DISCOVERY EDUCATION ASSESSMENT

Table of Contents

About Discovery Education Assessment

Assessment for Learning

ThinkLink Learning, founded by Vanderbilt University in 2000, became part of Discovery Education in 2006. **Discovery Education** provides engaging digital resources to schools and homes with the goal of making educators more effective, increasing student achievement, and connecting classrooms and families to a world of learning.

Discovery Education is a division of **Discovery Communications, LLC** the leading global nonfiction media company. The leader in digital video-based learning, Discovery Education produces and distributes high-quality digital resources in easy-to-use formats in all core-curricular subject areas. Discovery Education is committed to creating scientifically proven, standards-based digital resources for teachers, students, and parents that make a positive impact on student learning. Through solutions like Discovery Education streaming, Discovery Education Science, Discovery Education Health and **Discovery Education Assessment,** LLC. Discovery Education helps over one million educators and 35 million students harness the power of broadband and media to connect to a world of learning.

From the beginning, Discovery Education Assessment has focused on **assessment *for* learning** more than on assessment ***of*** learning (Black, Paul et al. 2004. Working inside the Black Box: Assessment of Learning in the Classroom. *Phi Delta Kaplan* 86, no. 1. pp. 8-21.) Our focus on research-based assessments that drive instruction is inherent in our **Predictive Benchmark** series which was conceptualized as a **teacher-centered, student-focused** assessment tool.

Discovery's assessments are **developed by teachers** for schools. Our founder was a teacher in an urban area. Our current vice president taught and was a principal. The directors of research and content have backgrounds that include teaching, school administration, district administration, state based school reform, and special education. The raw elements of the Discovery Education Assessment tools were initially mined from the school experiences of these varied teaching careers.

Foundations in Research

Tests are useful only to the degree they are used to improve schools. Perhaps no single publication has exceeded the impact of Marzano's "What Works in Schools" (2003, ASCD) in underscoring the importance of systematic school consensus around challenging goals and effective feedback at least every nine weeks on specific knowledge and skills for individual students (p. 180). The importance of data that parents and students can understand and use to assess individual academic growth can not be over estimated.

Schools have been responsible for systematic academic screening since EHA 1976. Since 1998, schools have been responsible for systematic school improvement keyed to state NCLB plans. To be most effective, school NCLB achievement targets simply must be tied to state specific NCLB assessment standards. The only way to consistently meet NCLB targets has been site based action research. School improvement is not a top-down administrative/curriculum process. It is founded in building problem solving teams that efficiently apply relevant data to improvement.

Student learning is best supported by informed teaching. All time spent testing or worse, practice testing, distracts from time students could spend learning. Teacher time spent testing, scoring and interpreting test results distracts from teaching and learning. Discovery assessments are designed to quickly and accurately produce reports that teacher can easily use to determine how to best use their class time, identify better instructional approaches, and gauge which students need additional support.

OUR ASSESSMENTS

Each series of Discovery Education Assessment Predictive Benchmark Assessments provides state specific screening data, using each state's curriculum standards and subskills for each test item. State specific predictive benchmark assessments are provided for students in grades three and above in the following states: **Alabama, California, District of Columbia, Florida, Illinois, Kentucky, Missouri, Mississippi, North Carolina, New York, Ohio, South Carolina, Tennessee, Virginia, Wisconsin**. A US test series developed to measure student performance on standards that generally used across the United States is sold in these states for grades that do take a high stakes state test. This test series is also used in Arkansas, New Mexico, and West Virginia and predicts performance on the state high stakes test. This series is used in other states based with mastery scoring. As larger numbers of schools participate in the US tests, assessments in those states are moved to state specific assessments.

Discovery Education Assessment in 2008-2009 will distribute over 1400 benchmark tests that will be used by over **1,000,000 students** in **18 states**. These benchmark tests have been used to improve instruction, help strengthen students' academic skills, and increase proficiency levels as measured under No Child Left Behind. Discovery Education Assessment subscribes to the *Standards for Educational and Psychological Testing* articulated by the consortium of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

Schools typically administer up to four predictive benchmark assessments per year. The predictive benchmark tests are designed to predict student performance on the next high-stakes test the student will experience. Many schools administer the first assessment during the first month of school and the fourth assessment during the last month of school, spanning

9 – 12 weeks between each benchmark assessment. The benchmark assessments are designed to be administered in a class period for grades 3-high school. Most districts specify that the tests should be summative; measure skills for an entire assessment year. In some states the assessment year begins on May 1, in others it begins on December 1.

Locally developed curriculum probes are frequently used for measuring student progress on skills targeted in the district curriculum map for a specific period (pacing tests). These can be specified by the district or by the school. Still more curriculum probes are sometimes selected by teachers to assess mastery of skills being taught at the classroom level to students who may be working on prerequisites to the state specified skill. Curriculum probes enable educators to customize the assessment to meet individual school needs and individual student needs. Curriculum probes are built from a pool of more than **30,000 items** that are mapped to state specific curriculum/assessment standards.

Test results from the Discovery Benchmark series are compared each year to specific state criterion referenced test results and demonstrate over 80% predictive validity in grades 2 through high school (highest grade depends on the state's testing structure). For each subject, a vertical scale is calculated using Rasch single parameter Item Response Theory and links all benchmark assessments; kindergarten through high school. This equal interval vertical scale is used to indicate academic growth both within and between years. When coupled with proficiency predictions, the vertical scale provides a dual discrepancy basis for academic screening.

Discovery Education Assessment produces detailed documents for each series of state benchmark tests entitled "What is Predictive Assessment?" These documents outline several technical criteria for these benchmark tests: content validity, test reliability, criterion validity, predictive validity, and consequential validity. Furthermore, the results of experimental and quasi-experimental studies showing the use of these benchmark tests to improve instruction are described. Technical criteria for tests and items are also enumerated: descriptions of vertical scaling techniques, item discrimination and difficulty indices, DIF criteria, and bias and sensitivity review material.

Our **R**eports

The Discovery Predictive Assessment Series provides a comprehensive approach to screening. **District staff** can compare schools and efficiently target instructional approaches proven to work in their particular schools. **Teachers** see how their classes are responding to various instructional approaches and progressing toward meeting state specific NCLB proficiency standards and levels. **Parents** and students see individual student strengths and opportunities and set targets for achievement. Problem solving teams have a solid set of data from which to recommend additional assessments, interventions, or tiers.

Discovery Education Assessment

Research

All reports are available for individual students, classes, grades within each school with district summaries and drill down options to provide efficient views of data for teachers, principals, district personnel and parents/students. By using the state's curriculum for each test, teachers have reports that support their daily instruction and guide changes.

The reports allow our Predictive Assessment Series to be used as a screener. Schools can screen students by proficiency prediction, by rate of academic growth, by specific skill, or by a combination of these. School level teams determine the best approach for their school and apply that approach to the benchmark data. Most schools use a combination of approaches to the data, along with anecdotal information from the classroom, in site based problem solving. For use with Response to Intervention (RTI), teachers and administrators can quickly identify general education instruction changes that are needed before specific student interventions are considered. Then, for each student, specific subskills are identified as problem areas that serve as the basis of intervention. Subskill gaps are identified by their state specific definitions and codes to facilitate RTI supports in addition to general education classroom strategies.

Occasional DIF studies have suggested that the tests are appropriate for use with diverse populations. Discovery staff members are trained to develop items that are fair and each item passes through many reviews. Most items have been systematically reviewed by teams of educators/customers purchasing the tests through large contracts. Individual teacher and administrators also regularly comment on items. This collective body of educator feedback produces changes in benchmark test items or item bank items as needed.

Schools are encouraged to administer the predictive benchmarks the same way they will administer the next state high-stakes test to each student. This includes IEP and 504 Plan adjustments to standard administration.

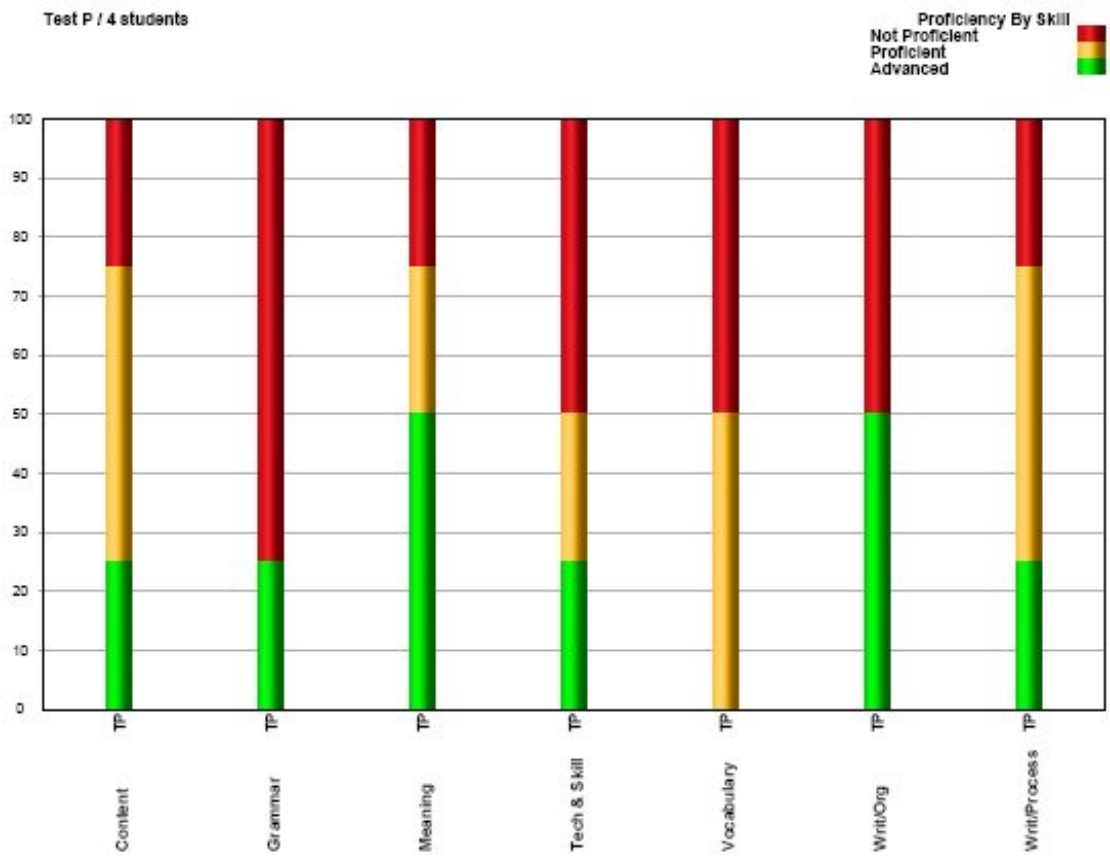
The following pages include some of the reports included in the Discovery Education Assessment reports. Each report is provided three or four times per year, following each benchmark assessment.

The **Class Summary Report** enables teachers to quickly determine which skills need to be emphasized in their class.

Class Summary Report

School: Demo TN

Test P of PAB (4 students)
Teacher: 6th grade Teacher
Class: Go Go Group 6 R (virtual class)
Grade: Grade 6
Subject: Reading/Language Arts



	Content	Grammar	Meaning	Tech & Skill	Vocabulary	Writ/Org	Writ/Process
Non	25.0%	75.0%	25.0%	50.0%	50.0%	50.0%	25.0%
Prof	50.0%	0.0%	25.0%	25.0%	50.0%	0.0%	50.0%
Adv	25.0%	25.0%	50.0%	25.0%	0.0%	50.0%	25.0%

Discovery Education Assessment
Research

The **Objectives and Subskills Report** allows teachers to rapidly identify the state specific skills and test items that need attention.
Administrators typically use the **Objectives and Subskills Report** to identify school-wide instructional targets and to properly allocate new instructional purchases and staff.

