	75%	25%	0%				N
Colorado	8	4	0%	100%	0	/	/	/	/	4	0%	100%	0%	0%				N
Colorado	9	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Colorado	10	6	0%	100%	0	/	/	/	/	6	33%	67%	0%	0%				N
Connecticut	10	10	0%	100%	0	/	/	/	/	10	10%	30%	60%	0%				N
Delaware	3	5	0%	100%	0	/	/	/	/	5	0%	100%	0%	0%				N
Delaware	4	5	40%	60%	2	50%	50%	0%	0%	3	33%	0%	67%	0%	N	N		
Delaware	5	5	20%	80%	1	100%	0%	0%	0%	4	0%	50%	50%	0%	N	N		
Delaware	6	9	22%	78%	2	50%	50%	0%	0%	7	0%	86%	14%	0%	N	N		

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Delaware	7	7	43%	57%	3	33%	67%	0%	0%	4	0%	75%	25%	0%	N	N		
Delaware	8	9	33%	67%	3	100%	0%	0%	0%	6	0%	83%	17%	0%	N	N		
Delaware	10	10	50%	50%	5	60%	40%	0%	0%	5	0%	40%	60%	0%	N	N		
Kentucky	5	11	0%	100%	0	/	/	/	/	11	9%	82%	9%	0%				N
Kentucky	8	12	0%	100%	0	/	/	/	/	12	8%	83%	8%	0%				N
Kentucky	11	13	0%	100%	0	/	/	/	/	13	8%	92%	0%	0%				N
Maryland	3	14	71%	29%	10	30%	70%	0%	0%	4	25%	75%	0%	0%	N	N		
Maryland	4	14	71%	29%	10	70%	30%	0%	0%	4	25%	75%	0%	0%	N	N		
Maryland	5	15	67%	33%	10	50%	50%	0%	0%	5	40%	40%	20%	0%	N	N		
Maryland	6	15	67%	33%	10	70%	30%	0%	0%	5	20%	80%	0%	0%	N	N		
Maryland	7	15	67%	33%	10	60%	40%	0%	0%	5	20%	80%	0%	0%	N	N		
Maryland	8	15	67%	33%	10	60%	40%	0%	0%	5	40%	60%	0%	0%	N	N		
Maryland	12	52	81%	19%	42	33%	67%	0%	0%	10	60%	40%	0%	0%	N	N		
Massachusetts	3	20	65%	35%	13	77%	23%	0%	0%	7	57%	43%	0%	0%	N	N		
Massachusetts	4	26	62%	38%	16	63%	38%	0%	0%	10	30%	70%	0%	0%	N	N		
Massachusetts	5	25	68%	32%	17	71%	29%	0%	0%	8	25%	75%	0%	0%	N	N		
Massachusetts	6	25	64%	36%	16	69%	31%	0%	0%	9	11%	89%	0%	0%	N	N		
Massachusetts	7	26	62%	38%	16	75%	25%	0%	0%	10	20%	70%	10%	0%	N	N		
Massachusetts	8	26	62%	38%	16	63%	38%	0%	0%	10	40%	60%	0%	0%	N	N		
Massachusetts	10	59	54%	46%	32	56%	44%	0%	0%	27	19%	81%	0%	0%	N	N		
Missouri	3	24	79%	21%	19	63%	37%	0%	0%	5	0%	100%	0%	0%	N	N		
Missouri	4	31	58%	42%	18	56%	44%	0%	0%	13	23%	77%	0%	0%	N	N		
Missouri	5	24	83%	17%	20	60%	40%	0%	0%	4	0%	100%	0%	0%	N	N		
Missouri	6	23	78%	22%	18	61%	39%	0%	0%	5	100%	0%	0%	0%	N	N		
Missouri	7	26	77%	23%	20	55%	45%	0%	0%	6	17%	83%	0%	0%	N	N		

