edmentum™

Predictive Validity Paper

Predictive Validity of Edmentum Study Island Benchmark Assessments

Predicting End-of-Year State Test Scores

Study Island®

edmentum™

Introduction

The purpose of this paper is to report on the findings of the quality review of the Edmentum Study Island Benchmark Assessments and, in particular, to highlight a study of the predictive validity of the benchmarks to end-of-year state test results.

The detailed results presented here are offered as empirical evidence that the Study Island math and English language arts (ELA) benchmark tests are statistically strong predictors of end-of-year state test scores. The magnitude of the validity coefficients are very high (.70s–.80s), which suggests the benchmarks are well suited to the intended criterion, and teachers using the benchmarks can be assured that they offer meaningful progress indicators toward end-of-year performance on summative assessments.

The design of this report is carefully aligned to the Standards for Supporting Documentation from the latest edition of the *Standards for Educational and Psychological Testing (2014)*, and it contains technical information for the Study Island benchmarks for math and English language arts for grades 3 through 8. According to Standard 7.0, "Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which tests to use for a specific purpose, how to administer the chosen test, and how to interpret test scores" (American Psychological Association (APA), National Council on Measurement in Education (NCME), Joint Committee on Standards for Educational and Psychological Testing, p. 125).

The detailed findings in this report demonstrate that Study Island benchmark tests are a statistically strong indicator of end-of-year state test scores. This data analysis shows that all grades (3–8) and all subjects (English language arts and math) included in the study demonstrate high validity coefficients in the range of .70s and .80s.

Study Island Benchmark Test Overview

The National Council on Measurement in Education uses the following definition for benchmark assessment:

Short assessments used by teachers at various times throughout the school year to monitor student progress in some area of the school curriculum. These tests also are known as interim assessments (Glossary of Important Assessment and Measurement, n.d.).

The Study Island benchmark tests are composed of fixed sets of items. Typically, the tests are administered online to students who are monitored by a classroom teacher. The benchmarks are a set of four forms per grade level. They are designed to be taken periodically throughout the school year. The tests are designed to mirror the structure and item formats found in state assessments. The result of each benchmark test is expected to reflect how students would perform on high-stakes assessments (i.e., predictive validity) and to target areas for instructional support. Educators can use data from the Study Island benchmark tests to allow for more efficient use of classroom time and resources.

Study Island Benchmark Test Development Process

Edmentum's rigorous test development process involves several stages, including but not limited to the following:

1. verification of validity and reliability, according to the identified purpose of the test
2. alignment of assessment content and depth of knowledge parameters with state test blueprint and standards
3. building of test forms with high item effects and construct validity
4. matching of distributions of item types, as well as grade-level form difficulty within each set of four grade-level forms

Subject matter experts from across the United States work with Edmentum curriculum specialists and writers to create test specifications for each subject area of the Study Island benchmark tests by synthesizing information from state and national standards. Subject matter experts take the following steps to define each test structure:

1. **Analyze state test blueprints.** Curriculum specialists analyze the state test blueprints to identify the structure of the test, the reporting categories, distribution of items, and overall test design to understand the requirements for the Study Island benchmark.
2. **Draft test designs.** The Study Island blueprint is aligned to the state test blueprint. Subject matter experts create a design for the Study Island benchmarks, making sure to include key topics and objectives found in the state standards. If state test blueprints are not available, objectives are organized into reporting categories using the breakdown of strands, domains, and clusters in the standards.
3. **Develop test items.** Test items are developed to meet the stated objectives. Subject matter experts define the approximate number of test items within each objective.
4. **Determining the test framework.** The test specifications are analyzed and recorded in the test framework. The framework represents the target objectives for testing.
5. **Determining reporting categories.** Reporting categories are based on frameworks and test specifications. These reporting categories represent organized, summative headings of the objectives recorded in the frameworks. Subject matter experts define reporting category titles and objectives within each reporting category.
6. **Designing student reports.** Reporting categories serve as the basis of the student reports that indicate progress and needs, as determined by the benchmark score.

Predictive Validity of Study Island Benchmark Tests

"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests." (American Psychological Association, 2014, p. 11)

Predictive validity describes how well an assessment predicts a student's performance in the future. When the alignment of learning standards and assessments is sound, then the extent to which one test score can predict another is elevated. The relationship between the two test scores can be called predictive or criterion validity.