Objectives and Subskills Report

School: Demo TN

Test P of PAB (4 students)
Teacher: 6th grade Teacher
Class: Go Go Group 6 R (virtual class)
Grade: Grade 6
Subject: Reading/Language Arts

Q.#	Ans	Right		Wrong		Code	Reporting Category	Reporting Subcategory	Code	Reporting Category	Reporting Subcategory	Level
		#	%	#	%	TN State Reporting Categories 2008			CRT Reporting Categories			
Reading/Language Arts												
1	B	1	25	3	75	0601.8.2	Literature	Determine the main ideas of plots	6.1.21	Content	6.1.21 Main idea of a plot	Easy
2	A	2	50	2	50	0601.5.6	Logic, Informational Text	Identify case and effect relationships	6.1.16	Meaning	6.1.16 Cause/ effect	Easy
3	D	1	25	3	75	0601.8.1	Literature	Distinguish among various literary ge	6.1.20	Content	6.1.20 Distinguish literary genres	Mod.
4	C	1	25	3	75	0601.1.13	Language & Communicat	Use dictionaries, thesauruses, electron	6.1.11	Techniques and Skills	6.1.11 Locate information	Easy
5	D	2	50	2	50	0601.5.5	Logic, Informational Text	Draw inferences from selected texts	6.1.17	Meaning	6.1.17 Inferences	Mod.
6	B	3	75	1	25	0601.8.1	Literature	Distinguish among various literary ge	6.1.20	Content	6.1.20 Distinguish literary genres	Hard
7	C	2	50	2	50	0601.8.5	Literature	Determine the author's purpose	6.1.26	Content	6.1.26 Author purpose	Mod.
8	B	1	25	3	75	0601.3.12	Writing & Research	Select a concluding sentence	6.2.8	Writing/Organization	6.2.8 Appropriate concluding sentence	Hard
9	A	1	25	3	75	0601.3.2	Writing & Research	Identify the audience	6.2.4	Writing/Writing Process	6.2.4 Identify intended audience	Mod.
10	C	2	50	2	50	0601.5.4	Logic, Informational Text	Evaluate text for fact and opinion	6.1.9	Meaning	6.1.9 Fact/ opinion	Mod.
11	A	1	25	3	75	0601.6.3	Logic, Informational Text	Locate and verify information	6.1.15	Techniques and Skills	6.1.15 Supporting information	Hard
12	D	0	0	4	100	0601.3.10	Writing & Research	Rearranged in logical and coherent ord	6.2.6	Writing/Organization	6.2.6 Paragraph order	Mod.
13	C	2	50	2	50	0601.3.9	Writing & Research	Select appropriate title	6.2.2	Writing/Organization	6.2.2 Appropriate text title	Mod.
14	A	0	0	4	100	0601.1.12	Language & Communicat	Determine the meaning of unfamiliar w	6.1.1	Vocabulary	6.1.1 Word meaning from affixes, syllabic	Mod.
15	D	2	50	2	50	0601.6.1	Logic, Informational Text	Recognize that purpose determines te	6.1.12	Meaning	6.1.12 Purpose & text format	Hard

Just before high stakes tests, teachers tend to focus on the “bubble” skills (yellow). At the beginning of the year, teachers tend to work more with the skills that are clearly not proficient (red).

Discovery Education Assessment
Research

The **Student Report** allows teachers to first identify students who have overall poor performance in the subject (pink in the right column) and then to speedily describe the skills that most need attention.

Student Report

School: Demo TN

Discovery EDUCATION ASSESSMENT

Test P of PAB (4 students)
Teacher: **6th grade Teacher**
Class: **Go Go Group 6 R** (virtual class)
Grade: **Grade 6**
Subject: **Reading/Language Arts**

Subject Proficiency
Not Proficient (0-7 correct)
Proficient (8-15 correct)
Advanced (16-28 correct)

determined by # correct

	Content	Grammar	Meaning	Tech & Skill	Vocabulary	Writ/Org	Writ/Process		
Adams, J	Prof	Non	Non	Non	Non	Non	Prof	Non	3
Adams, J	Prof	Non	Adv	Prof	Prof	Adv	Prof	Prof	12
Berry, S F	Non	Non	Prof	Non	Non	Non	Non	Non	1
Carrey, A L	Adv	Adv	Adv	Adv	Prof	Adv	Adv	Adv	17

Many schools use the **Student Report** with students to help them set learning targets that will lead to good performance on end of course tests or the high stakes assessment.

Schools frequently use the **Student Report** with parents to assist them in understanding that formal assessments are clearly indicating success and/or opportunities to improve in specific areas.

These same easy to use reports are available to district and school administrators to design school improvement plans that work for students and to meet the requirements of NCLB. In today's push for maximized instructional time / achievement, each test must efficiently meet the data needs of administrators, teachers, students, and parents.

Overview of Standards and Scientifically-Based Evidence Supporting the Discovery Education Assessment Predictive Benchmark Test Series

Since its inception in 2000 by Vanderbilt University as ThinkLink Learning, Discovery Education Assessment has focused on the use of formative assessments to improve K-12 student learning and performance. Bridging the gap between university research and classroom practice, Discovery Education Assessment offers effective and user-friendly assessment products that provide classroom teachers and students with the feedback needed to strategically adapt their teaching and learning activities throughout the school year.

Discovery Education Assessment pioneered a unique approach to formative assessments using a scientifically research-based continuous improvement model that maps diagnostic assessments to each state's high stakes test. Discovery Education Assessment: *Predictive Benchmark* tests are aligned to the content assessed by each state test allowing teachers to track student progress toward the standards and objectives used for accountability purposes.

The subsequent sections detail the evidence Discovery Education Assessment has accumulated for each of the following quality testing standards:

1. **Are Discovery's *Predictive Benchmark* assessments reliable?**

   Test reliability provides evidence that test questions are consistently measuring a given construct, such as mathematics ability or reading comprehension. Furthermore, high test reliability indicates that the measurement error for a test is low.

2. **Do Discovery's *Predictive Benchmark* assessments have content validity?**

   Content validity evidence shows that test content is appropriate for the particular constructs that are being measured. Content validity is measured by (a) agreement among subject matter experts about test material and alignment to state standards, (b) highly reliable training procedures for item writers, (c) thorough reviews of test material for accuracy and lack of bias, and (d) examination of depth of knowledge of test questions.

3. **Do Discovery's *Predictive Benchmark* assessments match state standardized tests?**

   Criterion validity evidence demonstrates that test scores predict scores on an important criterion variable, such as a state's standardized test.

4. **Can Discovery's *Predictive Benchmark* assessments predict proficiency levels?**

   Proficiency predictive validity evidence supports the claim that a test can predict a state's proficiency levels. High accuracy levels show that a high degree of confidence can be placed in the vendor's prediction of student proficiency.

5. **Can the use of Discovery's *Predictive Benchmark* assessments improve student learning?**

Consequential validity outlines how the use of these predictive assessments facilitates important consequences, such as the improvement of student learning and student performance on state standardized tests.

6. **Can Discovery's *Predictive Benchmark* assessments be used to measure growth over time?**

   Growth models depend on a highly rigorous and valid vertical scale to measure student performance over time. A vendor's vertical scales should be constructed using advanced statistical methodologies such as Rasch measurement models and other state-of-the-art psychometric techniques.

7. **Are Discovery's *Predictive Benchmark* assessments based on scientifically-based research advocated by the U. S. Department of Education?**

   In the *No Child Left Behind Act of 2001*, the U.S. Department of Education outlined six major criteria for "scientifically-based research" to be used by consumers of educational measurements and interventions. Accordingly, a vendor's test

   (i) *employs systematic, empirical methods that draw on observation and experiment;*

   (ii) *involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;*

   (iii) *relies on measurements or observational methods that provide reliable and valid data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;*

   (iv) *is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition control.*

   (v) *ensures experimental studies are presented in sufficient detail and clarity to allow for replication or, at a minimum, offer the opportunity to build systematically on their finding;*

   (vi) *has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective and scientific review;*

TEST RELIABILITY

1. **Are Discovery's *Predictive Benchmark* assessments reliable?**

Test reliability provides evidence that test questions are consistently measuring a given construct, such as mathematics ability or reading comprehension. Furthermore, high test reliability indicates that the measurement error for a test is low. Reliabilities are calculated using Cronbach's alpha.

Table 1 through 6 present test reliabilities and sample sizes for six states—Florida, Tennessee, Kentucky, Illinois, New York, Alabama—that utilized Discovery Education Assessment *Predictive Benchmark* tests during the Spring 2008 test cycle in the subject areas of Reading, Mathematics, and Science.

The overall median Reading reliability across all six states was .85 with a median sample size of 6104. The overall median Mathematics reliability was .85 with a sample size of 5945. The overall median Science reliability was .75 with a median sample size of 5726.

Table 1: Florida Test Reliabilities for Reading, Mathematics, and Science Spring 2008.

Florida – Spring 2008 (Test B)						
	Reading	N	Mathematics	N	Science	N
Grade 2	.87	4,421	.86	4,513		
Grade 3	.87	5,232	.84	5,261	.79	1,566
Grade 4	.86	4,634	.86	4,593	.75	2,392
Grade 5	.86	4,607	.84	4,609	.71	4,251
Grade 6	.86	3,872	.83	4,076	.76	2,305
Grade 7	.85	4,112	.84	4,296	.79	2,430
Grade 8	.82	3,696	.86	3,863	.79	4,004
Grade 9	.83	3,265	.87	3,227		
Grade 10	.85	3,875	.87	3,377		
Grade 11					.78	4,442
Median	**.86**	**4,112**	**.86**	**4,296**	**.78**	**2,430**

Table 2: Tennessee Test Reliabilities for Reading, Mathematics, and Science Spring 2008.

Tennessee – Spring 2008 (Test B)						
	Reading	N	Mathematics	N	Science	N
Grade 3	.83	14,676	.83	18,679	.75	11,991
Grade 4	.82	19,256	.83	19,230	.75	12,874
Grade 5	.82	20,286	.82	20,692	.76	8,465
Grade 6	.81	18,533	.77	18,163	.78	14,211

Grade 7	.74	18,730	.83	18,874	.72	14,402
Grade 8	.83	18,703	.85	18,398	.74	13,832
Gateway	.89	5,746	.82	7,751	.86	3,358
Median	**.82**	**18,703**	**.83**	**18,679**	**.75**	**12,874**

Table 3: Kentucky Test Reliabilities for Reading, Mathematics, and Science Spring 2008.

Kentucky – Spring 2008 (Test B)						
	Reading	N	Mathematics	N	Science	N
Grade 3	.85	6,878	.86	6,775		
Grade 4	.86	6,593	.84	6,505	.80	6,369
Grade 5	.84	6,416	.84	6,486		
Grade 6	.86	7,727	.81	7,472		
Grade 7	.86	8,560	.86	8,537	.82	8,120
Grade 8	.84	8,723	.86	8,576		
Grade 9	.88	3,355	.81	2,977		
Grade 10	.84	3,313	.82	2,796		
Grade 11			.76	3,490		
Median	**.86**	**6,736**	**.84**	**6,505**	**.81**	**7,245**

Table 4: Illinois Test Reliabilities for Reading, Mathematics, and Science Spring 2008.

Illinois – Spring 2008 (Test B)						
	Reading	N	Mathematics	N	Science	N
Grade 3	.88	5,537	.85	5,360		
Grade 4	.85	5,902	.82	5,792	.75	1,106
Grade 5	.80	5,851	.82	5,814		
Grade 6	.84	5,472	.81	5,407		
Grade 7	.82	4,842	.81	4,714	.69	842
Grade 8	.85	4,803	.82	4,639		
Grade 11	.82	700			.75	173
Median	**.84**	**5,472**	**.82**	**5,384**	**.75**	**842**

Table 5: New York Test Reliabilities for Reading and Mathematics Spring 2008.

New York – Spring 2008 (Test B)				
	Reading	N	Mathematics	N
Grade 3	.88	140	.85	513
Grade 4	.84	218	.84	510

Grade 5	.84	332	.81	633
Grade 6	.79	289	.85	339
Grade 7	.85	164	.87	157
Grade 8	.82	273	.84	176
Median	**.84**	**246**	**.85**	**425**

Table 6: Alabama Test Reliabilities for Reading, Mathematics, and Science Spring 2008.

Alabama – Spring 2008 (Test B)						
	Reading	N	Mathematics	N	Science	N
Grade 2	.87	9,138	.86	8,814		
Grade 3	.89	16,496	.89	16,815	.69	5,871
Grade 4	.88	17,431	.88	17,472	.72	5,726
Grade 5	.86	17,207	.87	17,233	.74	10,291
Grade 6	.87	13,926	.88	13,536	.73	5,309
Grade 7	.89	12,888	.88	13,107	.69	7,313
Grade 8	.85	12,750	.88	12,103	.68	5,042
Grade 11	.82	2,561	.82	2,286	.71	447
Median	**.87**	**13,407**	**.88**	**13,322**	**.71**	**5,726**

CONTENT VALIDITY

2. **Do Discovery's *Predictive Benchmark* assessments have content validity?**

Content validity evidence shows that test content is appropriate for the particular constructs that are being measured. Content validity is measured by agreement among subject matter experts about test material and alignment to state standards, by highly reliable training procedures for item writers, by thorough reviews of test material for accuracy and lack of bias, and by examination of depth of knowledge of test questions.

To ensure **content validity** of all tests, Discovery Education Assessment carefully aligns the content of its assessments to a given state's content standards and the content sampled by the respective high stakes test. For customized contracts, Discovery Education Assessment also employs one of the leading alignment research methodologies, the **Webb Alignment Tool (WAT)**, which has supported the alignment of our tests to state specific content standards both in breadth (i.e., amount of standards and objectives sampled) and depth (i.e., cognitive complexity of standards and objectives). All Discovery Education Assessment tests are **state specific** and feature **matching reporting categories** of a given state's large-scale assessment used for accountability purposes.

Discovery Education Assessment typically completes two **content verification** steps following the initial blueprint development. Below is a summary of the procedural steps:

1. **Test blueprint development** by content experts that appropriately samples the state's assessment standards with comparable balance of representation and range of difficulty.
2. **First verification** of content validity via statistical analysis of previously tested items. Another verification of content validity via the research-validated Webb Alignment Tool (WAT) may occur for customized contracts.
3. **Second verification** of content validity by public school teachers and district personnel through subsequent use in the classroom.