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Missouri	8	30	70%	30%	21	29%	71%	0%	0%	9	22%	67%	11%	0%	N	N		
Missouri	10	30	67%	33%	20	55%	45%	0%	0%	10	10%	70%	20%	0%	N	N		
NECAP[3]	3	19	53%	47%	10	50%	50%	0%	0%	9	44%	33%	22%	0%	N	N		
NECAP	4	19	53%	47%	10	70%	30%	0%	0%	9	67%	33%	0%	0%	N	N		
NECAP	5	18	56%	44%	10	70%	30%	0%	0%	8	38%	50%	13%	0%	N	N		
NECAP	6	16	56%	44%	9	78%	22%	0%	0%	7	57%	43%	0%	0%	N	N		
NECAP	7	18	56%	44%	10	50%	50%	0%	0%	8	63%	38%	0%	0%	N	N		
NECAP	8	19	53%	47%	10	60%	40%	0%	0%	9	33%	56%	11%	0%	N	N		
NECAP	11	27	44%	56%	12	92%	8%	0%	0%	15	47%	47%	7%	0%	N	N		
New Jersey	3	22	82%	18%	18	67%	33%	0%	0%	4	50%	50%	0%	0%	N	N		
New Jersey	4	31	65%	35%	20	75%	25%	0%	0%	11	9%	82%	9%	0%	N	N		
New Jersey	5	16	0%	100%	0	/	/	/	/	16	25%	75%	0%	0%				N
New Jersey	8	49	69%	31%	34	41%	59%	0%	0%	15	7%	87%	7%	0%	N	N		
New Jersey	12	11	0%	100%	0	/	/	/	/	11	27%	73%	0%	0%				N
New York	3	33	79%	21%	26	81%	19%	0%	0%	7	29%	71%	0%	0%	N	N		
New York	4	56	54%	46%	30	70%	30%	0%	0%	26	38%	62%	0%	0%	N	N		
New York	5	42	60%	40%	25	76%	24%	0%	0%	17	24%	76%	0%	0%	N	N		
New York	6	40	63%	38%	25	80%	20%	0%	0%	15	33%	67%	0%	0%	N	N		
New York	7	43	70%	30%	30	70%	30%	0%	0%	13	15%	85%	0%	0%	N	N		
New York	8	53	51%	49%	27	78%	22%	0%	0%	26	15%	85%	0%	0%	N	N		
New York	12	152	67%	33%	102	34%	66%	0%	0%	50	8%	80%	12%	0%	N	N		
Ohio	3	19	79%	21%	15	73%	27%	0%	0%	4	0%	75%	25%	0%	N	N		
Ohio	4	18	94%	6%	17	76%	24%	0%	0%	1	0%	100%	0%	0%	N	N		
Ohio	5	15	100%	0%	15	73%	27%	0%	0%	0	/	/	/	/			N	
Ohio	6	18	72%	28%	13	85%	15%	0%	0%	5	0%	80%	20%	0%	N	N		

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Ohio	7	16	88%	13%	14	36%	64%	0%	0%	2	0%	100%	0%	0%	N	N		
Ohio	8	19	58%	42%	11	55%	45%	0%	0%	8	13%	88%	0%	0%	N	N		
Ohio	12	39	85%	15%	33	33%	67%	0%	0%	6	0%	67%	33%	0%	N	N		
Texas	3	40	98%	3%	39	62%	38%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	4	42	98%	2%	41	76%	24%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	5	44	98%	2%	43	81%	19%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	6	46	98%	2%	45	76%	24%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	7	48	98%	2%	47	77%	23%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	8	50	98%	2%	49	82%	18%	0%	0%	1	100%	0%	0%	0%	N	N		
Texas	9	52	100%	0%	52	69%	31%	0%	0%	0	/	/	/	/			N	
Texas	10	56	100%	0%	56	82%	18%	0%	0%	0	/	/	/	/			N	
Texas	12	60	100%	0%	60	72%	28%	0%	0%	0	/	/	/	/			N	
Washington	3	7	43%	57%	3	33%	67%	0%	0%	4	0%	100%	0%	0%	N	N		
Washington	4	7	43%	57%	3	100%	0%	0%	0%	4	25%	75%	0%	0%	N	N		
Washington	5	6	50%	50%	3	100%	0%	0%	0%	3	0%	100%	0%	0%	N	N		
Washington	6	6	50%	50%	3	67%	33%	0%	0%	3	0%	67%	33%	0%	N	N		
Washington	7	7	43%	57%	3	67%	33%	0%	0%	4	25%	50%	25%	0%	N	N		
Washington	8	6	50%	50%	3	33%	67%	0%	0%	3	0%	67%	33%	0%	N	N		
Washington	10	10	70%	30%	7	29%	71%	0%	0%	3	0%	100%	0%	0%	N	N		

NOTE:

1. Criterion A and Criterion B are applied to tests with both MC and OE items. Criterion MC and Criterion OE refer to the criterion applicable to tests with MC items only or OE items only, respectively.
2. Y = A test met the criteria for a deeper learning assessment; N = otherwise.
3. NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

