What is Predictive Assessment?

FLORIDA EXECUTIVE SUMMARY

1. Are Discovery Education Predictive Assessments reliable?

These benchmark assessments are highly reliable. For Grades 3 to 10 Reading tests over three time periods (Fall, Winter, Spring), the median reliability was .82 with a median sample size of 3,119. The median Mathematics reliability was .82 with a sample size of 3,167. The median Science reliability was .76 with a median sample size of 4,178.

2. Do Discovery Education Predictive Assessments have content validity?

These benchmark assessments model the objectives and skills in the Sunshine State Standards for Reading, Mathematics, and Science.

3. Do Discovery Education Predictive Assessments match state standardized tests?

The Gilchrest County school system participated in a criterion validity study during the school year 2006-2007. Individual student scores from the 2007 FCAT administration were obtained from the school system. The median correlation for the Reading assessments was .73 and the median correlation for the Mathematics assessments was .77. All correlations were significant at $p < .01$. There is substantial evidence that total scores on Discovery Education Predictive Assessments predict scale scores on the FCAT for both Reading and Mathematics.

4. Can Discovery Education Predictive Assessments predict proficiency levels?

Yes, there is a greater than 80% accuracy rate for predicting state proficiency levels. Approximately 6,800 students in the Putnam County School system participated in a proficiency prediction study during the 2006-2007 school year. The median Proficiency Prediction Score for Reading was 86.49%, and the median Proficiency Prediction Score for Mathematics was 94.5%.

5. Can the use of Discovery Education Predictive Assessments improve student learning?

Many factors contribute to the improvement of student learning. A comparison of improvement in student proficiency from the 2006 FCAT to the 2007 FCAT was conducted for Gilchrest County. Bell School had significant improvement in Grades 3, 6, 8, and 9 Reading and in Grades 3, 6, 7, 8, and 9 Mathematics. Trenton School had significant improvement in Grades 4, 7, and 10 Reading and in Grades 7 and 9 Mathematics.

6. Can Discovery Education Predictive Assessments be used to measure growth over time?

Yes. These benchmark assessments are scored on a vertical scale using state-of-the-art Rasch psychometric modeling. Thus, reliable estimates of student growth can be made over time.

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

Two matched control group studies—one in Birmingham, Alabama, and the other in Nashville, Tennessee—support the claim that Discovery Education Predictive Assessments help schools demonstrate significant improvement in student proficiency.

What is Predictive Assessment?

FLORIDA

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

Since its inception in 2000 by Vanderbilt University, ThinkLink Learning, now a part of Discovery Education, 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 through ThinkLink Learning has 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's Predictive State-Specific 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.

Furthermore, 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. This document, "What is Predictive Assessment?", outlines how Discovery Education Assessment addresses the following quality testing standards:

1. Are Discovery Education Predictive 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 Education Predictive 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.

3. Do Discovery Education Predictive 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 Education Predictive 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 Education Predictive 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 Education Predictive 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 Education Predictive 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 Education Predictive 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, Table 2, and Table 3 present test reliabilities and sample sizes for Discovery Education Predictive Assessments for three time periods—Fall, Winter, and Spring—in the subject areas of Reading, Mathematics, and Science.

The median Reading reliability was .82 with a median sample size of 3,119. The median Mathematics reliability was .82 with a sample size of 3,167. The median Science reliability was .76 with a median sample size of 4,178.

Table 1: Test Reliabilities for Reading and Mathematics Fall 2006 and Science Fall 2007.

	Fall 2006				Fall 2007	
	Reading	N	Mathematics	N	Science	N
Grade 2	0.87	3,240	0.85	3,249		
Grade 3	0.85	3,202	0.81	3,201	0.82	3,919
Grade 4	0.82	3,192	0.83	3,198	0.72	4,625
Grade 5	0.79	2,937	0.81	2,898	0.78	7,762
Grade 6	0.82	2,761	0.82	2,737	0.79	4,786
Grade 7	0.79	2,679	0.78	2,798	0.80	5,754
Grade 8	0.83	2,785	0.83	2,899	0.72	8,842
Grade 9	0.74	2,586	0.82	2,698		
Grade 10	0.65	2,821	0.82	2,655		

Table 2: Test Reliabilities for Reading and Mathematics Winter 2007 and Science Winter 2008.