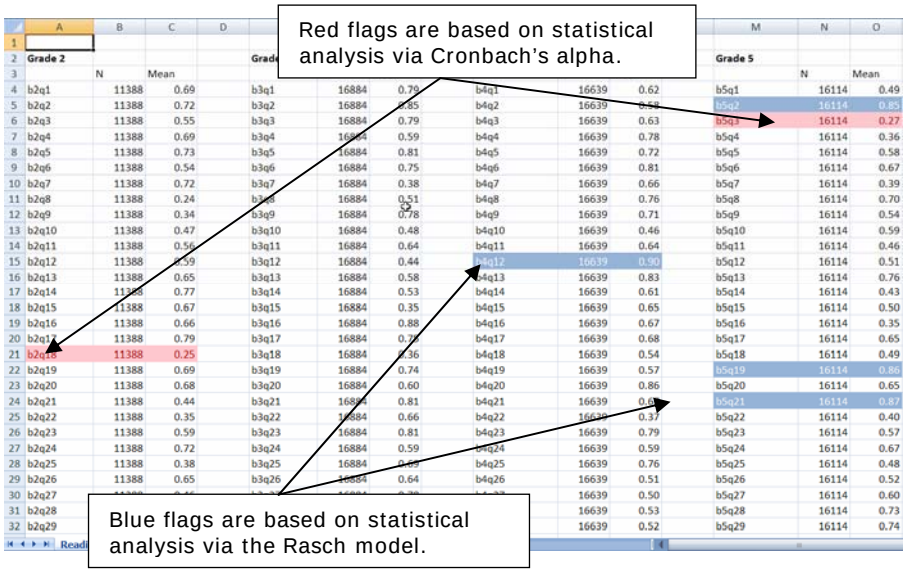
Discovery Education Assessment further employs rigorous quality standards during the **item writing and review processes**. All item writing at Discovery is completed by experienced teachers with familiarity in varied subjects, age groups, and ability levels. Item writers and test developers are certified subject matter experts with **graduate degrees** and a minimum of 3 years of **teaching experience**. All content experts have received supervised standardized test construction training and work on an ongoing basis with psychometric staff to systematically review and align items to state standards. The content review, copy editing, and quality control departments are also staffed by competent, qualified teachers with graduate degrees.

After completion of the initial state test blueprint, the Discovery test development team uses the online Assessment Manager to pull existing items from the Discovery item pool. Selected items match state assessment standards and have undergone previous field testing and/or feature actual use reliability statistics. Afterwards, item writers begin to develop new items based on the objectives and subskills put forth in the test blueprint.

Discovery Education Assessment
Research

Each test cycle is analyzed by psychometric staff to determine the p-value for each test item as well as overall test reliability. Items that do not meet the necessary psychometric criteria are removed or, when appropriate, rewritten. The **p-value** is typically referred to as the item difficulty index and indicates the proportion of examines who answered the item correctly. While it is standard procedure in educational testing to subject extreme p-values to item review, Discovery Education Assessment utilizes additional psychometric analyses such as **internal consistency reliability** measures and **Rasch modeling** to ensure customers high-quality assessments that yield reliable scores and valid test score inferences. Test reliability is measured via Cronbach's alpha, which represents a measure of internal consistency indicating to what extent a given item is measuring the same construct in relation to other items on the same test.

Discovery Education Assessment psychometric staff provides the test development team with state-of-the-art psychometric analyses to guide their item review process. The example below shows part of the information provided to the test development team. Red and blue items are flagged for review and receive close scrutiny by our content and test development experts.



Test developers and item writers guide their reviews along the following understanding:

- ***Red Items:*** Items are flagged red based on a measure of internal consistency indicating to what extent a given item is measuring the same construct in relation to other items on the same test. Red items are reviewed because, for some reason, they have failed to correlate sufficiently with other items measuring the same construct. Red items are confusing to students of both high and low ability levels. Red items may have an ambiguously worded stem, may have more than one correct answer, may have no correct answer, or may not match the content standards of a state. Red items can be revised, if the problem with the item can be identified. Otherwise, red items are replaced.
- ***Blue Items:*** Items are flagged blue based on the Rasch person-item map providing a graphic representation of person ability estimates and item difficulty estimates. Blue items are reviewed because they are too easy even for students of the lowest ability level.

Blue items are revised to make them harder. If this is not possible or appropriate, then blue items are replaced.

To further ensure **adequate content alignment** between Discovery Benchmark tests and state assessment standards, Discovery Education Assessment also utilizes Norman Webb's method of alignment and the Web Alignment Tool (WAT). Webb's alignment methodology has received sustained attention in the research literature. It is capable of assessing the breadth (i.e., matching topics) and depth (i.e., matching student expectations) of alignment between assessments and standards, and is commonly viewed as one of the "best practice" options in the field of alignment (e.g., Blank, 2002; Blank, Porter, &; Porter, 2002; Roach, Niebling, & Kurz, 2008; Smithson, 2001; Webb, 1997a; Webb, Herman, & Webb, 2006). The following bullets describe WAT in greater detail:

- Alignment among the three elements of the educational environment—standards, instruction, assessments—represents a necessary condition for optimal student learning and the validity of test score interpretations. The WAT is a research-validated alignment methodology recommended by the Council of Chief School State Officers (CCSSO), and represents an in-depth process of measuring the alignment between tests and standards.

- The Webb model was conceptualized by Norman L. Webb at the University of Wisconsin, and has been successfully used in over a dozen states to assess the alignment between standards and assessments in language arts, mathematics, science, and social studies (CCSSO, 2006). Webb's model is primarily concerned with the alignment of standards, frameworks, and assessments. His method features expert review panels that provide qualitative judgments as well as quantified coding and analysis of standards and assessment.

- The model begins by training teams of four to six reviewers (e.g., teachers, content specialists) on judging the depth-of-knowledge required to answer test items and meet content objectives. The model's four depth-of-knowledge levels indicate increasingly demanding and complex cognitive tasks: level 1, the recall level, requires the recollection of facts, definitions, terms, or simple procedures; level 2, the skill/concept level, requires students to go beyond one-step problem solving; level 3, the strategic thinking level, asks students to explain their thought processes, make conjectures, and utilize evidence; and level 4, the extended thinking level, demands complex problem solving and the drawing of connections within and across subject domains.

- After the training process, review panels begin to evaluate individual test items judging its depth-of-knowledge level. Reviewers then identify all content standard objectives that correspond to a particular test item. Consensus decisions typically resolve divergent opinions amongst reviewers; however, differences in opinion may also be due to a lack of clarity within the test item or content objective, which can provide the impetus for subsequent revision of the item or objective. Finally, reviewer ratings are used to create descriptive statistics and tabular reports on four criteria of alignment for each item/objective: (a) *categorical concurrence*, (b) *range-of-knowledge correspondence*, (c) *balance of representation*, (d) and *depth-of-knowledge consistency*. The first three criteria allow the model to reflect the breadth dimension of alignment (i.e., the extent of matching content coverage), while the depth-of-knowledge consistency offers information on its depth (i.e., the extent of matching academic difficulty).

Discovery Education Assessment
Research

Below is a selection of Discovery reporting categories from four states—Florida, Tennessee, Kentucky, Illinois—that were carefully designed to **match the content and reporting categories** of the respective state tests. Discovery Education Assessment continually updates its state-specific assessments to reflect the most current version of a state's standards.

Florida Reading Reporting Categories (FCAT)	
Words & Phrases in Context	Reference & Research
Main Idea, Plot, & Purpose	Writing Skills
Comparisons & Cause/Effect	Language

Florida Mathematics Reporting Categories (FCAT)	
Number Sense, Concepts, and Operations	Algebraic Thinking
Measurement	Data Analysis and Probability
Geometry and Spatial Sense	

Florida Science Reporting Categories (FCAT)	
Physical and Chemical	Life and Environmental
Earth and Space	Scientific Thinking

Tennessee Reading Reporting Categories (TCAP)	
Content	Vocabulary
Grammar Conventions	Writing/Organization
Meaning	Writing/Writing Process
Techniques and Skills	

Tennessee Mathematics Reporting Categories (TCAP)	
Number Sense/Number Theory	Data Analysis and Probability
Computation	Measurement
Algebraic Thinking	Geometry
Real World Problem Solving	Graphs and Graphing

Discovery Education Assessment
Research

Tennessee Science Reporting Categories (TCAP)

- Structure and Function of Organisms
- Ecology
- Interactions Between Living Things and Their Environment
- Diversity and Adaptation Among Living Things
- Heredity and Reproduction
- Earth's Features and Resources
- Biological Change
- Energy
- Cell Structure and Function
- Structure and Properties of Matter
- Life Cycles and Biological Changes
- Space, Weather, and Climate
- Motion and Forces, Forms of Energy
- Forces and Motion
- Interactions of Matter
- Matter
- Earth and Its Place in the Universe
- Food Production and Energy for Life
- Atmospheric Cycles
- Earth Features

Kentucky Reading Reporting Categories (KCCT)

- Forming a Foundation for Reading
- Developing an Initial Understanding
- Interpreting Text
- Demonstrating a Critical Stance

Kentucky Mathematics Reporting Categories (KCCT)

- Number/Computation
- Geometry/Measurement
- Probability/Statistics
- Algebraic Thinking

Kentucky Science Reporting Categories (KCCT)

- Physical Science
- Earth & Space Science
- Life Science
- Practical Living

Kentucky Social Studies Reporting Categories (KCCT)

- Government and Civics
- Culture and Society
- Economics
- Geography
- History

Discovery Education Assessment
Research

Illinois Reading Reporting Categories (ISAT)	
Vocabulary Development/ Reading Strategies	Writing Organization/ Purpose
Reading Comprehension	Acquire, Assess, and Communicate Information
Literary Elements/ Literary Works	Reading Strategies
Grammar, Usage and Structure	Variety of Literary Works

Illinois Mathematics Reporting Categories (ISAT)	
Number	Geometry
Measurement	Data Analysis and Probability
Algebra	

Illinois Science Reporting Categories (ISAT)	
Scientific Inquiry/ Tech Design	Earth & Resources/ Universe
Living Things/ Environment	Practices/ Interaction
Matter & Energy/ Force & Motion	

CRITERION VALIDITY

3. **Do Discovery's *Predictive Benchmark* assessments *match* state standardized tests?**

Criterion validity evidence demonstrates that test scores predict scores on an important criterion variable, such as a state's standardized test. Scientifically-based research presents evidence that there is a significant correlation between Discovery Education Assessment *Predictive Benchmark* assessments and a state test, at the overall test score level and also at a specific skill level. Significant correlations show that high scores on these predictive assessments predict high scores on a state's test.

Florida
The Gilchrest County school system participated in a criterion validity study during the 2006/2007 school year. Approximately 1500 students in grades 3 to 10 took Discovery's *Predictive Benchmark* tests. Individual student scores from the 2007 FCAT administration were obtained from the school system. Table 7 shows the correlation for Reading between Discovery and FCAT. Table 8 shows similar results for Mathematics. The median correlation for the Reading assessments was .73 and the median correlation for the Mathematics assessments was .77. All correlations were significant at $p < .01$. Thus, there is substantial evidence that total scores on Discovery's *Predictive Benchmark* assessments predict scale scores on the FCAT for both Reading and Mathematics.

Table 9 shows correlations at the objective level for Reading, and Table 10 shows similar correlations at the objective level for Mathematics. Median correlations were mostly in the .50 range (and all are significant at $p < .01$). Since the number of questions that comprise objectives are much smaller compared to total test score, there is an expectation that these correlations would be somewhat lower than those for total test score but still significant. Thus, there is evidence that objective scores on Discovery's *Predictive Benchmark* assessments predict objective scale scores on the FCAT for both Reading and Mathematics.

Table 7: Correlation of Discovery Reading Growth Score and FCAT Reading Scale Score.

Test B Discovery and FCAT 2007 Spring Reading		
	N	**Correlation**
Grade 3	176	0.74
Grade 4	175	0.73
Grade 5	216	0.66
Grade 6	199	0.72
Grade 7	192	0.72
Grade 8	188	0.73
Grade 9	195	0.74
Grade 10	164	0.74
Median		**0.73**

**All correlations are significant at p < .01*

Table 8: Correlation of Mathematics Growth Score and FCAT Mathematics Scale Score.

Test B Discovery and FCAT 2007 Spring Mathematics		
	N	Correlation
Grade 3	177	0.75
Grade 4	174	0.74
Grade 5	216	0.80
Grade 6	194	0.74
Grade 7	188	0.81
Grade 8	182	0.78
Grade 9	190	0.83
Grade 10	156	0.71
Median		**0.77**

**All correlations are significant at p < .01*

Table 9: Correlation of Reading Reporting Categories and FCAT Reading Objectives.

Test B Discovery and FCAT 2007 Spring Reading				
	Words	Main Idea	Comparison	Reference
Grade 3	0.49	0.61	0.6	0.22
Grade 4	0.36	0.65	0.57	0.26
Grade 5	0.31	0.56	0.44	0.18
Grade 6	0.49	0.56	0.58	0.50
Grade 7	0.38	0.61	0.47	0.33
Grade 8	0.27	0.58	0.47	0.26
Grade 9	0.43	0.63	0.64	0.39
Grade 10	0.44	0.62	0.53	0.46
Median	**0.41**	**0.61**	**0.55**	**0.30**

**All correlations are significant at p < .01*

Table 10: Correlation of Mathematics Reporting Categories and FCAT Mathematics Objectives.