Table B.2. Percentage of Released State Reading Test Items Rated at Each DOK Level, by State and Grade Level, and Deeper Learning Assessment Classification Results

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
California	3	96	65%	35%	62	42%	48%	10%	0%	34	71%	26%	3%	0%	N[2]	N		
California	4	114	55%	45%	63	37%	46%	14%	3%	51	63%	33%	4%	0%	N	N		
California	5	114	100%	0%	114	46%	44%	10%	0%	0	/	/	/	/			N	
California	6	114	100%	0%	114	41%	52%	7%	0%	0	/	/	/	/			N	
California	7	114	100%	0%	114	47%	38%	15%	0%	0	/	/	/	/			N	
California	8	114	100%	0%	114	42%	39%	16%	4%	0	/	/	/	/			N	
California	11	114	100%	0%	114	49%	26%	22%	3%	0	/	/	/	/			N	
Colorado	3	3	0%	100%	0	/	/	/	/	3	67%	33%	0%	0%				N
Colorado	4	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Colorado	5	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Colorado	6	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Colorado	7	3	0%	100%	0	/	/	/	/	3	0%	67%	33%	0%				N
Colorado	8	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Connecticut	10	6	100%	0%	6	83%	17%	0%	0%	0	/	/	/	/			N	
Delaware	3	6	67%	33%	4	75%	25%	0%	0%	2	0%	50%	0%	50%	N	Y		
Delaware	4	6	67%	33%	4	75%	25%	0%	0%	2	0%	0%	100%	0%	N	N		
Delaware	6	5	60%	40%	3	0%	100%	0%	0%	2	0%	50%	50%	0%	N	N		
Delaware	7	6	67%	33%	4	25%	75%	0%	0%	2	0%	100%	0%	0%	N	N		
Delaware	8	5	60%	40%	3	33%	33%	33%	0%	2	0%	50%	50%	0%	N	N		
Delaware	10	5	60%	40%	3	33%	0%	67%	0%	2	0%	0%	50%	50%	N	Y		
Kentucky	4	20	80%	20%	16	50%	31%	19%	0%	4	0%	50%	50%	0%	N	N		
Kentucky	7	20	80%	20%	16	31%	50%	19%	0%	4	0%	25%	75%	0%	N	N		

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Kentucky	10	15	80%	20%	12	58%	17%	25%	0%	3	0%	0%	100%	0%	N	N		
Maryland	3	33	82%	18%	27	7%	93%	0%	0%	6	0%	33%	67%	0%	N	N		
Maryland	4	29	83%	17%	24	13%	88%	0%	0%	5	0%	0%	100%	0%	N	N		
Maryland	5	23	83%	17%	19	16%	68%	11%	5%	4	0%	50%	50%	0%	N	Y		
Maryland	6	25	80%	20%	20	20%	80%	0%	0%	5	0%	20%	80%	0%	N	N		
Maryland	7	24	83%	17%	20	0%	85%	15%	0%	4	0%	0%	50%	50%	N	Y		
Maryland	8	22	86%	14%	19	16%	68%	16%	0%	3	0%	0%	100%	0%	N	N		
Maryland	12	60	100%	0%	60	37%	37%	27%	0%	0	/	/	/	/			N	
Massachusetts	3	17	88%	12%	15	47%	53%	0%	0%	2	0%	50%	50%	0%	N	N		
Massachusetts	4	18	83%	17%	15	53%	27%	20%	0%	3	0%	0%	67%	33%	N	Y		
Massachusetts	5	16	94%	6%	15	53%	40%	7%	0%	1	0%	0%	100%	0%	N	N		
Massachusetts	6	17	88%	12%	15	40%	47%	7%	7%	2	0%	50%	50%	0%	N	Y		
Massachusetts	7	20	80%	20%	16	13%	44%	44%	0%	4	0%	25%	75%	0%	N	N		
Massachusetts	8	12	92%	8%	11	27%	55%	18%	0%	1	0%	0%	100%	0%	N	N		
Massachusetts	10	42	88%	12%	37	27%	57%	16%	0%	5	0%	0%	100%	0%	N	N		
Missouri	3	6	33%	67%	2	100%	0%	0%	0%	4	0%	25%	75%	0%	N	N		
Missouri	4	6	33%	67%	2	50%	50%	0%	0%	4	0%	50%	50%	0%	N	N		
Missouri	5	6	33%	67%	2	50%	50%	0%	0%	4	0%	100%	0%	0%	N	N		
Missouri	6	6	33%	67%	2	0%	100%	0%	0%	4	0%	0%	100%	0%	N	N		
Missouri	7	6	33%	67%	2	0%	50%	50%	0%	4	0%	50%	50%	0%	N	N		
Missouri	8	6	33%	67%	2	50%	50%	0%	0%	4	0%	50%	50%	0%	N	N		
Missouri	11	6	50%	50%	3	0%	67%	33%	0%	3	0%	33%	67%	0%	N	N		
NECAP[3]	3	12	83%	17%	10	90%	10%	0%	0%	2	0%	0%	100%	0%	N	N		
NECAP	4	12	83%	17%	10	80%	10%	10%	0%	2	50%	0%	50%	0%	N	N		
NECAP	5	12	83%	17%	10	70%	20%	10%	0%	2	0%	0%	100%	0%	N	N		