	Winter 2007				Winter 2008	
	Reading	N	Mathematics	N	Science	N
Grade 2	0.81	3,699	0.81	3,717		
Grade 3	0.83	4,518	0.79	4,388	0.74	3,216
Grade 4	0.84	4,093	0.8	4,163	0.73	4,104
Grade 5	0.81	4,229	0.8	4,286	0.78	7,510
Grade 6	0.83	2,842	0.8	2,853	0.62	4,006
Grade 7	0.81	2,730	0.81	2,694	0.73	5,072
Grade 8	0.89	2,955	0.82	2,946	0.71	8,051
Grade 9	0.77	2,403	0.81	2,622		
Grade 10	0.79	2,891	0.81	2,775		

Table 3: Test Reliabilities for Reading and Mathematics Spring 2007 and Science Winter 2008.

	Spring 2007				Spring 2008	
	Reading	N	Mathematics	N	Science	N
Grade 2	0.85	3,638	0.81	3,690		
Grade 3	0.83	4,350	0.85	4,353	.79	1,566
Grade 4	0.83	3,686	0.86	3,743	.75	2,392
Grade 5	0.82	4,212	0.82	4,301	.71	4,251
Grade 6	0.82	3,320	0.84	3,348	.76	2,305
Grade 7	0.86	3,119	0.86	3,167	.79	2,430
Grade 8	0.79	3,394	0.86	3,382	.79	4,004
Grade 9	0.83	2,312	0.84	2,308		
Grade 10	0.82	2,537	0.86	2,449		

CONTENT VALIDITY

2. Do Discovery Education Predictive 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. Discovery Education Assessment hereby employs one of the leading alignment research methodologies, the **Webb Alignment Tool (WAT)**, which has continually 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 thus **state specific** and feature **matching reporting categories** of a given state's large-scale assessment used for accountability purposes.

The following objectives are used on Discovery Education Predictive Assessments for Florida in Reading, Mathematics, and Science. These objectives and reporting categories are based on Sunshine State Standards. We continually update our assessments to reflect the most current version of a state's standards.

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

Florida Math Reporting Categories	
Number Sense, Concepts, and Operations	Algebraic Thinking
Measurement	Data Analysis and Probability
Geometry and Spatial Sense	

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

CRITERION VALIDITY

3. Do Discovery Education Predictive 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 Predictive 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.

The Gilchrest County school system participated in a criterion validity study during the school year 2006-2007. Approximately 1500 students in grades 3 to 10 took Discovery Education Predictive Assessments. Individual student scores from the 2007 FCAT administration were obtained from the school system. Table 4 shows the correlation for Reading between Discovery Education Assessment and FCAT. Table 5 shows similar results for Mathematics. The median correlation for the Reading assessments is .73 and the median correlation for the Mathematics assessments is .77. All correlations are significant at $p < .01$. Thus, there is substantial evidence that total scores on Discovery Education Predictive Assessments predict scale scores on the FCAT for both Reading and Mathematics.

Table 6 shows correlations at the objective level for Reading, and Table 7 shows similar correlations at the objective level for Mathematics. Median correlations are 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 Education Predictive Assessments predict objective scale scores on the FCAT for both Reading and Mathematics.

Figure 1-4 illustrate these correlations comparing the means of the Discovery Education Assessment Reporting Categories with the respective FCAT Objectives. Figure 1 and 2 present comparisons for Reading Grades 4 and 7, and Figure 3 and 4 present comparisons for Mathematics Grades 4 and 7.

Table 4: Correlation of Discovery Education Assessment Reading Growth Score and FCAT Reading Scale Score.