"Predictive validity is determined by calculating the correlation coefficient between the results of the assessment and the subsequent targeted behavior. The stronger the correlation between the assessment data and the target behavior, the higher the degree of predictive validity the assessment possesses (Clause, n.d.)."

The correlations between the benchmark test scores and the state test scores provide evidence of the predictive validity. Predictive correlations can range from 0 to $^{+/-}1$. The higher the correlation coefficient, the more the two tests measure the same construct, increasing the chance of accurately predicting one from the other. In general, tests, such as the Study Island benchmark tests, are said to have predictive validity when they demonstrate their effectiveness in predicting the criterion or, in this case, in predicting the end-of-year state test scores.

In this particular examination of Study Island's benchmark tests, predictive validity was studied for English language arts and math in grades 3 through 8. The study was performed at the request of a school district that administered Study Island benchmark tests congruently with a statewide accountability measure. For the sake of maintaining confidentiality, the district will not be identified in this paper. To provide context, some characteristics of the district sample are presented below.

The district includes approximately 1,200 students with one elementary school (grades K–4) with 450 students; one middle school (grades 5–8) with 402 students; and one high school (grades 9–12) with 385 students. The race/ethnicity of the school district consists of the following: 94% white, 3% Hispanic, and 3% Black, Asian, and American Indian. One hundred percent of the teachers are designated as "highly qualified" by the state department of education. The percent of identified gifted and talented students is less than 1%. The identified special education students account for 13% of all students. Students who are eligible for free or reduced-price lunch represent 42% of the district.

A sample of 574 English language arts and 574 math state test scores of unidentifiable students from the district was shared with Edmentum. Available benchmark test scores of grade level forms 1, 2, 3 and 4 were also extracted from the Edmentum database and were matched with the state test scores using an Edmentum student identification number. Using a cohort approach to analysis, only students with scores for all four benchmarks[1], or in the case of grade 8, three benchmarks, were included in the analyses.

The distribution of matching state scale scores and benchmark test percent correct scores by grade and content area (i.e., math and ELA) is provided in Table 1. A total of 57 and 30 students, respectively, were not included in the sample due to missing scores.

Table 1: Number of Students by Grade Level with Complete Set of State and Benchmark Test Scores for Math and ELA

Grade	Number of Matched Student Scores for Math	Number of Matched Student Scores for ELA
3	95	96
4	81	83
5	88	91
6	63	82
7	98	103
8	92	89
Total Sample	517	544

Results

Math

Table 2 provides the Pearson product-moment correlations between the Study Island math benchmark test scores and the state math test scores by grade level. All of the correlations are statistically significant at the .01 level. For math, the correlations range from .594 for grade 8 students on benchmark 1, usually taken at the beginning of the school year, to .862, also for grade 8 students, on benchmark 3, taken at the end of the school year[2].

[1] *Grade 8 students were not assigned to take benchmark 4.*
[2] *Benchmark 3 was the last benchmark taken by grade 8 students before the state test was administered.*

edmentum™

Table 2: Predictive Validity Correlation Coefficients
by Grade Level Cohort for Math

Correlation Coefficients* for a State** Math Test and Study Island Benchmark Scores By Grade School Year 2013-2014					
	Percent of Study Island Benchmark Items Correct				
Grade	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4	Sample Size
3	.692	.609	.737	.753	n=95
4	.764	.766	.821	.830	n=81
5	.647	.810	.835	.831	n=88
6	.603	.688	.827	.710	n=63
7	.717	.720	.810	.800	n=98
8	.594	.732	.862	-	n=92
*All correlations are significant at the .01 level (2-tailed).					

** The state is not identified to maintain the anonymity of the district. See the description of the district demographics in a previous section.

In general, the predictive validity of the math benchmark tests increases from benchmark 1 to benchmark 4. For example, for grade 4, shown in red, the predictive validity of benchmark 1, taken in the fall, is .764; benchmark 2 is .766; benchmark 3 is .821; and benchmark 4 is .830. One possible, logical interpretation of the increases is that students have learned more standards-related material across the school year, so the probability of predicting the state score from the benchmark score increases.

Illustrative Interpretation

Example: What does a correlation of .83 mean?