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
NECAP	6	12	83%	17%	10	80%	10%	10%	0%	2	0%	0%	100%	0%	N	N		
NECAP	7	12	83%	17%	10	80%	20%	0%	0%	2	0%	0%	100%	0%	N	N		
NECAP	8	12	83%	17%	10	60%	10%	30%	0%	2	0%	0%	100%	0%	N	N		
NECAP	11	17	82%	18%	14	36%	50%	14%	0%	3	0%	0%	100%	0%	N	N		
New Jersey	3	7	86%	14%	6	50%	33%	17%	0%	1	0%	0%	0%	100%	N	Y		
New Jersey	4	7	86%	14%	6	50%	50%	0%	0%	1	0%	100%	0%	0%	N	N		
New Jersey	8	24	83%	17%	20	20%	55%	25%	0%	4	0%	25%	50%	25%	N	Y		
New York	3	28	89%	11%	25	44%	44%	12%	0%	3	0%	100%	0%	0%	N	N		
New York	4	35	80%	20%	28	39%	57%	4%	0%	7	0%	71%	14%	14%	N	Y		
New York	5	27	93%	7%	25	44%	52%	4%	0%	2	0%	50%	50%	0%	N	N		
New York	6	34	76%	24%	26	46%	50%	4%	0%	8	0%	63%	25%	13%	N	Y		
New York	7	35	89%	11%	31	35%	61%	3%	0%	4	0%	75%	25%	0%	N	N		
New York	8	34	76%	24%	26	31%	42%	27%	0%	8	0%	38%	50%	13%	N	Y		
New York	12	28	89%	11%	25	44%	16%	36%	4%	3	0%	0%	33%	67%	N	Y		
Ohio	3	16	88%	13%	14	21%	71%	7%	0%	2	0%	100%	0%	0%	N	N		
Ohio	4	15	80%	20%	12	42%	42%	17%	0%	3	0%	100%	0%	0%	N	N		
Ohio	5	17	76%	24%	13	15%	62%	23%	0%	4	25%	75%	0%	0%	N	N		
Ohio	6	16	81%	19%	13	23%	54%	23%	0%	3	0%	100%	0%	0%	N	N		
Ohio	7	16	81%	19%	13	31%	54%	15%	0%	3	0%	100%	0%	0%	N	N		
Ohio	8	18	83%	17%	15	7%	60%	33%	0%	3	0%	33%	67%	0%	N	N		
Ohio	12	38	84%	16%	32	9%	69%	22%	0%	6	0%	33%	67%	0%	N	N		
Texas	3	36	100%	0%	36	39%	44%	17%	0%	0	/	/	/	/			N	
Texas	4	40	100%	0%	40	25%	55%	20%	0%	0	/	/	/	/			N	
Texas	5	42	100%	0%	42	21%	52%	26%	0%	0	/	/	/	/			N	
Texas	6	42	100%	0%	42	24%	52%	24%	0%	0	/	/	/	/			N	