Test B Discovery and FCAT 2007 Spring Mathematics				
	Number	Measurement	Geometry	Algebra
Grade 3	0.66	0.46	0.42	0.52
Grade 4	0.61	0.56	0.24	0.36
Grade 5	0.57	0.52	0.41	0.59
Grade 6	0.55	0.57	0.57	0.57
Grade 7	0.56	0.55	0.43	0.69
Grade 8	0.7	0.66	0.59	0.59

Grade 9	0.59	0.59	0.57	0.67
Grade 10	0.57	0.39	0.56	0.57
Median	**0.58**	**0.56**	**0.50**	**0.58**

**All correlations are significant at p < .01*

Illinois

The Harlem County school system participated in a criterion validity study during the 2006/2007 school year. Approximately 3500 students in grades 3 to 11 took Discovery's *Predictive Benchmark* tests. Individual student scores from the 2007 ISAT and PSAE administration were obtained from the school system. Table 11 shows the correlation for Reading between Discovery Education Assessment and ISAT/PSAE. Table 12 shows similar results for Mathematics. The median correlation for the Reading assessments was .75 and the median correlation for the Mathematics assessments was .80. All correlations were significant at p < .01. Thus, there is substantial evidence that total scores on Discovery's *Predictive Benchmark* assessments predict scale scores on the ISAT/PSAE for both Reading and Mathematics.

Table 13 shows correlations at the objective level for Reading, and Table 14 shows similar correlations at the objective level for Mathematics. Median correlations are mostly in the .40 to .60 range (and all are significant at p < .01). Since the number of questions that comprise objectives are much smaller compared to total test score, there is an expectation that these correlations would be somewhat lower than those for total test score but still significant. Thus, there is evidence that objective scores on Discovery's *Predictive Benchmark* assessments predict objective scale scores on the ISAT for both Reading and Mathematics.

Table 11: Correlation of Discovery Education Assessment and ISAT/PSAE Reading Score.

Discovery and ISAT/PSAE 2007 Spring Reading		
	N	**Correlation***
Grade 3	476	0.55
Grade 4	475	0.79
Grade 5	495	0.76
Grade 6	525	0.75
Grade 7	532	0.75
Grade 8	537	0.75
Grade 11	410	0.20
Median		**0.75**

**All correlations are significant at p < .01*

Table 12: Correlation of Discovery Education Assessment and ISAT/PSAE Mathematics Score.

Discovery and ISAT/PSAE 2007 Spring Mathematics		
	N	Correlation*
Grade 3	471	0.57
Grade 4	477	0.81
Grade 5	494	0.80
Grade 6	525	0.80
Grade 7	531	0.85
Grade 8	524	0.81
Grade 11	176	0.15
Median		**0.80**

**All correlations are significant at p < .01 except Grade 11 significant at p < .05*

Table 13: Correlation of Reading Reporting Categories and ISAT Reading Objectives.

Discovery and ISAT 2007 Spring Reading				
	Vocabulary	Reading Strategies	Reading Comprehension	Literary Elements
Grade 3	0.33	0.47	0.53	0.40
Grade 4	0.52	0.43	0.38	0.52
Grade 5	0.24	0.37	0.38	0.21
Grade 6	0.27	0.30	0.40	0.56
Grade 7	0.31	0.33	0.45	0.36
Grade 8	0.39	0.23	0.45	0.48
Median	0.32	0.35	0.43	0.44

**All correlations are significant at p < .01*

Table 14: Correlation of Mathematics Reporting Categories and ISAT Mathematics Objectives.

Discovery and ISAT 2007 Spring Mathematics				
	Number	Measurement	Algebra	Geometry
Grade 3	0.58	0.62	0.54	0.42
Grade 4	0.64	0.64	0.56	0.48
Grade 5	0.57	0.58	0.63	0.55
Grade 6	0.58	0.56	0.50	0.52
Grade 7	0.68	0.53	0.67	0.59
Grade 8	0.61	0.53	0.67	0.49
Median	0.60	0.57	0.60	0.51

**All correlations are significant at p < .01*

PROFICIENCY PREDICTIVE VALIDITY

4. Can Discovery's *Predictive Benchmark* assessments *predict* state proficiency levels?
Proficiency predictive validity supports the claim that a test can predict a state's proficiency levels. High accuracy levels show that a high degree of confidence can be placed in our test predictions of student proficiency. Two measures of predictive validity are calculated. If only summary data for a school or district are available, the *Proficiency Prediction Score* is tabulated. When individual student level data are available, then an additional index, the *Proficiency Success Rate*, is also calculated. Both measures are explained in the following sections with examples drawn from actual data from Illinois schools.

Proficiency Prediction Score
The Proficiency Prediction Score is used to determine the accuracy of predicted proficiency status. Under the NCLB legislation, it is important that states and school districts help students progress from a "Not Proficient" status to one of "Proficient". The Proficiency Prediction Score is based on the percentage of correct proficiency classifications (Not Proficient/Proficient). If a state uses two or more classifications for "Proficient" (such as "Proficient" and "Advanced"), the percentage of students in these two or more categories would be added together. Also, if a state uses two or more categories for "Not Proficient" (such as "Below Basic" and "Basic"), the percentage of students in these two or more categories would be added together. To see how to use this score, let's assume a school district had the following data based on its annual state test and a Discovery Education Assessment Spring benchmark assessment. Let's use data from a Grade 4 Mathematics Test as an example:

Predicted Percent Proficient or higher = 70%
Actual Percent Proficient or higher on the State Test = 80%

The error rate for these predictions is as follows:

Error Rate = /Actual Percent Proficient minus Predicted Percent Proficient/
Error Rate = /80% - 70%/ = 10%

In this example, Discovery Education Assessment under predicted the percent of students proficient by 10%. The absolute value (shown by the symbols / /) of the error rate is used to account for cases where Discovery Education Assessment over predicts the percent of students proficient and the calculation is negative (e.g., Actual - Predicted = 70% - 80% = -10%; absolute value is 10%).

The Proficiency Prediction Score is calculated as follows:

Proficiency Prediction Score = 100% minus Error Rate

In this example, the score is as follows:

Proficiency Prediction Score = 100% - 10% = 90%.

A higher Proficiency Prediction Score indicates a larger number or percentage of correct proficiency predictions. In this example, Discovery Education Assessment had a score of 90%, which indicates 9 correct classifications for every 1 misclassification. Discovery Education Assessment uses information from these scores to improve its benchmark assessments every year.

Florida
The Putnam County School system participated in a proficiency prediction study during the 2006/2007 school year. Comparisons of Discovery Education Assessment proficiency predictions on the Spring 2007 tests with actual FCAT 2007 results were made for grades 3 to 10 in Reading and Mathematics. Approximately 6800 students participated in this study.

The Proficiency Prediction Scores for all grades in Reading and Mathematics are presented in Table 15. The median Proficiency Prediction Score for Reading was 86.49%, and the median Proficiency Prediction Score for Mathematics was 94.5%.

Table 15: Putnam County Proficiency Prediction Scores for Reading and Mathematics.

	Reading		Mathematics	
	N	Proficiency Prediction Score	N	Proficiency Prediction Score
Grade 3	857	95%	858	86.6%
Grade 4	852	88.73%	852	98%
Grade 5	837	90.47%	837	90.42%
Grade 6	828	72.35%	878	80.81%
Grade 7	799	84%	805	99.6%
Grade 8	850	96%	837	91.62%
Grade 9	789	82.5%1	715	99.42%
Grade 10	1025	84.25%	930	97.38%
Median		**86.49%**		**94.5%**

Proficiency Success Rate
When individual student data are available, an additional measure, the *Proficiency Success Rate*, can also be calculated. After taking Discovery's *Predictive Benchmark* assessment, a student receives a prediction of his or her proficiency status: *Proficient* (Level 3, 4, or 5) or *Not Proficient* (Level 1 or 2). The percentage of students predicted as proficient by Discovery Education Assessment that actually scored proficient on the FCAT is called the Proficiency Success Rate. For instance, a Proficiency Success Rate of 90% indicates that ninety percent of the students that Discovery Education predicted as proficient actually achieved this status on the FCAT.

The Gilchrist County School District also participated in a Proficiency Success Rate study during the 2006/2007 school year. Individual student proficiency scores were obtained for Reading an

Mathematics in grades 3 to 10 and compared with proficiency predictions on Discovery's *Predictive Benchmark* assessments. Table 16 and Table 17 present the Proficiency Success Rates for Reading and Mathematics. The median Proficiency Success Rate for Reading was 82.37%, and the median Proficiency Success Rate for Mathematics was 89.29%.

Table 16: Results of the Proficiency Success Rate Study in Gilchrist County for Reading.

Proficiency Success Rate in Gilchrist County 2006-2007 Reading		
	N	Proficiency Success Rate
Grade 3	176	91.73%
Grade 4	175	85.62%
Grade 5	216	87.89%
Grade 6	199	78.7%
Grade 7	192	82.32%
Grade 8	188	82.41%
Grade 9	195	72.79%
Grade 10	164	53.7%
Median		**82.37%**

Table 17: Results of the Proficiency Success Rate Study in Gilchrist County for Mathematics.

Proficiency Success Rate in Gilchrist County 2006-2007 Mathematics		
	N	Proficiency Success Rate
Grade 3	177	97.84%
Grade 4	174	87.23%
Grade 5	216	81.33%
Grade 6	194	76.43%
Grade 7	188	88.89%
Grade 8	182	90.78%
Grade 9	190	90.38%
Grade 10	156	89.68%
Median		**89.29%**

Tennessee
Due to our representation throughout the state of Tennessee, direct comparisons of Spring 2007 (Test B) and actual 2007 TCAP proficiency percentages were made for grades 3 to 8 in Reading and Mathematics.

The Proficiency Prediction Scores were calculated via the aforementioned formulas using the combined percentages of “Proficient” and “Advanced”. The results for all grades in Reading and Mathematics are presented in Table 18. The median Proficiency Prediction Score for Reading was 96%, and the median Proficiency Prediction Score for Mathematics was 92%.

Table 18: Proficiency Prediction Scores for Reading and Mathematics.

	Reading Proficient & Advanced Combined	Mathematics Proficient & Advanced Combined
	Proficiency Prediction Score	Proficiency Prediction Score
Grade 3	100%	99%
Grade 4	99%	98%
Grade 5	89%	86%
Grade 6	93%	93%
Grade 7	98%	90%
Grade 8	86%	91%
Median	**96%**	**92%**

Kentucky

Due to our representation throughout the state of Kentucky, direct comparisons between the Discovery Test B (Spring 2007) and actual 2007 KCCT proficiency percentages were made for Grades 3 through 8 in Reading and Mathematics.

The Proficiency Prediction Scores were calculated via the aforementioned formulas. The results for all grades in Reading and Mathematics are presented in Table 19. The median Proficiency Prediction Score for Reading was 97%, and the median Proficiency Prediction Score for Mathematics was 94%.

Table 19: Proficiency Prediction Scores for Reading and Mathematics.

	Reading	Mathematics
	Proficiency Prediction Score	Proficiency Prediction Score
Grade 3	92%	99%
Grade 4	98%	92%
Grade 5	95%	100%
Grade 6	96%	89%
Grade 7	99%	87%
Grade 8	100%	96%
Median	**97%**	**94%**

Illinois

The Harlem County School system participated in a proficiency prediction study during the 2006/2007 school year. Comparisons of Discovery Education Assessment proficiency predictions on the Spring 2007 tests with actual ISAT and PSAE 2007 results were made for grades 3 to 11 in Reading and Mathematics. Approximately 3500 students participated in this study.

The Proficiency Prediction Scores for all grades in Reading and Mathematics are presented in Table 20. The median Proficiency Prediction Score for Reading was 97%, and the median Proficiency Prediction Score for Mathematics was 96%.

Table 20: Harlem County Proficiency Prediction Scores for Reading and Mathematics.

	Reading		Mathematics	
	N	Proficiency Prediction Score	N	Proficiency Prediction Score
Grade 3	475	97%	470	96%
Grade 4	475	97%	477	98%
Grade 5	495	98%	494	100%
Grade 6	525	98%	525	96%
Grade 7	532	93%	531	89%
Grade 8	537	93%	524	88%
Grade 11	410	92%	176	98%
Median		**97%**		**96%**

Proficiency Success Rate

When individual student data are available, an additional measure, the *Proficiency Success Rate*, can also be calculated. After taking Discovery's *Predictive Benchmark* assessment, a student receives a prediction of his or her proficiency status: *Proficient* (Meets or Exceeds) or *Not Proficient* (Below or Warning). The percentage of students predicted as proficient by Discovery Education Assessment that actually scored proficient on the ISAT or PSAE is called the Proficiency Success Rate. For instance, a Proficiency Success Rate of 90% indicates that ninety percent of the students that Discovery Education predicted as proficient actually achieved this result on the ISAT or PSAE.