Test B Discovery Education Assessment 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 5: Correlation of Mathematics Growth Score and FCAT Mathematics Scale Score.

Test B ThinkLink 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 6: Correlation of Reading Reporting Categories and FCAT Reading Objectives.

Test B ThinkLink 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 7: Correlation of Mathematics Reporting Categories and FCAT Mathematics Objectives.

Test B ThinkLink 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*

Figure 1: Discovery Education Assessment Reading Reporting Categories versus FCAT Reading Objectives.

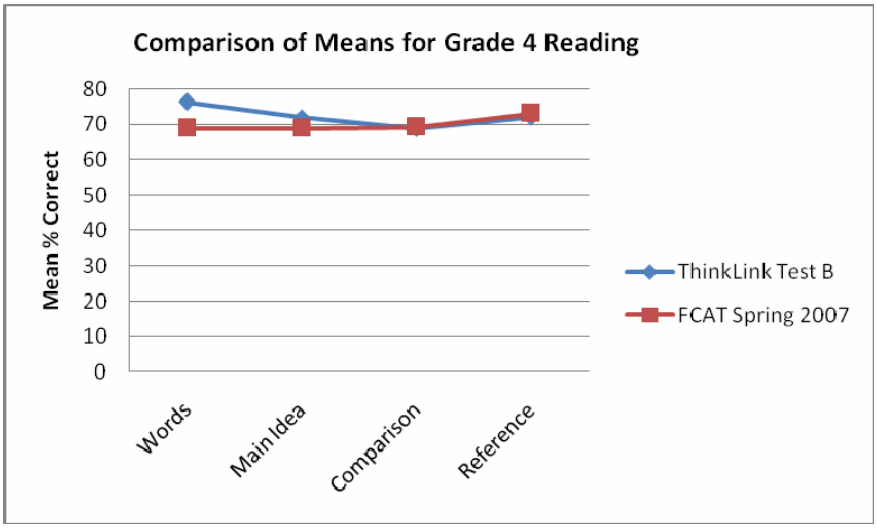


Figure 2: Discovery Education Assessment Reading Reporting Categories versus FCAT Reading Objectives.

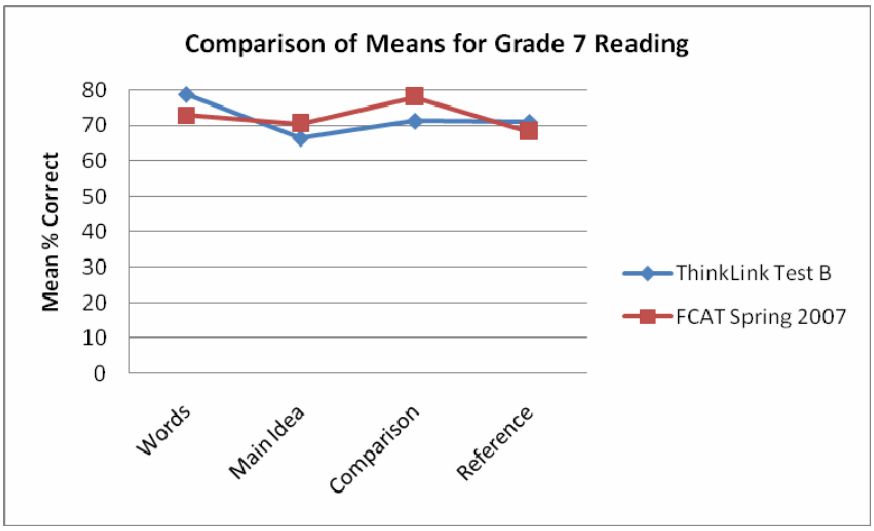


Figure 3: Discovery Education Assessment Reading Reporting Categories versus FCAT Mathematics Objectives.