A correlation coefficient does not have a simple interpretation, but researchers like to square the correlation coefficient to yield an interesting and interpretable quantity. When we square .83, we arrive at .69. The number .69 indicates that the predictor (benchmark) and the criterion (end-of-year test) share 69% of their variance in common.

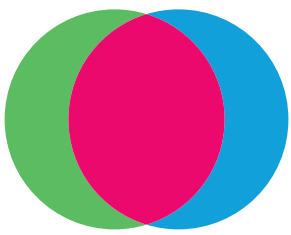


Figure 1. 69% overlap

Figure 2. 20% overlap

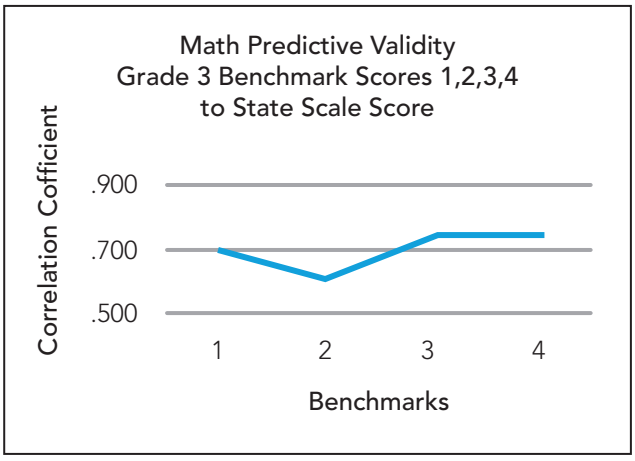
The overlap region of the Venn diagram in Figure 1 shows what 69% shared variance looks like. For purposes of comparison, Figure 2 shows what 20% shared variance looks like. This is the degree of shared variance implied by a correlation of .45. In this way, we can interpret what a "good enough" correlation is.

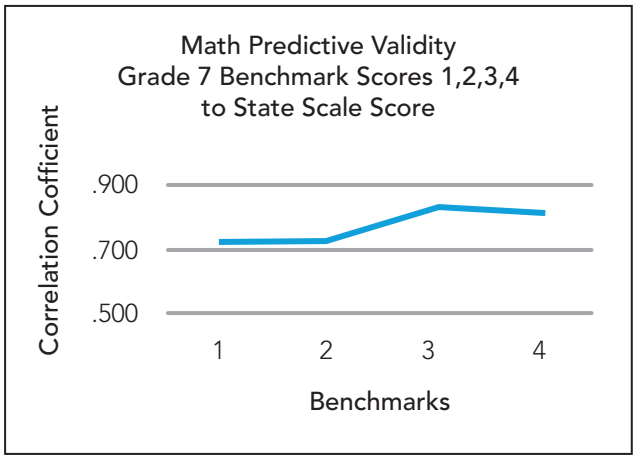
What does it mean to share 69% of variance in common? It means that both the benchmarks and the state tests show variability in student scores. There are individual differences. We infer that an underlying ability makes those scores vary from student to student. The Venn diagram in Figure 1 means that the overlap region is what the benchmark and the state test capture in common. We infer that this underlying trait is in fact a student's true ability in the specified domain.