The Harlem County School District also participated in proficiency success rate study during the 2006-2007 school year. Individual student proficiency scores were obtained for Reading and Mathematics in grades 3 to 11 and compared with proficiency predictions on Discovery Education Predictive Assessments. Table 21 and Table 22 present the Proficiency Success Rates for Reading and Mathematics. The median Proficiency Success Rate for Reading was 91%, and the median Proficiency Success Rate for Mathematics was 94%.

Table 21: Results of the Proficiency Success Rate Study in Harlem County for Reading.

Proficiency Success Rate in Harlem County 2006-2007 Reading		
	N	Proficiency Success Rate
Grade 3	475	91%
Grade 4	475	87%
Grade 5	495	88%
Grade 6	525	91%
Grade 7	532	92%
Grade 8	537	93%
Grade 11	410	71%
Median		**91%**

Table 22: Results of the Proficiency Success Rate Study in Harlem County for Mathematics.

Proficiency Success Rate in Harlem County 2006-2007 Mathematics		
	N	Proficiency Success Rate
Grade 3	470	96%
Grade 4	477	94%
Grade 5	494	92%
Grade 6	525	94%
Grade 7	531	97%
Grade 8	524	98%
Grade 11	176	80%
Median		**94%**

New York

Comparisons of Discovery Education Assessment proficiency predictions between the 0607 Test A and Test B results and actual 2007 NY State test results were made for our two largest middle school customers' grades 6-8 in English Language Arts and Mathematics.

The Proficiency Prediction Scores for Test A grades 6-8 in Mathematics at Albion Middle are presented in Table 23. The median Proficiency Prediction Score for Test A at Albion Middle was 91%. Table 24 provides the Proficiency Prediction Scores and proficiency averages for Test B grades 6-8 in Mathematics at Albion Middle. The median Proficiency Prediction Score for Test B at Albion Middle was 88%. Table 25 provides the Proficiency Prediction Scores and proficiency averages for Test B grades

6-8 in English Language Arts (ELA) at William H. Golding. The median Proficiency Prediction Score for Test B at William H. Golding was 82%.

Table 23: Albion Middle Test A Proficiency Prediction Scores for Mathematics.

	Mathematics Discovery Test A	
	N	**Proficiency Prediction Score**
Grade 6	170	96%
Grade 7	187	89%
Grade 8	207	91%
Median		**91%**

Table 24: Albion Middle Test B Proficiency Prediction Scores for Mathematics.

	Mathematics Discovery Test B	
	N	**Proficiency Prediction Score**
Grade 6	165	87%
Grade 7	185	96%
Grade 8	185	88%
Median		**88%**

Table 25: William H. Golding Middle Test B Proficiency Prediction Scores for English Language Arts.

	English Language Arts Discovery Test B	
	N	**Proficiency Prediction Score**
Grade 6	52	65%
Grade 7	128	96%
Grade 8	161	82%
Median		**82%**

Alabama
Due to our representation throughout the state of Alabama, direct comparisons between Discovery Test B (Spring 2007) and actual 2007 ARMT proficiency percentages were made for grades 3 to 8 in Reading and Mathematics.

The Proficiency Prediction Scores were calculated via the aforementioned formulas using the combined percentages of "Level III" (Meets Academic Content Standards) and "Level IV" (Exceeds Academic Content Standards). The results for grade 3 through 8 in Reading and Mathematics are presented in Table 26. The median Proficiency Prediction Score for Reading was 95%, and the median Proficiency Prediction Score for Mathematics was 96%.

Table 26: Proficiency Prediction Scores for Reading and Mathematics.

	Reading Level III & Level IV		Mathematics Level III & Level IV	
	N	Proficiency Prediction Score	N	Proficiency Prediction Score
Grade 3	13,561	95%	12,413	96%
Grade 4	13,033	94%	11,402	91%
Grade 5	11,827	89%	10,758	90%
Grade 6	10,563	95%	9,370	98%
Grade 7	10,235	97%	7,600	96%
Grade 8	9,287	97%	8,542	97%
Median		**95%**		**96%**

CONSEQUENTIAL VALIDITY

5. Can the use of Discovery's *Predictive Benchmark* assessments improve student learning?

Consequential validity outlines how the use of benchmark assessments facilitates important consequences, such as the improvement of student learning and student performance on state standardized tests.

Florida
The Gilchrist County School System also participated in a consequential validity study. This system used Discovery's *Predictive Benchmark* assessments during the 2006/2007 school year. The percent of students that were classified as Proficient (Levels 3, 4, or 5) on the 2007 FCAT was tabulated and compared with the percent of students that were classified as Proficient on the 2006 FCAT. The results for grades 3 to 10, Reading and Mathematics, for the two years, 2006 and 2007 are presented in the following tables. The results are presented separately for six schools in Gilchrist County, the Bell schools Table 27 and 28) and Trenton schools (Table 29 and 30). The "Difference" between 2007 and 2006 was also tabulated; a positive score indicates an increase in the percent of students proficient from 2006 to 2007. As a reference point, the improvement (or decline) in the percent of students proficient in the state of Florida was compared to this Difference score.

Table 27: Results of Consequential Validity Study for Bell Schools in Reading.

Bell Elementary, Middle, and High School in Gilchrist County 2006-2007 Reading					
Grade	**2006**	**2007**	**Difference**	**Bell FL State**	
3	78%	82%	4%	10%	
4	76%	69%	-7%	-9%	
5	78%	81%	3%	-2%	
6	72%	72%	0%	2%	
7	71%	66%	-5%	-7%	
8	50%	58%	8%	5%	
9	41%	50%	9%	8%	
10	39%	34%	-5%	-7%	

Take a look at grade 3 Reading for Bell Elementary. The percent of students proficient in 2006 was 78, and the percent proficient in 2007 was 82, a difference or improvement of 4%. The state of Florida actually had a decline of 6% for these years in grade 3 Reading. So the "Bell FL State" calculation is actually 10%; the Bell grade 3 Reading classes improved 10% in the percent of students proficient compared to the state of Florida.

Table 28: Results of Consequential Validity Study for Bell Schools in Mathematics.

Bell Elementary, Middle, and High School in Gilchrist County 2006-2007 Mathematics					
Grade	2006	2007	Difference	Bell	FL State
3	83%	91%	8%	6%	
4	70%	69%	-1%	-3%	
5	66%	68%	-2%	-4%	
6	59%	65%	6%	9%	
7	52%	61%	9%	5%	
8	70%	77%	8%	5%	
9	64%	71%	7%	6%	
10	76%	73%	-3%	-3%	

Table 29: Results of Consequential Validity Study for Trenton Schools in Reading.

Trenton Elementary, Middle, and High School in Gilchrist County 2006-2007 Reading					
Grade	2006	2007	Difference	Trenton	FL State
3	88%	72%	-16%	-10%	
4	69%	82%	13%	11%	
5		81%			
6	72%	67%	-5%	-3%	
7	63%	72%	9%	7%	
8	55%	56%	1%	-2%	
9	51%	53%	2%	1%	
10	34%	52%	18%	16%	

Table 30: Results of Consequential Validity Study for Trenton Schools in Mathematics.

Trenton Elementary, Middle, and High School in Gilchrist County 2006-2007 Mathematics					
Grade	2006	2007	Difference	Trenton	FL State
3	82%	83%	1%	-1%	
4	84%	84%	0%	-2%	
5		66%			

6	61%	56%	-5%	-2%
7	58%	75%	17%	13%
8	75%	72%	-3%	-6%
9	74%	79%	5%	4%
10	81%	80%	-1%	-1%

Many factors contribute to the improvement of the percent of students proficient from year to year. Discovery's *Predictive Benchmark* assessments are usually just one factor in school-wide and district-wide improvement plans. Thus, these results should be considered in the light of these many factors.

Tennessee

The Grainger County school system participated in a consequential validity study. This system used Discovery's *Predictive Benchmark* assessments during the 2006/2007 school year. The percent of students that were classified as "Proficient" and "Advanced" on the 2007 TCAP was tabulated and compared with the percent of students that were classified as "Proficient" and "Advanced" on the 2006 TCAP. The results for grades 3 to 8, Reading and Mathematics, for the two years—2006 and 2007—are presented in Table 31 and 32.The "Difference" between 2007 and 2006 was also tabulated; a positive score indicates an increase in the percent of students proficient from 2006 to 2007. As a reference point, the improvement (or decline) in the percent of students classified as "Proficient" and "Advanced" in the state of Tennessee was compared to this Difference score.

The percentages are to be understood as follows. Take a look below at grade 3 Mathematics. The percent of students proficient in 2006 was 87, and the percent proficient in 2007 was 93, a difference or improvement of 5% (using exact not rounded percentages). However, grade 3 Mathematics in the state of Tennessee improved by only 1% during the same time. Therefore, the "Grainger TN State" calculation is actually 4%. That is, the Grainger County grade 3 Reading classes improved 4% in the percent of students proficient compared to the state of Tennessee.

Table 31: Results of Consequential Validity Study for Grainger County in Mathematics.

Grainger County, TN Mathematics				
Grade	**2006**	**2007**	**Difference***	**Grainger TN State***
3	87%	93%	5%	4%
4	90%	96%	6%	4%
5	94%	97%	2%	2%
6	87%	92%	4%	4%
7	92%	91%	-1%	-2%
8	89%	92%	3%	0%

**Calculated based on exact not rounded percentages listed under 2006 and 2007.*

Table 32: Results of Consequential Validity Study for Grainger County in Reading.

Grainger County, TN Reading					
Grade	2006	2007	Difference*	Grainger State*	TN
3	87%	93%	5%	2%	
4	92%	88%	-4%	-4%	
5	92%	97%	5%	2%	
6	91%	94%	3%	-2%	
7	90%	93%	2%	0%	
8	92%	94%	2%	0%	

**Calculated based on exact not rounded percentages listed under 2006 and 2007.*

Many factors contribute to the improvement of the percent of students proficient from year to year. Discovery's *Predictive Benchmark* assessments are usually just one factor in school and district-wide improvement plans. Thus, these results should be considered in the light of these many factors.

Illinois

The Harlem County School System participated in a consequential validity study. This system used Discovery's *Predictive Benchmark* assessments during the 2006/2007 school year. The percent of students that were classified as Proficient (Meets or Exceeds Standards) on the 2007 ISAT was tabulated and compared with the percent of students that were classified as Proficient on the 2006 ISAT. The results for grades 3 to 8, Reading and Mathematics, for the two years, 2006 and 2007 are presented in the following tables. The results are presented separately for three schools in Harlem County: Harlem Middle School (Table 33 and 34) and Olson (Table 35 and 36) and Ralston Elementary (Table 37 and 38). The "Difference" between 2007 and 2006 is also tabulated; a positive score indicates an increase in the percent of students proficient from 2006 to 2007. As a reference point, the improvement (or decline) in the percent of students proficient in the state of Illinois was compared to this Difference score.

For Harlem Middle School, there was significant improvement in grades 7 and 8 Reading and grade 4 Mathematics. For Olson Elementary, there were significant improvements in grades 4 through 6 Reading and grade 6 Mathematics. For Ralston Elementary, there were improvements in grade 3 Reading and grades 3, 4, and 6 Mathematics.

Table 33: Results of Consequential Validity Study for Harlem Middle in Reading.

Harlem Middle in Harlem County 2006-2007 Reading					
Grade	2006	2007	Difference	Harlem	IL State
7	72%	76%	4%	3%	
8	73%	78%	5%	2%	

Table 34: Results of Consequential Validity Study for Harlem Middle in Mathematics.

Harlem Middle in Harlem County 2006-2007 Mathematics					
Grade	**2006**	**2007**	**Difference**	**Harlem**	**IL State**
3	86%	86%	0%	-1%	
4	73%	78%	5%	3%	

Table 35: Results of Consequential Validity Study for Olson Elementary in Reading.

Olson Elementary in Harlem County 2006-2007 Reading					
Grade	**2006**	**2007**	**Difference**	**Olson**	**IL State**
3	84%	83%	-1%	-3%	
4	70%	80%	10%	9%	
5	68%	78%	10%	9%	
6	75%	85%	10%	9%	

Table 36: Results of Consequential Validity Study for Olson Elementary in Mathematics.

Olson Elementary in Harlem County 2006-2007 Mathematics					
Grade	**2006**	**2007**	**Difference**	**Olson**	**IL State**
3	95%	91%	-4%	-5%	
4	91%	91%	0%	-2%	
5	89%	87%	-2%	-6%	
6	82%	89%	7%	5%	

Table 37: Results of Consequential Validity Study for Ralston Elementary in Reading.

Ralston Elementary in Harlem County 2006-2007 Reading					
Grade	**2006**	**2007**	**Difference**	**Ralston**	**IL State**
3	78%	89%	11%	9%	
4	79%	77%	-2%	-3%	
5	85%	78%	-7%	-8%	
6	86%	87%	1%	0%	

Table 38: Results of Consequential Validity Study for Ralston Elementary in Mathematics.

Ralston Elementary in Harlem County 2006-2007 Mathematics				
Grade	**2006**	**2007**	**Difference**	**Ralston IL State**
3	87%	94%	7%	6%
4	86%	93%	7%	5%
5	88%	88%	0%	-4%
6	88%	95%	7%	5%

Many factors contribute to the improvement of the percent of students proficient from year to year. Discovery's *Predictive Benchmark* assessments are usually just one factor in school-wide and district-wide improvement plans. Thus, these results should be considered in the light of these many factors.