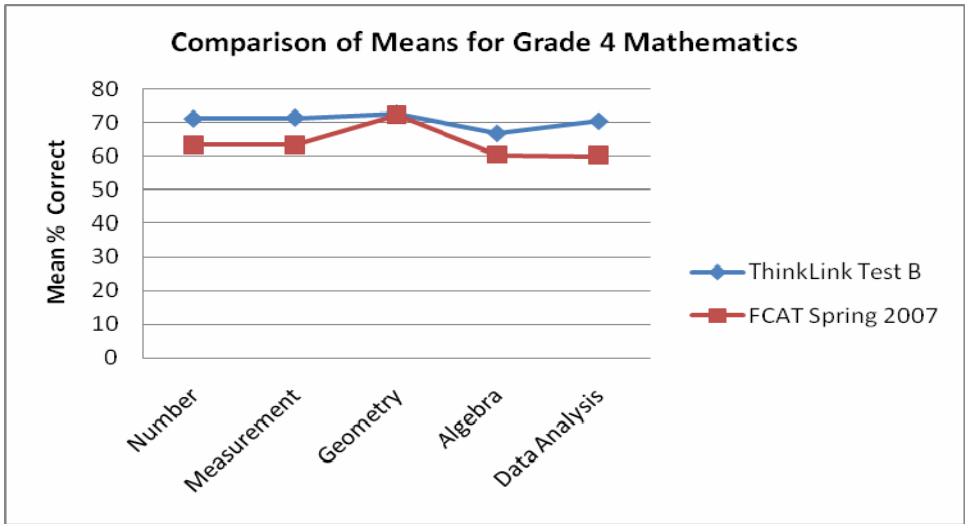
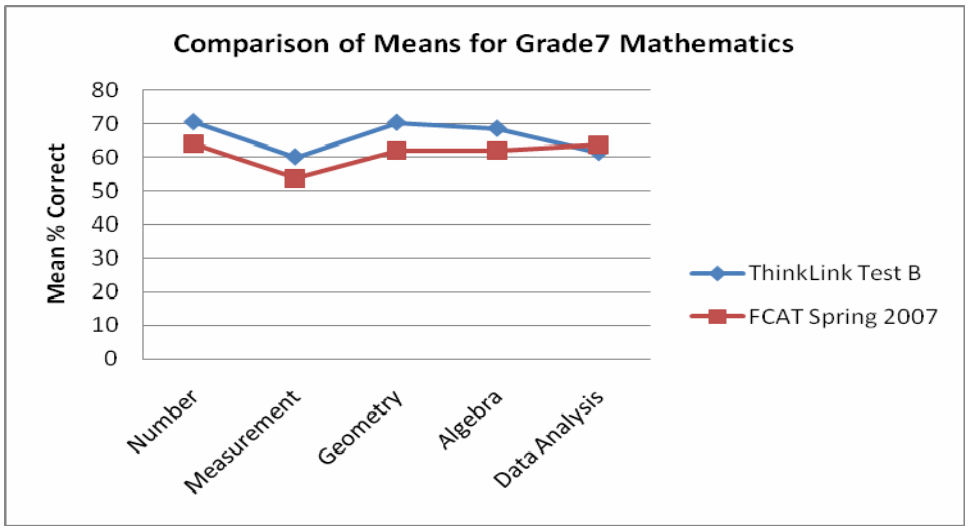


Figure 4: Discovery Education Assessment Reading Reporting Categories versus FCAT Mathematics Objectives.



PROFICIENCY PREDICTIVE VALIDITY

4. Can Discovery Education Predictive 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 is 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 Florida 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 - 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 (the symbols / /) of the error rate is used to account for cases where Discovery Education Assessment overpredicts 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% - 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.

Case Study: Putnam County Schools

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 8. The median Proficiency Prediction Score for Reading was 86.49%, and the median Proficiency Prediction Score for Mathematics was 94.5%.

Table 8: 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%**

To illustrate these Proficiency Prediction Scores, Figure 5-10 show the percentage of students at each proficiency level for Discovery Education Assessment and FCAT for Grades 3, 8, and 10 Reading, and Grades 3, 7, and 10 Mathematics.

