Working Paper

Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items Through the State Achievement Tests

KUN YUAN & VI-NHUAN LE

RAND Education

WR-967-WFHF
November 2012
Prepared for the William and Flora Hewlett Foundation

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 was 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 Selected Deeper Learning Skills

The manner in which students are assessed on the state exams restricted our analysis to two types of deeper learning skills: 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.

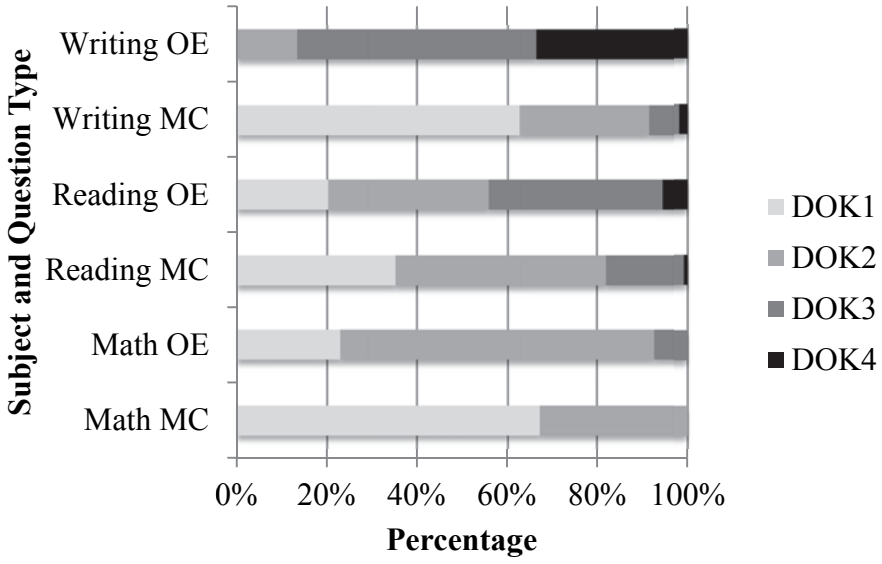
FINDINGS

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 Selected 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

As described above, the types of assessments analyzed and the range of deeper learning skills evaluated in this study were limited due to the requirements on access to both test items and the number of students taking the test and the current status of the assessment landscape for deeper learning skills. The criterion used to define what counts as deeper learning also has limitations in its capacity to capture deeper learning comprehensively. Moreover, cognitive rigor represents only one dimension of deeper

learning. Thus, the study enacted is somewhat different from the one envisioned at the beginning.

In addition, 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 and resource constraints did not allow us to analyze mastery of core content, collaboration, oral communication, or learn-how-to-learn skills. Although omitting these 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.

NEXT STEPS FOR THE FOUNDATION

This study has implications for the Foundation, as it moves to gauge progress towards the Deeper Learning Initiative's goal of increasing the percentage of students assessed on deeper learning skills to at least 15% by 2017. First, due to a number of constraints with the state achievement tests, we were able to assess only limited aspects of deeper learning. Future evaluations of the Deeper Learning Initiative may encounter the same types of challenges as this benchmark study, such that only a limited type of deeper learning skills can be examined.

Second, given the lack of credible standardized measures of intrapersonal or interpersonal competencies within the Deeper Learning Initiative and information about the number of students taking such measures, the Foundation may not be able to assess the full range of deeper learning skills outlined in the Deeper Learning Initiative without making trade-offs with respect to psychometric properties, costs, and other considerations.

Third, because of the interdependence between critical thinking and problem solving skills and fluency with the core concepts, practices, and organizing principles that constitute a subject domain, it is necessary to develop an analytic framework that would allow an analysis of the mastery of core conceptual content as integrated with critical thinking and problem solving. Although this task was beyond the scope of the time and resources available for this study, future studies examining the Foundation's Deeper Learning Initiative should consider frameworks that define fundamental concepts and knowledge for each subject area.

Finally, it is unclear whether there are any conceptual frameworks that can account for interpersonal competencies such as complex communication and collaboration, and intrapersonal competencies such as persistence and self-regulation. Because these dimensions were not included on the state tests, we did not examine any frameworks that assessed these skills. However, considering the lack of standardized measures that assess these skills, there is also likely to be a dearth of analytic frameworks that can be used to study these competencies. As curriculum, instructional practices, and assessment methodologies evolve that are responsive to a full range of deeper learning competencies, the analytic frameworks will need to evolve as well.