Alabama

The Birmingham City Schools participated in a consequential validity study. This system used Discovery's *Predictive Benchmark* assessments during the 2006/2007 school year. The percent of students that were classified as "Level III" and "Level IV" on the 2006 ARMT was tabulated and compared with the percent of students that were classified as "Level III" and "Level IV" on the 2007 ARMT. The results for grades 3 to 8, Reading and Mathematics, for the two years—2006 and 2007—are presented in Table 39 and 40. The "Difference" between 2006 and 2007 was also tabulated; a positive score indicates an increase in the percent of students proficient from 2006 to 2007. As a reference point, the improvement (or decline) in the percent of students classified as "Level III" and "Level IV" in the state of Alabama was compared to this Difference score.

The percentages are to be understood as follows. Take a look below at grade 6 Reading. The percent of students proficient in 2006 was 68, and the percent proficient in 2007 was 73, a difference or improvement of 5%. However, grade 6 Reading in the state of Alabama improved by only 2% during the same time. Therefore, the "Birmingham AL State" calculation is actually 3%. That is, the Birmingham City grade 3 Reading classes improved 3% in the percent of students proficient compared to the state of Alabama.

Table 39: Results of Consequential Validity Study for Birmingham City Schools in Reading.

Birmingham City, AL Reading				
Grade	**2006**	**2007**	**Difference***	**Birmingham AL State***
3	74%	74%	0%	-1%
4	72%	73%	1%	1%
5	70%	74%	4%	0%
6	68%	73%	5%	3%
7	64%	66%	3%	0%
8	60%	62%	2%	1%

**Calculated based on exact not rounded percentages listed under 2006 and 2007.*

Table 40: Results of Consequential Validity Study for Birmingham City Schools in Mathematics.

Birmingham City, AL Mathematics				
Grade	**2006**	**2007**	**Difference***	**Birmingham AL State***
3	68%	68%	0%	0%
4	69%	67%	-2%	-2%
5	69%	69%	0%	-1%
6	58%	60%	2%	4%
7	42%	48%	6%	5%
8	57%	58%	2%	3%

**Calculated based on exact not rounded percentages listed under 2006 and 2007.*

Many factors contribute to the improvement of the percent of students proficient from year to year. Discovery's *Predictive Benchmark* assessments are usually just one factor in school and district-wide improvement plans. Thus, these results should be considered in the light of these many factors.

GROWTH MODELS

6. Can Discovery's *Predictive Benchmark* assessments be used to measure growth over time?

Growth models depend on a highly rigorous and valid vertical scale to measure student performance over time. Discovery Education Assessment vertical scales are constructed using Rasch measurement models with state-of-the-art psychometric techniques.

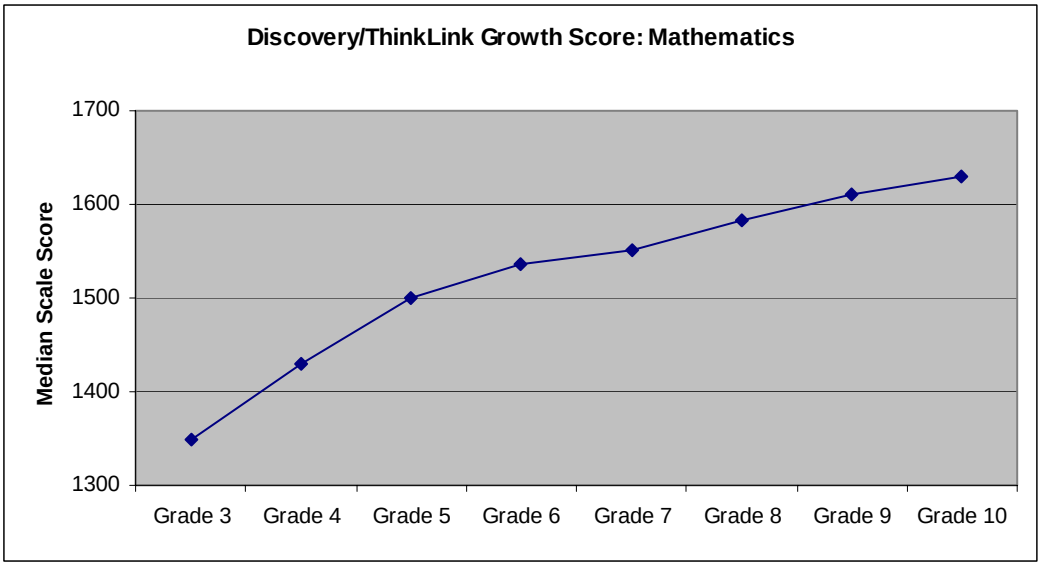
The accurate measurement of student achievement over time is becoming increasingly important to parents, teachers, and school administrators. **Student "growth" within a grade and across grades** has also been sanctioned by the U. S. Department of Education as a reliable way to measure student proficiency in Reading and Mathematics and to **satisfy the requirements of Adequate Yearly Progress (AYP)** under the No Child Left Behind Act. Accurate measurement and recording of individual student achievement can also help with **issues of student mobility**: as students move within a district or state, records of individual student achievement can help new schools administer to the needs of this mobile population.

The assessment of student achievement over time is even more important with the use of benchmarks tests. Discovery's *Predictive Benchmark* assessments provide a snapshot of student progress toward state standards at up to four points during the school year. These benchmark tests are scientifically linked, so that the reporting of student proficiency levels is both reliable and valid.

How is the growth score created?
Discovery Education Assessment has added a scientifically based vertical scaled growth score to its family of benchmark tests in 2007-08. These growth scores are based on the Rasch measurement model, a state-of-the-art psychometric technique for scaling ability (e.g., Wright & Stone, 1979; Wright & Masters, 1982; Linacre 1999; Smith & Smith, 2004; Wilson, 2005). To accomplish vertical scaling, common items are embedded across assessments to enable the psychometric linking of tests at different points in time. For example, a grade 3 mathematics benchmark test administered mid-year might contain below grade level and above grade level items. Performance on these off grade level items provides an accurate measurement of how much growth occurs across grades. Furthermore, benchmark tests within a grade are also linked with common items, once again to assess change at different points in time within a grade. Discovery Education Assessment is using established psychometric procedures to build calibrated item banks and linked tests (i.e., Ingebo, 1997; Kolen & Brennan, 2004).

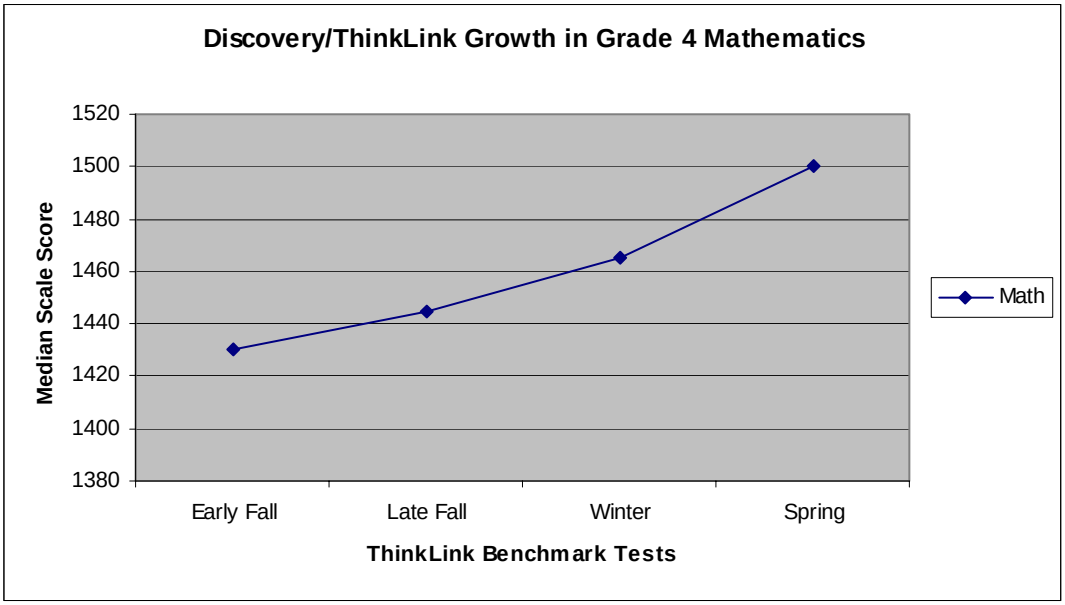
Why use such a rigorous vertical scale?
Isn't student growth similar across grades? Don't students change as much from grade 3 to grade 4 as they do from grade 7 to grade 8? Previous research on the use of vertical scales has demonstrated that **student growth is not linear**; that is, growth in student achievement is different from grade to grade (see Young 2006). For instance, Figure 1 on the next page shows preliminary Discovery Education Assessment vertically scaled growth results. This graph shows growth from grades 3 to 10 in Mathematics as measured by Discovery's Spring benchmark tests. Typically, students have larger gains in mathematics achievement in elementary grades with growth somewhat slowing in middle and high school, as published by other major testing companies.

Figure 1: Vertically Scaled Growth Results for Discovery Education Assessment Mathematics Tests.



What is unique about the Discovery Education Assessment vertical growth scores?
Student growth can now be accurately measured at four points in time in each grade level. Discovery's *Predictive Benchmark* assessments are administered up to four times yearly: Early Fall, Late Fall, Winter, and Spring. For each time period, we report scale scores and accompanying statistics. Most testing companies only allow the measurement of student growth at two points in time: Fall and Spring. Discovery's *Predictive Benchmark* assessments provide normative information to measure student growth multiple times each year. Figure 2 illustrates this growth for grade 4 Mathematics using our benchmark assessments.

Figure 2: Within-Year Growth Results for Discovery Education Assessment Mathematics Tests.



Florida Growth Scale

The following tables illustrate the Test Difficulty on the Discovery Education Assessment vertical growth scale for Reading and Mathematics tests between two time periods, Fall 2007 and Spring 2008.

Table 41: Vertical Growth Score Comparisons for Fall 2007 and Spring 2008 in Reading.

Florida 0708 Test Difficulty Comparisons Reading									
	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 9	Gr. 10
Fall	1327	1395	1418	1486	1504	1541	1562	1620	1623
Spring	1364	1409	1477	1495	1548	1555	1598	1627	1639

Table 42: Vertical Growth Score Comparisons for Fall 2007 and Spring 2008 in Mathematics.

Florida 0708 Test Difficulty Comparisons Mathematics									
	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 9	Gr. 10
Fall	1267	1330	1408	1466	1510	1566	1581	1609	1622
Spring	1305	1401	1440	1510	1565	1588	1598	1604	1647

Tennessee Growth Scale
The following tables illustrate the Test Difficulty on the Discovery Education Assessment vertical growth scale for the 0708 Reading and Mathematics tests between two time periods, Fall and Winter 2007.

Table 43: Vertical Growth Score Comparisons for Fall 2007 and Winter 2007 in Reading.

Tennessee 0708 Test Difficulty Comparisons Reading						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test P (Fall)	1416	1432	1483	1523	1556	1547
Test A (Winter)	1429	1481	1515	1535	1565	1584

Table 44: Vertical Growth Score Comparisons for Fall 2007 and Winter 2007 in Mathematics.

Tennessee 0708 Test Difficulty Comparisons Mathematics						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test P (Fall)	1348	1387	1441	1516	1575	1593
Test A (Winter)	1366	1441	1496	1557	1568	1598

Kentucky Growth Scale

The following tables illustrate the Test Difficulty on the Discovery Education Assessment vertical growth scale for the 0708 Reading and Mathematics tests across three time periods: Fall 2007 (Test P), Winter 2007 (Test A), and Spring 2008 (Test B).

Table 45: Vertical Growth Score Comparisons for Fall, Winter, and Spring 0708 in Reading.

Kentucky 0708 Test Difficulty Comparisons Reading								
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 9	Gr. 10
Test P (Fall)	1385	1417	1477	1497	1550	1529	1581	1620
Test A (Winter)	1403	1485	1511	1520	1542	1588	1602	1625
Test B (Spring)	1424	1487	1525	1534	1571	1595	1622	1639

Table 46: Vertical Growth Score Comparisons for Fall, Winter, Spring 0708 in Mathematics.

Kentucky 0708 Test Difficulty Comparisons Mathematics								
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 9	Gr. 10
Test P (Fall)	1341	1409	1424	1496	1561	1577	1605	1652
Test A (Winter)	1355	1415	1493	1529	1554	1598	1628	1664
Test B (Spring)	1379	1457	1495	1551	1581	1601	1638	1667

Illinois Growth Scale

The following tables illustrate the test difficulty on the Discovery Education Assessment vertical growth scale for Reading and Mathematics tests between three time periods, Fall 2007 (Test P), Winter 2008 (Test A), and Spring 2008 (Test B).

Table 47: Vertical Growth Score Comparisons for Fall, Winter, and Spring of 0708 in Reading.