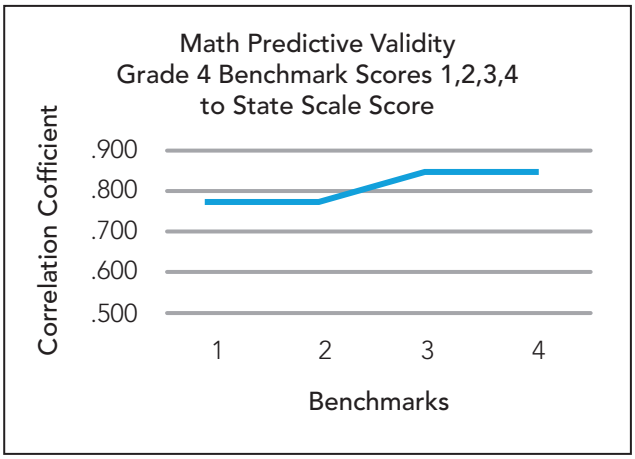
Math Validity Coefficients over Time

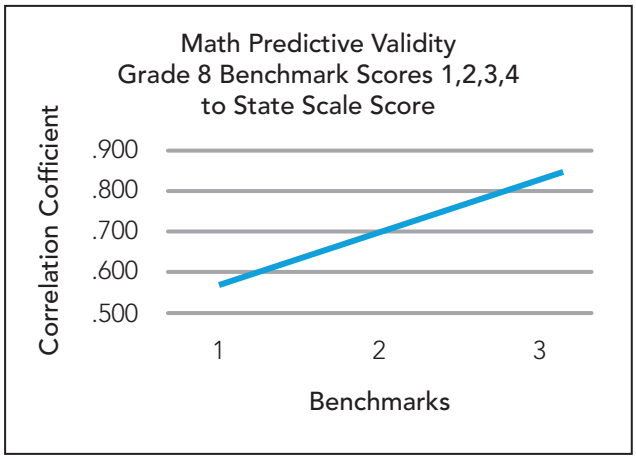
The four graphs in Figure 3 show a sample of math correlation coefficients by grade, across the school year. Although each benchmark test may not have been administered on exactly the same dates, the administration schedule tends to be spread out evenly across time, beginning in fall and ending a month or so before the end-of-year state test is administered.

Figure 3. Math correlation coefficients by grade









It is important not to over-interpret the general upward trend of the correlation between the benchmarks and the end-of-year test. However, it is noted that one would expect a population of students on a learning trajectory to acquire an increasingly coherent mastery of the skills. This expectation might explain the increasing correlation between the touchpoints along the way and the fully developed skill construct measured at termination.

English Language Arts

The correlations between the benchmark and the state test scores for English language arts are similar to those for math and are, on average, higher (.63 and .76, respectively). All of the correlations are statistically significant at the .01 level, ranging from .688 for grade 7, benchmark 1, to .836 for grade 6, benchmark 2.

Table 3: Predictive Validity Correlation Coefficients by Grade Level Cohort for ELA

Correlation Coefficients* for State** ELA Test and Study Island Benchmark Scores By Grade School Year 2013–2014					
Percent of Study Island Benchmark Items Correct					
Grade	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4	Sample Size
3	.693	.808	.775	.791	n=96
4	.735	.801	.726	.748	n=83
5	.810	.800	.765	.801	n-91
6	.773	.836	.721	.788	n=82
7	.688	.770	.750	.727	n=103
8	.757	.698	.778	-	n=89

*All correlations are significant at the .01 level (two-tailed).

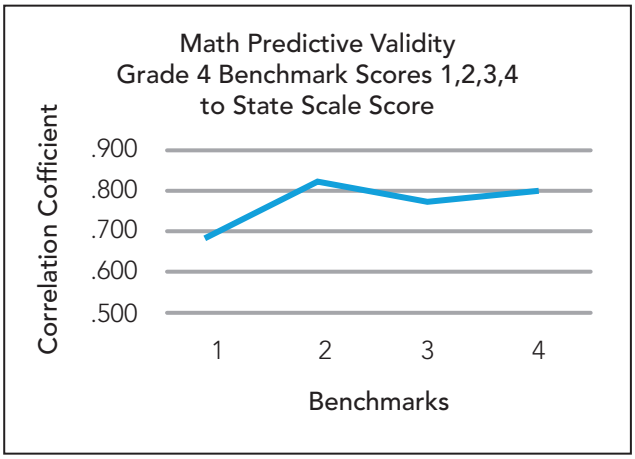
**The state is not identified to maintain the anonymity of the district. See the description of the district demographics in a previous section.