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Texas	7	48	100%	0%	48	19%	46%	35%	0%	0	/	/	/	/			N	
Texas	8	48	100%	0%	48	25%	42%	33%	0%	0	/	/	/	/			N	
Texas	9	36	92%	8%	33	18%	52%	30%	0%	3	0%	0%	67%	33%	N	Y		
Texas	10	31	100%	0%	31	13%	42%	35%	10%	0	/	/	/	/			Y	
Texas	12	32	88%	13%	28	11%	29%	61%	0%	4	0%	0%	25%	75%	N	Y		
Washington	3	10	70%	30%	7	0%	57%	43%	0%	3	0%	67%	33%	0%	N	N		
Washington	4	4	75%	25%	3	67%	33%	0%	0%	1	0%	0%	100%	0%	N	N		
Washington	5	5	80%	20%	4	50%	50%	0%	0%	1	0%	100%	0%	0%	N	N		
Washington	6	6	67%	33%	4	0%	75%	25%	0%	2	0%	0%	100%	0%	N	N		
Washington	7	13	77%	23%	10	30%	70%	0%	0%	3	0%	0%	100%	0%	N	N		
Washington	8	14	71%	29%	10	10%	70%	20%	0%	4	25%	50%	25%	0%	N	N		
Washington	10	23	74%	26%	17	24%	71%	6%	0%	6	0%	33%	50%	17%	N	Y		

NOTE:

1. Criterion A and Criterion B are applied to tests with both MC and OE items. Criterion MC and Criterion OE refer to the criterion applicable to tests with MC items only or OE items only, respectively.
2. Y = A test met the criteria for a deeper learning assessment; N = otherwise.
3. NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

Table B.3. Percentage of Released State Reading Test Items Rated at Each DOK Level, by State and Grade Level, and Deeper Learning Assessment Classification Results

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
California	7	2	0%	100%	0	/	/	/	/	2	0%	0%	0%	100%				Y[2]
Colorado	3	1	0%	100%	0	/	/	/	/	1	0%	100%	0%	0%				N
Colorado	4	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
Colorado	5	2	0%	100%	0	/	/	/	/	2	0%	50%	50%	0%				N
Colorado	6	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
Colorado	8	1	0%	100%	0	/	/	/	/	1	0%	100%	0%	0%				N
Delaware	5	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
Delaware	8	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
Delaware	10	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
Kentucky	4	2	0%	100%	0	/	/	/	/	2	0%	0%	0%	100%				Y
Kentucky	7	2	0%	100%	0	/	/	/	/	2	0%	0%	0%	100%				Y
Kentucky	12	2	0%	100%	0	/	/	/	/	2	0%	0%	0%	100%				Y
Missouri	3	1	100%	0%	1	0%	0%	100%	0%	0	/	/	/	/			N	
Missouri	7	1	0%	100%	0	/	/	/	/	1	0%	0%	0%	100%				Y
Missouri	11	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
NECAP[3]	5	14	100%	0%	14	57%	14%	21%	7%	0	/	/	/	/			Y	
NECAP	8	14	100%	0%	14	64%	7%	21%	7%	0	/	/	/	/			Y	
NECAP	11	1	100%	0%	1	0%	0%	0%	100%	0	/	/	/	/			Y	
New Jersey	3	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
New Jersey	4	1	0%	100%	0	/	/	/	/	1	0%	0%	100%	0%				N
New Jersey	8	2	0%	100%	0	/	/	/	/	2	0%	0%	100%	0%				N
New Jersey	12	2	0%	100%	0	/	/	/	/	2	0%	0%	100%	0%				N

		All items			MC Items					OE Items					Criterion[1]			
State	Grade	N	MC	OE	N	DOK1	DOK2	DOK3	DOK4	N	DOK1	DOK2	DOK3	DOK4	A	B	MC	OE
Ohio	4	11	91%	9%	10	100%	0%	0%	0%	1	0%	0%	100%	0%	N	N		
Ohio	7	7	100%	0%	7	29%	57%	14%	0%	0	/	/	/	/	N	N		
Ohio	12	13	77%	23%	10	60%	20%	20%	0%	3	0%	33%	67%	0%	N	N		
Texas	4	29	97%	3%	28	50%	50%	0%	0%	1	0%	0%	0%	100%	N	Y		
Texas	7	40	100%	0%	40	60%	40%	0%	0%	0	/	/	/	/			N	
Texas	10	20	100%	0%	20	80%	15%	5%	0%	0	/	/	/	/			N	
Texas	12	20	100%	0%	20	70%	30%	0%	0%	0	/	/	/	/			N	

NOTE:

1. Criterion A and Criterion B are applied to tests with both MC and OE items. Criterion MC and Criterion OE refer to the criterion applicable to tests with MC items only or OE items only, respectively.
2. Y = A test met the criteria for a deeper learning assessment; N = otherwise.
3. NECAP is administered in four states, including Maine, New Hampshire, Rhode Island, and Vermont.