Figure 5: Results of Proficiency Prediction Study in Putnam County for Grade 3 Reading.

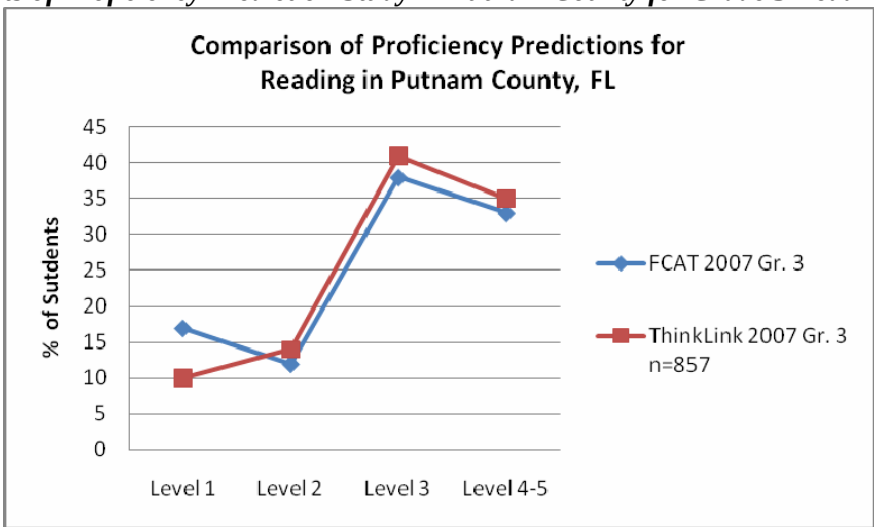


Figure 6: Results of Proficiency Prediction Study in Putnam County for Grade 8 Reading.

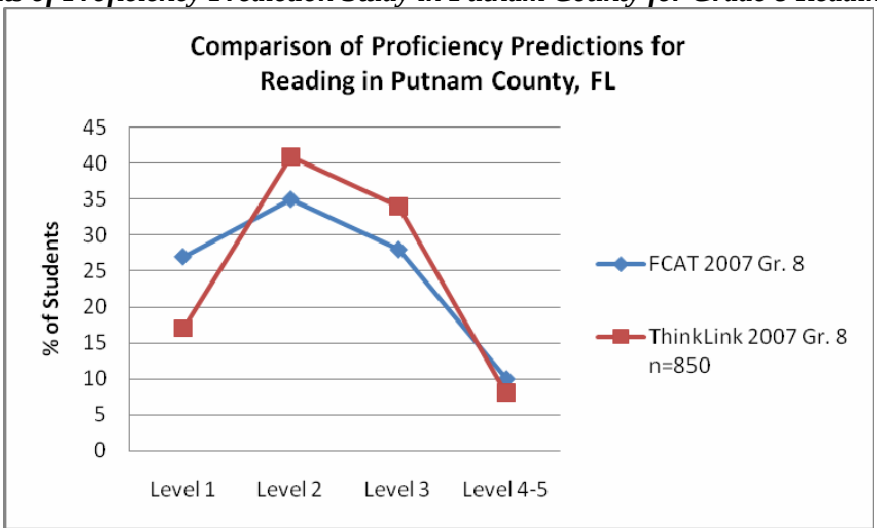


Figure 7: Results of Proficiency Prediction Study in Putnam County for Grade 10 Reading.