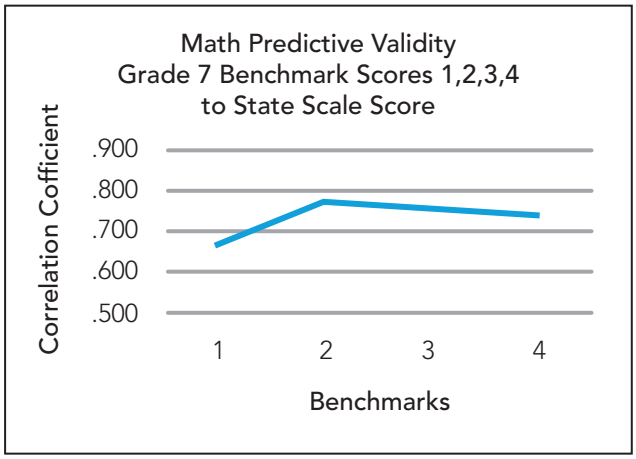
In general, the predictive validity of the math benchmark tests increases from benchmark 1 to benchmark 4. For example, The correlation coefficients for ELA appear to be more stable than the math correlations across benchmarks in terms predictive validity, with the majority falling in the mid .7 to low .8 range. Thirteen out of twenty-three correlations (57%) fall in the .7 range. Another interesting observation is that the correlations for benchmark 2 are higher than benchmarks 1, 3, and 4, except for grades 5 and 8. One possible explanation for benchmark 2 acting as the best predictor could be a better alignment of the standards tested in benchmark 2 compared to the standards tested on the state test. Although the overall blueprints of the state and benchmark tests are matched, a match of the item difficulty levels could add yet another dimension to the precision of the prediction. In all cases, except for grade 5, benchmark 4 is a better predictor of the state test score than benchmark 1.

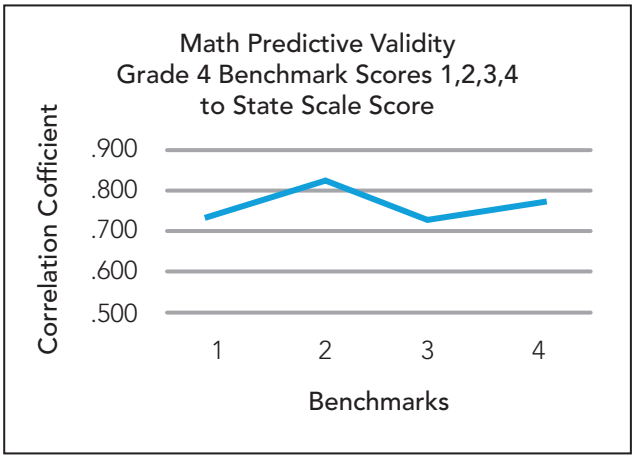
ELA Validity Coefficients Over Time

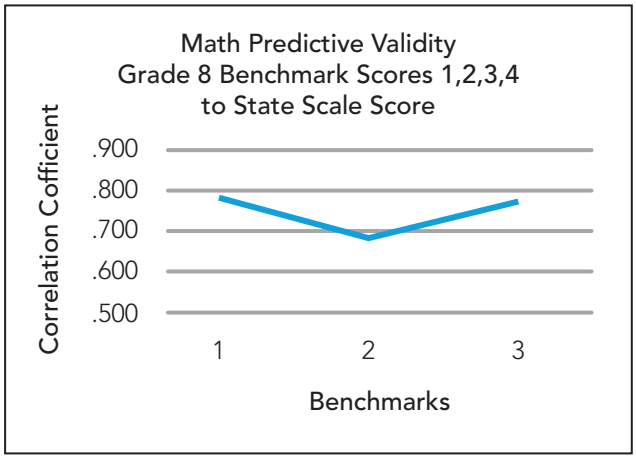
Figure 4 below shows a sample of ELA correlation coefficients by grade, across the school year.

Figure 4. English language arts correlation coefficients by grade









Alignment of Study Island Assessment Content and Depth of Knowledge Parameters with State Standards

Study Island benchmark tests are customized to match the blueprints of the state tests. This blueprint matching is aligned at the level of the reporting categories of the standards. Typically, tests do not cover every standard that students are expected to learn; most likely, state tests focus on what the state determines to be key standards categories. Thus, the alignment of Edmentum tests to the state's blueprint, which is public information, is critical for supporting students for test taking and educators for instructional planning.

The distribution of emphasis for ELA benchmark reporting categories is shown in Tables 4 and 5.