References

Darling-Hammond, Linda, and Frank Adamson, *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*, Stanford, Calif.: Stanford Center for Opportunity Policy in Education, Stanford University, 2010.

Goertz, Margaret E., Leslie Nabors Olah, and Matthew Riggan, *Can Interim Assessments Be Used for Instructional Change?* Philadelphia, Pa.: Consortium for Policy Research in Education, 2009.

Great Lakes West Comprehensive Center, *Overview of Selected State Assessment Systems*, Naperville, Ill., January 2009.

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 20, 2012: http://www.nciea.org/beta-site/publication_PDFs/cognitiverigorpaper_KH11.pdf

Koretz, Daniel, Brian Stecher, Stephen Klein, and Daniel McCaffrey, "The Vermont Portfolio Assessment Program: Findings and Implications," *Educational Measurement: Issues and Practice,* Vol. 13, No. 3, Fall 1994, pp. 5–16.

Le, Huy, Alex Casillas, Steven B. Robbins, and Ronelle Langley, "Motivational and Skills, Social, and Self-Management Predictors of College Outcomes: Constructing the Student Readiness Inventory, *Educational and Psychological Measurement,* Vol. 65, No. 3, June 2005, pp. 482–508.

Maryland State Department of Education, *Maryland School Assessment: Mathematics Public Release Items (Grade 4)*, 2009. As of February 20, 2012: http://mdk12.org/share/msa_publicrelease/2008/2008MatMSAGr4.pdf

Matsumura, Lindsay C., 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,* No. 669, Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, 2006.

Matsumura, Lindsay C., Helen E. Garnier, Sharon Cadman Slater, and Melissa D. Boston, "Toward Measuring Instructional Interactions "At-Scale," *Educational Assessment,* Vol. 13, No. 4, October 2008, pp. 267–300.

Mitchell, Karen, Jamie Shkolnik, Mengli Song, Kazuaki Uekawa, Robert Murphy, Mike Garet, and Barbara Means, *Rigor, Relevance, and Results: The Quality of Teacher Assignments and Student Work in New and Conventional High Schools*, Washington, D.C.: American Institutes for Research, July 2005.

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, October 1998.

Ohio Department of Education, *Ohio Achievement Tests: Reading Grade 4*, undated. As of February 20, 2012: http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=4681

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

Reckase, Mark D., "Portfolio Assessment: A Theoretical Estimate of Score Reliability," *Educational Measurement: Issues and Practice*, Vol. 14, No. 1, September 1995, pp. 12–14.

Rhode Island Department of Elementary and Secondary Education, *New England Common Assessment Program Released Items: Grade 8 Writing,* 2010. As of February 20, 2012: http://www.ride.ri.gov/assessment/DOCS/NECAP/2010_Items/Gr8/NECAP_2010_Gr8_Writing_Released_Items.pdf

Rockford Public Schools, "4th Grade Extended Response: Mathematics," 2009–2010. As of February 20, 2012: http://webs.rps205.com/curriculum/k5math/files/B38EDB0AF01641FD943E9BD405C7F26B.pdf

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

State of Washington Office of Superintendent of Public Instruction, *Washington Assessment of Student Learning: Mathematics Grade 4*, undated.

Stecher, Brian, *Performance Assessment in an Era of Standards-Based Educational Accountability*, Stanford, Calif.: Stanford Center for Opportunity Policy in Education, Stanford University, 2010.

Texas Education Agency, *Texas Assessment of Knowledge and Skills: Grade 4 Writing*, March 2009. As of February 20, 2012: http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147499164&libID=2147499161

Webb, Norman L., *Alignment of Science and Mathematics Standards and Assessments in Four States*, Research Monograph No. 18, Washington, D.C.: Council of Chief State School Officers, 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 Officer, 2002a.

———, "Depth-of-Knowledge Levels for Four Content Areas," March 28, 2002b. As of August 24, 2011: 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, A Review of Model School Networks That Emphasize Deeper Learning, Santa Monica, Calif.: RAND Corporation, PM-3638-WFHF, 2010.

Zimmerman, Barry J., and Manuel Martinez-Pons, "Student Differences in Self-Regulated Learning: Relating Grades, Sex, and Giftedness to Self-Efficacy and Strategy Use," *Journal of Educational Psychology*, Vol. 82, No. 1, March 1990, pp. 51–59.