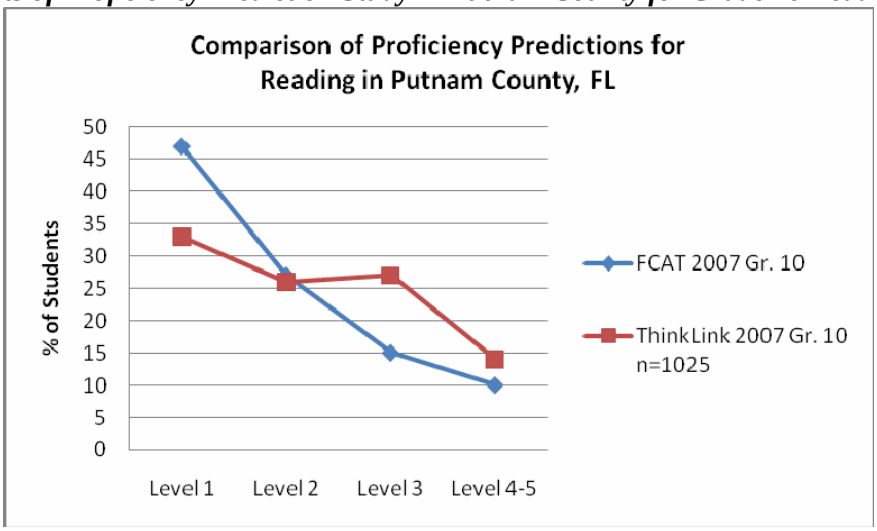


Figure 8: Results of Proficiency Prediction Study in Putnam County for Grade 3 Mathematics.

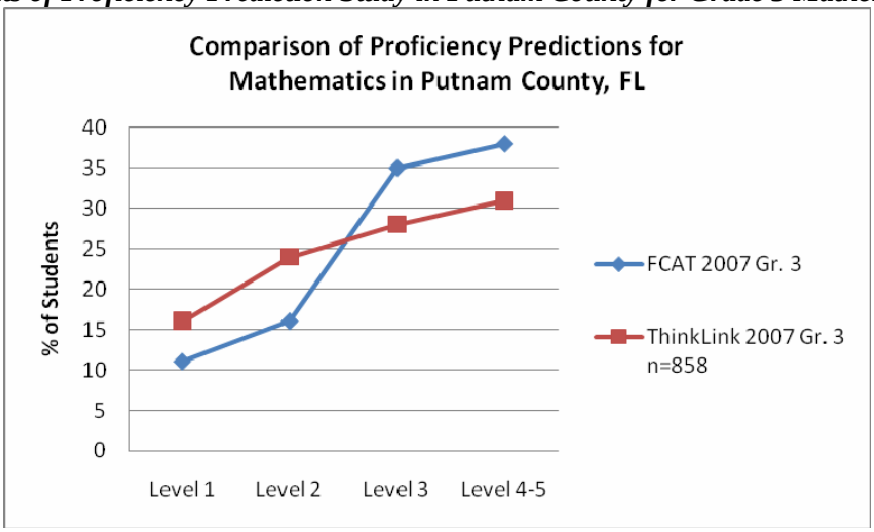


Figure 9: Results of Proficiency Prediction Study in Putnam County for Grade 7 Mathematics.