Illinois 0708 Test Difficulty Comparisons Reading						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Fall 2007	1395	1421	1460	1508	1538	1535
Winter 2008	1410	1463	1493	1529	1577	1588
Spring 2008	1395	1485	1496	1529	1570	1603

Table 48: Vertical Growth Score Comparisons for Fall, Winter, and Spring of 0708 in Mathematics.

Illinois 0708 Test Difficulty Comparisons Mathematics						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Fall 2007	1343	1379	1447	1519	1546	1567
Winter 2008	1372	1421	1499	1529	1558	1600
Spring 2008	1392	1442	1504	1550	1565	1606

New York Growth Scale

The following tables illustrate the Test Difficulty on the Discovery Education Assessment vertical growth scale for English Language Arts and Mathematics tests for two time periods, Test A 0708 and Test B 0708.

Table 49: Vertical Growth Score Comparisons for 0708 Test A and B in English Language Arts.

New York 0708 Test Difficulty Comparisons English Language Arts						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test A	1408	1442	1495	1511	1533	1596
Test B	1393	1457	1504	1532	1562	1589

Table 50: Vertical Growth Score Comparisons for 0708 Test and B in Mathematics.

New York 0708 Test Difficulty Comparisons Mathematics						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test A	1358	1443	1506	1552	1595	1636
Test B	1377	1456	1508	1566	1582	1621

Tables 51 and 52 illustrate the Student Test Averages on the Discovery Education Assessment vertical growth scale for English Language Arts and Mathematics tests for two time periods, Test A 0708 and Test B 0708.

Table 51: Vertical Growth Score Comparisons for 0708 Test A and B in English Language Arts.

New York 0708 Student Ability Comparisons English Language Arts						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test A	1436	1474	1525	1588	1566	1615
Test B	1452	1499	1534	1571	1609	1622

Table 52: Vertical Growth Score Comparisons for 0708 Test and B in Mathematics.

New York 0708 Student Ability Comparisons Mathematics						
	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Test A	1366	1454	1519	1534	1594	1607
Test B	1428	1485	1529	1605	1630	1662

Alabama Growth Scale

The following tables illustrate the test difficulty on the Discovery Education Assessment vertical growth scale for Reading and Mathematics tests between three time periods, Fall 0708 (Test P), Winter 0708 (Test A), and Spring 0708 (Test B).

Table 53: Vertical Growth Score Comparisons for Fall, Winter, and Spring of 0708 in Reading.

Alabama 0708 Test Difficulty Comparisons Reading							
	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Fall	1317	1409	1405	1480	1508	1543	1565
Winter	1393	1413	1463	1503	1516	1527	1571
Spring	1397	1396	1483	1512	1539	1536	1610

Table 54: Vertical Growth Score Comparisons for Fall, Winter, and Spring of 0708 in Mathematics.

Alabama 0708 Test Difficulty Comparisons Mathematics							
	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Fall	1285	1326	1411	1440	1507	1544	1572
Winter	1289	1388	1434	1507	1518	1587	1588
Spring	1315	1380	1458	1491	1538	1562	1588

NCLB Scientifically-Based Research

7. Are Discovery Education Predictive Assessments based on scientifically-based research advocated by the U. S. Department of Education?

Discovery Education Assessment has also adhered to the criteria for "scientifically-based research" put forth in the *No Child Left Behind Act of 2001*. "What is Predictive Assessment?" has outlined how Discovery Education Predictive Assessments test reliability and validity meets the following criteria for scientifically-based research set forth by NCLB:

(i) *employs systematic, empirical methods that draw on observation and experiment;*
(ii) *involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;*
(iii) *relies on measurements or observational methods that provide reliable and valid data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;*

Discovery Education Assessment also provides evidence of meeting the following scientifically-based research criterion:

(iv) *is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition control.*

Case Study One: Birmingham City Schools, Alabama
Larger schools and school districts typically do not participate in experimental or quasi-experimental studies due to logistical and ethical concerns. However, a unique situation in Birmingham, Alabama afforded Discovery Education Assessment with the opportunity to investigate the efficacy of its benchmark assessments in respect to a quasi-control group. In 2003/2004, approximately one-half of the poverty schools in Birmingham City used Discovery's *Predictive Benchmark* tests whereas the other half of poverty schools did not. Schools were selected and matched for representative characteristics of both student demographics and prior student achievement by the Birmingham Board of Education. Approximately 6500 students participated in each group. The State of Alabama used the Stanford Achievement Test (SAT10) to measure student academic ability. Results for the SAT10 from the 2003 school year were used as the baseline measure. Results from the SAT10 for the 2004 school year were used to measure the difference between the experimental schools and the matched control schools. At the end of the school year, achievement results for both groups were compared revealing a significant improvement on the SAT10 for those schools that used Discovery's *Predictive Benchmark* tests as opposed to those that did not. Discovery Education Assessment subsequently compiled a brief report titled the "Birmingham Case Study". Excerpts from the case study are included below:

This study is based on data from elementary and middle schools in the City of Birmingham, Alabama. In 2002-03, Birmingham Schools did not use the Discovery Education Assessment: *Predictive Benchmark* tests. Starting in 2003-04, 20 elementary (grades 3 to 5) and 9 middle schools (grades 6 to

8) used the Discovery Education Assessment: *Predictive Benchmark* program. An equal number of elementary and middle schools were used for the matched comparison group. All Birmingham schools took the Stanford Achievement Test Tenth Edition (SAT10) at the end of both school years. The SAT10 is administered yearly as part of the state's school accountability program. The state of Alabama uses improvement in SAT10 percentiles to gauge school progress and as part of its NCLB reporting. National percentiles on the SAT10 are reported by subject and grade level. The subjects are Language Arts, Reading, and Mathematics. A single national percentile is reported for all students within a subject and grade. Furthermore, national percentiles are disaggregated by various subgroups within a school. For the comparisons that follow, the national percentiles for students classified as utilizing free and reduced lunch were used. All percentiles have been converted to Normal Curve Equivalents (NCE) to allow for averaging of results.

The Discovery schools comprised the experimental group in this study. The Birmingham schools that did not use Discovery comprise the matched comparison group. Tables 55, 56, and 57 show SAT10 National Percentile changes for Discovery schools vs. Non-Discovery schools in grades 3 to 8 for the subjects of Language Arts, Reading, and Mathematics. The average NCE for 2003 and 2004 is shown along with the change from 2003 to 2004. The final column in the tables shows the difference in NCEs for Discovery schools compared to Non-Discovery schools.

With sample sizes this large, NCEs of 1.0 or greater are significant at $p<.01$ using t-tests. Using this criteria, Discovery schools outperformed Non-Discovery schools in all grades and subjects except for grade 5 Language Arts and grade 6 Reading. Of more practical significance is the change in NCEs for a school's status under the state of Alabama accountability criteria. Gains in NCEs of greater than 1.5 NCEs are considered practically significant and indicate progress at a school level and individual level.

As a result of the improvement that many of the Discovery schools had made in the 2003/2004 school year, the Birmingham City Schools adopted the Discovery Education Assessment: *Predictive Benchmark* program in *all* of the schools the following school year. The Birmingham City Schools also chose to provide professional development in each school to help all teachers become more familiar with the concepts of formative assessment and to better utilize data for guiding instructional changes.

Table 55: Comparison of SAT10 NCEs for Discovery and Non-Discovery Schools for Language Arts.

Birmingham Case Study, Alabama Language Arts							
	Discovery Schools			Non-Discovery Schools			
	2003	2004	Change	2003	2004	Change	Compare
Grade 3	33.6	35.2	1.6	40.0	35.5	-4.5	**6.1**
Grade 4	53.7	55.7	2.0	56.0	53.0	-3.0	**5.0**
Grade 5	45.2	45.0	-0.2	45.4	45.0	-0.4	**0.2**
Grade 6	39.7	42.2	2.5	44.1	42.9	-1.2	**3.7**
Grade 7	41.0	39.7	-1.3	48.7	45.9	-2.8	**1.5**
Grade 8	40.0	39.9	-0.1	46.1	44.0	-2.1	**2.0**

Table 56: Comparison of SAT10 NCEs for Discovery and Non-Discovery Schools for Reading.

Birmingham Case Study, Alabama Reading							
	Discovery Schools			Non-Discovery Schools			
	2003	2004	Change	2003	2004	Change	Compare
Grade 3	22.9	32.0	9.1	31.8	33.2	1.4	**7.7**
Grade 4	37.3	37.8	0.5	42.8	39.4	-3.4	**3.9**
Grade 5	42.5	42.8	0.3	45.4	43.7	-1.7	**2.0**
Grade 6	25.7	27.0	1.3	33.2	35.1	1.9	**-0.6**
Grade 7	32.2	28.2	-4.0	44.0	37.8	-6.2	**2.2**
Grade 8	32.3	30.3	-2.0	42.3	36.8	-5.5	**3.5**

Table 57: Comparison of SAT10 NCEs for Discovery and Non-Discovery Schools for Mathematics.

Birmingham Case Study, Alabama Mathematics							
	Discovery Schools			Non-Discovery Schools			
	2003	2004	Change	2003	2004	Change	Compare
Grade 3	28.3	33.3	5.0	37.3	34.3	-3.0	**8.0**
Grade 4	36.8	39.8	3.0	41.9	41.0	-0.9	**3.9**
Grade 5	41.0	42.1	1.1	42.4	42.4	0.0	**1.1**
Grade 6	24.8	26.5	1.7	33.0	32.6	-0.4	**2.1**
Grade 7	24.7	24.3	-0.4	36.0	32.0	-4.0	**3.6**
Grade 8	32.3	33.9	1.6	37.8	34.2	-3.6	**5.2**

Case Study Two: Metro Nashville Public Schools, Tennessee

During the 2004/2005 school year, sixty-five elementary and middle schools in Metro Nashville—representing over 20,000 students—used the Discovery Education Assessment: *Predictive Benchmark* program. Fifty-two elementary and middle schools—representing over 10,000 students—did not partner with Discovery Education Assessment. A comparison of the improvement in the percent of students at the Proficient/Advanced level from 2004 to 2005 for Reading and Mathematics is presented in tables 58 and 59. The results compare Discovery schools versus Non-Discovery schools in Metro Nashville. Discovery schools showed more improvement in AYP status from 2004 to 2005 when schools are combined and analyzed separately at the elementary and middle school level. Figures 3 and 4 illustrate the comparison between Discovery and Non-Discovery Schools.

Table 58: Comparison of Proficiencies for Discovery and Non-Discovery Schools for Reading.

Nashville Case Study, Tennessee Reading					
Grade Level	**Status**	**Year**	**N**	**% Prof/Adv**	**% Improve**
Combined	Discovery	2005	20,190	82.45	**10.75**
		2004	21,576	71.70	
	Non-Discovery	2005	10,167	88.41	**7.45**
		2004	10,143	80.96	
Elementary	Discovery	2005	5,217	86.20	**11.71**
		2004	5,640	74.49	
	Non-Discovery	2005	5,215	88.40	**8.91**
		2004	5,309	79.49	
Middle	Discovery	2005	14,948	81.19	**10.42**
		2004	15,917	70.77	
	Non-Discovery	2005	4,945	88.41	**5.80**
		2004	4,831	82.61	

Table 59: Comparison of Proficiencies for Discovery and Non-Discovery Schools for Mathematics.

Nashville Case Study, Tennessee Mathematics					
Grade Level	**Status**	**Year**	**N**	**% Prof/Adv**	**% Improve**
Combined	Discovery	2005	21,549	78.15	**7.21**
		2004	21,738	70.94	
	Non-Discovery	2005	10,490	83.43	**4.36**
		2004	10,172	79.07	
Elementary	Discovery	2005	5,765	80.59	**8.33**
		2004	5,702	72.26	
	Non-Discovery	2005	5,400	81.81	**5.21**
		2004	5,338	76.60	
Middle	Discovery	2005	15,759	77.31	**6.77**
		2004	16,017	70.54	
	Non-Discovery	2005	5,083	85.19	**3.45**
		2004	4,831	81.74	

Figure 3: Comparison of Proficiencies for Discovery and Non-Discovery Schools for Reading.