Table 4
Study Island's Blueprint Match for ELA, Grades 3–5

Reporting Category	Grade 3		Grade 4		Grade 5	
	Items	%	Items	%	Items	%
Reading Standards for Literature	7	23	7	23	7	23
Reading Standards for Informational Texts	7	23	7	23	7	23
Writing Standards	4	13	4	13	4	13
Language Standards	12	40	12	40	12	40
Total	30	100	30	100	30	100

Table 5
Study Islands Blueprint Match for ELA, Grades 6–8

Reporting Category	Grade 6		Grade 7		Grade 8	
	Items	%	Items	%	Items	%
Reading Standards for Literature	7	23	7	23	7	23
Reading Standards for Informational Texts	7	23	7	23	7	23
Writing Standards	4	13	4	13	4	13
Language Standards	12	40	12	40	12	40
Total	30	100	30	100	30	100

edmentum™

Key Metrics Used in the Benchmark Development Process

Item Discrimination Statistics

If an item discriminates well, getting an item correct should be correlated with the total score. Item discrimination is an essential item statistic in the test development process. Item discrimination is calculated using a point biserial correlation coefficient and ranges from 0 to +/-1. The point biserial compares each item, by itself, to the sum total of the rest of the items on that test.

Item discrimination "is often referred to as Item Effect since it is an index of an item's effectiveness at distinguishing those who know the content from those who do not" (Ramsay, n.d.)

If a student gets an item correct but does not know the content, there is a problem with the test. There is no hard-and-fast rule for interpreting a point biserial correlation. In general, psychometricians say that a point biserial coefficient of .2 to .3 and higher indicates a good item. All of the Study Island test forms show an average correlation of .3 or higher, as shown in Tables 6 and 7.

Table 6
Sample Size and Average Point Biserial Correlations by Grade and Form for Study Island Math Benchmarks

Grade	Form	Number of Students in Sample	Average Point Biserial Correlation Coefficient
3	1	34,501	.42
	2	24,702	.37
	3	17,930	.39
	4	7,997	.42
4	1	38,477	.36
	2	27,722	.37
	3	17,992	.39
	4	8,449	.42
5	1	39,297	.36
	2	34,627	.37
	3	18,179	.43
	4	7,885	.41

6	1	30,838	.35
	2	20,687	.35
	3	14,026	.36
	4	6,551	.36
7	1	27,201	.31
	2	18,207	.34
	3	11,496	.35
	4	4,950	.36
8	1	25,146	.35
	2	16,218	.37
	3	10,052	.38
	4	4,221	.37

Table 7
Sample Size and Average Point Biserial Correlations
by Grade and Form for Study Island ELA Benchmarks

Grade	Form	Number of Students in Sample	Average Point Biserial Correlation Coefficients
3	1	33,954	.42
	2	25,509	.42
	3	18,437	.43
	4	8,383	.43
4	1	37,330	.42
	2	29,648	.43
	3	20,316	.44
	4	8,976	.42

5	1	38,046	.44
	2	29,487	.44
	3	19,839	.44
	4	8,507	.44
6	1	32,773	.41
	2	21,804	.44
	3	13,221	.43
	4	7,132	.45
7	1	30,340	.43
	2	19,662	.41
	3	12,544	.43
	4	5,751	.44
8	1	28,748	.41
	2	18,543	.42
	3	11,654	.42
	4	4,858	.44

Comparable Content and Comparable Item Difficulty

All of the items are multiple choice. This particular item type distribution matches the benchmark tests used for the data analysis reported in this paper.

The item and form development described earlier in the Study Island Benchmark Test Development Process section includes an extensive item difficulty review with the following considerations:

- Content coverage and match to test design. Test developers complete an initial sorting of items into sets based on a balance of the reporting categories across forms as well as a match to the master test specifications.
- Visual balance. Item sets are reviewed to ensure similar length and density of items (e.g., length and complexity of reading selections, number of graphics).
- Response option balance. Each item set is checked to verify that it contains a roughly equivalent number of answer options (A, B, C, and D).
- Item difficulty and complexity. Difficulty of items is determined through a combination of expert judgment and the use of field tested item statistics. The resulting benchmark assessment forms are designed to be similar in level of difficulty and complexity across the forms. Summaries of average item difficulties per form, by content area, are presented in Tables 8 and 9.

Table 8
Study Island Math Benchmark Form Comparability

Grade	Form	Number of Students in Sample	Average Item Difficulty per Form
3	1	34,501	.62
	2	24,702	.62
	3	17,930	.71
	4	7,997	.69
4	1	38,477	.54
	2	27,722	.61
	3	17,992	.61
	4	8,449	.65
5	1	39,297	.43
	2	34,627	.51
	3	18,179	.61
	4	7,885	.64
6	1	30,838	.49
	2	20,687	.51
	3	14,026	.59
	4	6,551	.53
7	1	27,201	.40
	2	18,207	.40
	3	11,496	.53
	4	4,950	.47
8	1	25,146	.42
	2	16,218	.49
	3	10,052	.56
	4	4,221	.51

Table 9
Study Island ELA Benchmark Form Comparability

Grade	Form	Number of Students in Sample	Average Item Difficulty per Form
3	1	33,954	.51
	2	25,509	.61
	3	18,437	.65
	4	8,383	.58
4	1	37,330	.57
	2	29,648	.64
	3	20,316	.70
	4	8,976	.64
5	1	38,046	.66
	2	29,487	.68
	3	19,839	.71
	4	8,507	.72
6	1	32,773	.62
	2	21,804	.69
	3	13,221	.68
	4	7,132	.75
7	1	30,340	.70
	2	19,662	.71
	3	12,544	.68
	4	5,751	.72
8	1	28,748	.67
	2	18,543	.68
	3	11,654	.70
	4	4,858	.68

In the school year 2014–2015, Edmentum's analysis on form comparability was updated to reflect the change in item types and a consistent representation of standard and Depth of Knowledge (DOK) across forms. In cases where a single standard is broad and encompasses several concepts, the same concept is assessed on each form. The new Edmentum benchmark test forms are based on a combination of content similarity and test specifications as well as pretested item difficulty data to improve the statistical match of the overall difficulty level of the benchmark assessment forms. Equating studies are now in progress to measure the degree of comparability of the most recently written benchmark forms. This process will continue, as necessary, when new items are written, when standards are revised, and when benchmarks need to be customized by state.

Conclusion

In conclusion, the Study Island benchmark tests are statistically strong predictors of end-of-year state test scores. This is true for both the math and English language arts areas, and it is true for all grade levels included in this sample (grades 3–8). There are three important implications of this finding.

First, as described in this paper, the process of test construction at Edmentum involves a rigorous alignment of item content to statutory state outcomes. This alignment can be considered good evidence that the test development process used by Edmentum is highly effective.

Second, the magnitudes of the correlations between the benchmarks and the state test used in this study are to be considered high by any standard. To achieve validity coefficients in the range of the .70s and .80s is quite unequivocal evidence that the benchmarks are hitting the criterion.

Third, as described, the Study Island benchmark tests share 69% shared variance with the state test used in this study. Such a high quantity of shared variance suggests that the two assessments are tapping what appears to be a common underlying trait. This evidence can assure teachers using the benchmarks that their efforts are aligned to the right criterion.

References

American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Benchmark Assessment. (n.d.). In *Glossary of Important Assessment and Measurement.* Retrieved October 25, 2015 from http://www.ncme.org/ncme/NCME/Resource_Center/Glossary.

Clause, C. (n.d.). Predictive validity in psychology: definition and examples. In *Intro to psychology: help and review.* Retrieved from http://study.com

Dube', Paul J. (2011). *Attempting to improve standardized test results using Study Islands' web-based mastery program* (Master's report) Michigan Technological University, Houghton, MI. Retrieved from http://digitalcommons.mtu.edu/etds/524/.

Magnolia Consulting (2008). *Study island scientific research base.* Retrieved from http://www.magnoliaconsulting.org/Study%20Island%20Foundational%20Report.pdf

Mississippi Department of Education (n.d.). *Webb's depth of knowledge guide.* Retrieved from http://www.aps.edu/re/documents/resources/Webbs_DOK_Guide.pdf

"The Depth of Knowledge (DOK) is a model designed to analyze the cognitive expectation demanded by standards, curricular activities and assessment tasks (Webb, 1997). The model is based upon the assumption that curricular elements may be categorized based upon the cognitive demands required to produce an acceptable response"(Mississippi Department of Education, n.d.)

National Center for Education Statistics. (2015, October 24). Common Core of Data. Public school district data for the 2012-2013, 2013-2014 school years. Retrieved from http:/nces.ed.gov/ccd/districtsearch.

Ramsay, C. (n.d.). *Item analysis*. Retrieved from Sites at Penn State: http://sites.psu.edu/itemanalysis.

Webb, N. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education.* (NISE Research Monograph No. 6). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

edmentum.com
800.447.5286
info@edmentum.com
0612-36 111815

2425 North Central Expressway
Suite 1000
Richardson, TX 75080

© 2015 EDMENTUM, INC.