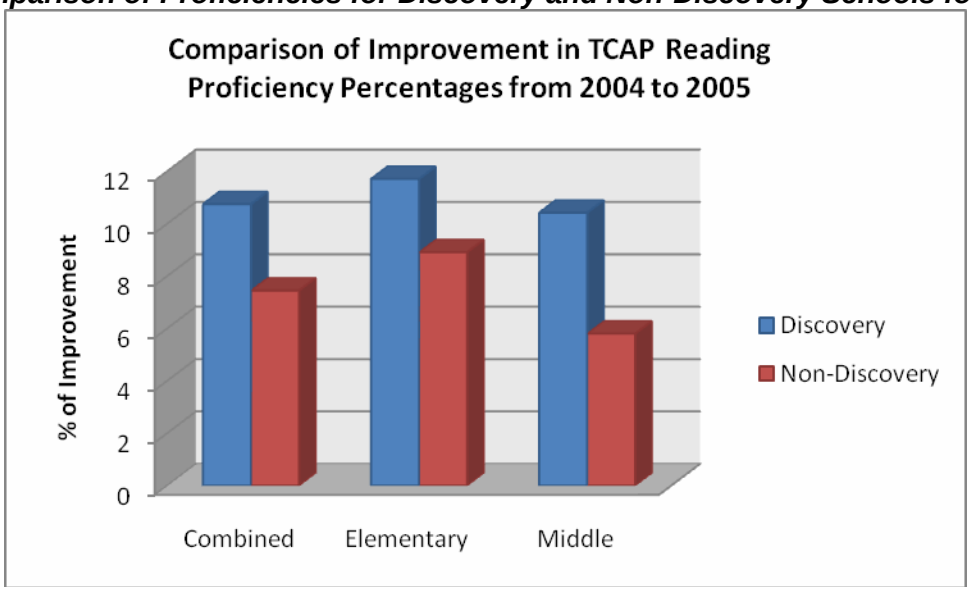
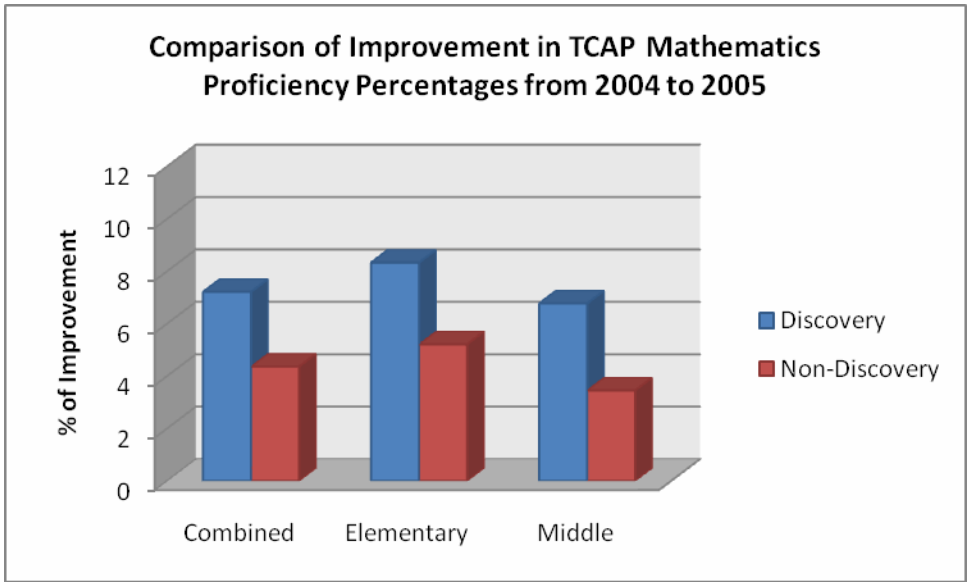


Figure 4: Comparison of Proficiencies for Discovery and Non-Discovery Schools for Mathematics.



(v) ensures experimental studies are presented in sufficient detail and clarity to allow for replication or, at a minimum, offer the opportunity to build systematically on their finding;

Consumers are encouraged to request additional data or further details for the examples listed in this overview. Discovery Education Assessment also compiles *Technical Manuals* specific to each school district and/or state. Accumulated data are of sufficient detail to permit adequate psychometric analyses, and their results have been consistently replicated across school districts and states. Past documents of interest include among others: "A Multi-State Comparison of Proficiency Predictions for Fall 2006" and "A Multi-State Look at 'What is Predictive Assessment?'." Furthermore, the "What is Predictive Assessment?" series of documents is available for multiple states.
Please check the Discovery website www.discoveryeducation.com/products/assessment/ for document updates.

(vi) has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective and scientific review;

TEST AND QUESTION STATISTICS, RELIABILITY, AND PERCENTILES

The following section reports test and question statistics, reliability, and percentiles for two benchmark tests, one each in grade 3 Reading and Mathematics. These benchmark tests were administered in Tennessee in Spring of 2008. These two benchmark tests are representative samples of over 1400 benchmark tests developed by Discovery Education Assessment. Benchmark tests are revised each year based on test and question statistics, particularly low item discrimination indices and significant DIF.

The following statistics are reported:

Number of Students:	Number of students used for calculation of test statistics.
Number of Items:	Number of items in each benchmark test (including common items used for scaling purposes).
Mean:	Test mean in terms of number correct.
Standard Deviation:	Test standard deviation.
Reliability:	Cronbach's alpha.
SEM:	Standard Error of Measurement (SEM) for the test.
Scale Score:	Discovery Education Assessment Scale Score for each number correct (Scale scores are vertically scaled using Rasch measurement. Scale scores from grades K-12 range from 1000 to 2000).
Percentiles:	Percentage of students below each number correct score
Question P-values:	The proportion correct for each item.
Biserial:	Item discrimination using biserial correlation.
Rasch Item Difficulty:	Rasch item difficulty parameter calculated using WINSTEPS.
DIF Gender:	Rasch item difficulty difference (Male vs. Female).
DIF Ethnicity:	Rasch item difficulty difference (White vs. Black).
DIF Size	
Negligible:	0 logits to .42 logits (absolute value).
Moderate:	.43 logits to .63 logits (absolute value).
Large:	.64 logits and up (absolute value).

(see p.1070 "An Adjustment for Sample Size in DIF Analysis", Rasch Measurement Transactions, 20:3, Winter 2006)

Technical Data Tennessee Spring Test 2007/2008 Reading Grade 3	
Test Statistics	
Number of Students	14,676
Number of Items	28
Mean	17.33
Standard Deviation	5.48
Reliability	0.83
Std Error Measurement	2.26

Scale Scores & Percentiles		
# Correct	Scale	Percentile
0	1063	1
1	1156	1
2	1213	1
3	1247	1
4	1273	1
5	1295	2
6	1313	3
7	1329	4
8	1344	7
9	1359	9
10	1372	12
11	1385	15
12	1398	19
13	1410	23
14	1422	27
15	1435	32
16	1447	37
17	1460	42
18	1473	48
19	1487	55
20	1502	62
21	1518	70
22	1535	77
23	1554	85
24	1576	91
25	1604	95
26	1640	98

Question Statistics					
Question	p-value	Biserial	Rasch Item Difficulty	DIF Gender	DIF Ethnicity
Q1	0.60	0.47	0.15	0.07	-0.29
Q2	0.57	0.45	0.32	-0.2	-0.4
Q3	0.31	0.21	1.64	0.01	0.26
Q4	0.73	0.46	-0.59	0.02	-0.11
Q5	0.76	0.47	-0.75	-0.22	-0.37
Q6	0.84	0.45	-1.42	-0.24	-0.72
Q7	0.60	0.51	0.13	0	0.04
Q8	0.61	0.45	0.08	-0.2	-0.11
Q9	0.70	0.49	-0.42	0	-0.21
Q10	0.81	0.5	-1.14	-0.07	-0.44
Q11	0.39	0.19	1.22	-0.22	0.46
Q12	0.43	0.32	1.03	-0.05	0.38
Q13	0.56	0.44	0.34	0.25	0.39
Q14	0.81	0.51	-1.11	0.03	-0.14
Q15	0.75	0.46	-0.73	0.24	-0.07
Q16	0.74	0.46	-0.61	0.15	-0.15
Q17	0.24	0.13	2.05	-0.11	0.58
Q18	0.47	0.34	0.78	0.05	0.23
Q19	0.56	0.34	0.33	-0.22	0.1
Q20	0.73	0.42	-0.6	0.13	0.06
Q21	0.70	0.56	-0.37	0.23	0
Q22	0.63	0.53	0	0.02	-0.12
Q23	0.71	0.45	-0.43	0.09	0.09
Q24	0.68	0.47	-0.26	0.05	-0.21
Q25	0.44	0.45	0.95	0.02	-0.52
Q26	0.73	0.48	-0.56	0.09	0.23

27	1697	99
28	1792	99

Q27	0.72	0.37	-0.51	-0.03	0.15
Q28	0.54	0.39	0.47	0.15	0.33

Technical Data
Tennessee Spring Test 2007/2008
Mathematics Grade 3

Test Statistics	
Number of Students	18,679
Number of Items	28
Mean	17.25
Standard Deviation	5.50
Reliability	0.83
Std Error Measurement	2.27

Scale Scores & Percentiles

# Correct	Scale	Percentile
0	1005	1
1	1099	1
2	1155	1
3	1190	1
4	1217	1
5	1238	1
6	1257	2
7	1274	3
8	1289	5
9	1303	8
10	1316	11
11	1329	15
12	1342	19
13	1354	24
14	1366	29
15	1378	34
16	1390	40
17	1403	46
18	1416	52
19	1429	58
20	1443	64
21	1458	71
22	1475	78
23	1494	84

Question Statistics

Question	p-value	Biserial	Rasch Item Difficulty	DIF Gender	DIF Ethnicity
Q1	0.83	0.37	-1.26	-0.17	-0.09
Q2	0.30	0.24	1.71	0.2	0.63
Q3	0.80	0.39	-1.05	0.28	0.03
Q4	0.48	0.45	0.75	-0.18	0.07
Q5	0.58	0.39	0.25	0.18	0.06
Q6	0.62	0.4	0.05	0.07	0.27
Q7	0.69	0.43	-0.35	0.28	0.28
Q8	0.51	0.35	0.60	0.18	0.32
Q9	0.58	0.37	0.21	-0.24	-0.33
Q10	0.64	0.42	-0.09	0.15	0.23
Q11	0.40	0.26	1.14	0.01	0.44
Q12	0.86	0.41	-1.54	0.2	-0.02
Q13	0.48	0.51	0.74	0.11	-0.22
Q14	0.67	0.48	-0.24	-0.05	-0.21
Q15	0.53	0.47	0.50	-0.39	-0.32
Q16	0.56	0.46	0.36	-0.13	0.03
Q17	0.73	0.46	-0.59	-0.32	-0.03
Q18	0.68	0.44	-0.31	-0.1	-0.03
Q19	0.71	0.48	-0.48	0.03	-0.26
Q20	0.67	0.4	-0.26	0.15	-0.13
Q21	0.38	0.31	1.23	0.09	0.34
Q22	0.67	0.43	-0.25	0.13	0.1
Q23	0.50	0.5	0.63	0.02	-0.02

24	1515	89
25	1541	93
26	1576	97
27	1633	99
28	1727	99

Q24	0.55	0.47	0.36	-0.13	-0.47
Q25	0.66	0.37	-0.18	-0.28	-0.24
Q26	0.66	0.46	-0.20	-0.22	0
Q27	0.84	0.43	-1.41	0.12	-0.11
Q28	0.68	0.33	-0.32	0.16	0.02

DISSEMINATION OF RESEARCH

Discovery Education Assessment tests and results have been incorporated and analyzed in the following publications, conference proceedings, dissertations, research documents, and tests:

1. Publications

Shrago, J. B., & Smith, M.K. (2006). Online assessment in the K-12 classroom: formative assessment model for improving student performance on standardized tests. In S. Howell & M. Hricko (Eds.), *Online assessment and measurement: case studies from higher education, K-12 and corporate* (pp. 181-194). Hershey, PA: Information Science Publishing.

2. Conference Proceedings

Shrago, J. B. chair. (2006, June). *Perspectives on large-scale formative assessment*. Presented at *36th annual nation conference on large-scale assessment hosted by the Council of Chief State School Officers.* San Francisco, CA.

Hass, J. (2006, June). *Algebra I pilot project: West Virginia department of education*. Presented at *36th annual nation conference on large-scale assessment hosted by the Council of Chief State School Officers.* San Francisco, CA.

Smith, M. K. (2006, June). *How can large scale formative assessment be research-based and valid?* Presented at *36th annual nation conference on large-scale assessment hosted by the Council of Chief State School Officers.* San Francisco, CA.

Thompson, E. (2006, June). *Selecting a formative reading assessment: guiding classroom reading instruction and intervention strategies*. Presented at *36th annual nation conference on large-scale assessment hosted by the Council of Chief State School Officers.* San Francisco, CA.

Vaughn-Neely, E., & Reed, M. (2005). *Reading findings*. Presented at *Society for Research & Child Development*. Atlanta, GA.

Vaughn-Neely, E., & Reed, M. (2006). *Reading findings*. Presented at *Society on Scientific Study of Reading*. Toronto, CA.

3. Dissertations

Johnson, J. (2005). *A multivariate analysis of the effects of the transition from elementary to middle school on the mathematics academic performance, personal achievement goal orientations, and achievement-related beliefs, perceptions and strategies of fifth grade students*. Unpublished doctoral dissertation, Union University, Jackson, TN.

4. Research Documents

Shrago, J. B., & Smith, M. K. (2006). The uses of benchmark tests to improve student learning. Nashville, TN: Discovery Education.

Smith, M. K. (2006). *Case study: Birmingham city school district, Ala.* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *Milwaukee Public Schools: 2008 Test Technical Manual f or Benchmark Assessment System.* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: Alabama?* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: Florida?* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: Illinois?* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: Kentucky?* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: New York?* Nashville, TN: Discovery Education.

Smith, M. K., & Kurz, A. (2008). *What is predictive assessment: Tennessee?* Nashville, TN: Discovery Education.

5. Tests

Discovery Education Benchmark Assessments. *Kentucky test b: reading language grade eight.* (2007). Nashville, TN: Discovery Education.