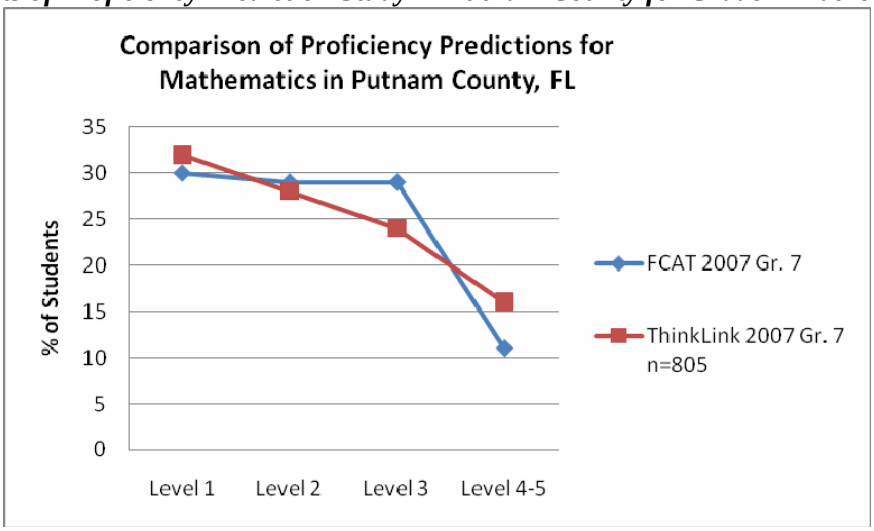
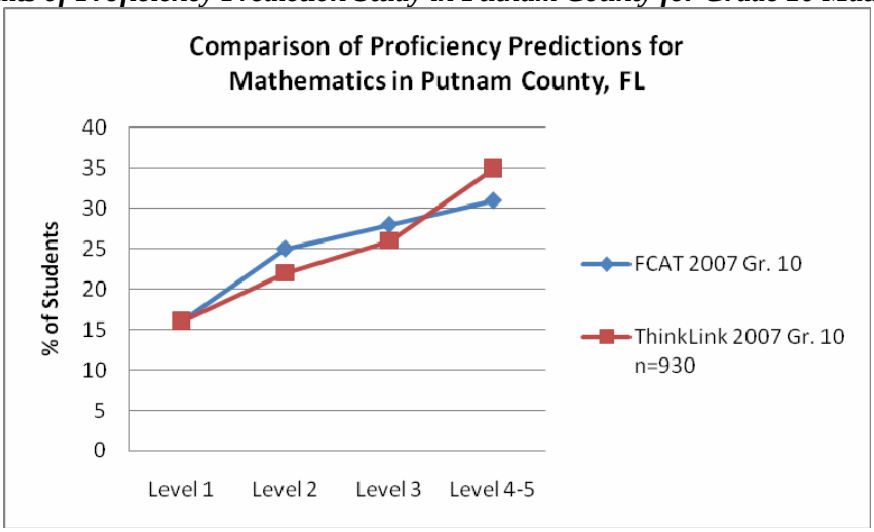


Figure 10: Results of Proficiency Prediction Study in Putnam County for Grade 10 Mathematics.



Proficiency Success Rate

When individual student data are available, an additional measure, the *Proficiency Success Rate*, can also be calculated. After taking a Discovery Education 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). When FCAT results are received, a comparison of this prediction with actual FCAT status can be made. The percentage of students predicted as proficient by Discovery Education Assessment that actually are 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 result on the FCAT.

Case Study: Gilchrest County

The Gilchrest 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 10 and compared with proficiency predictions on Discovery Education Predictive Assessments. Table 9 and Table 10 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 9: Results of the Proficiency Success Rate Study in Gilchrest County for Reading.

Proficiency Success Rate in Gilchrest 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 10: Results of the Proficiency Success Rate Study in Gilchrest County for Mathematics.

Proficiency Success Rate in Gilchrest 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%**

CONSEQUENTIAL VALIDITY

5. Can the use of Discovery Education Predictive 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.

Once again, the Gilchrest County School System participated in a consequential validity study. This system used Discovery Education Predictive 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 Gilchrest County, the Bell schools Table 11 and 12) and Trenton schools (Table 13 and 14). 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 Florida was compared to this Difference score.

Table 11: Results of Consequential Validity Study for Bell Schools in Reading.

Bell Elementary, Middle, and High School in Gilchrest 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 12: Results of Consequential Validity Study for Bell Schools in Mathematics.

Bell Elementary, Middle, and High School in Gilchrest 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 13: Results of Consequential Validity Study for Trenton Schools in Reading.

Trenton Elementary, Middle, and High School in Gilchrest 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 14: Results of Consequential Validity Study for Trenton Schools in Mathematics.

Trenton Elementary, Middle, and High School in Gilchrest 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 Education Predictive 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.

The Bell schools had significant improvement in Grades 3, 6, 8, and 9 Reading and in Grades 3, 6, 7, 8, and 9 Mathematics (see Figure 11 and 12). The Trenton schools had significant improvement in Grades 4, 7, and 10 Reading and in Grades 7 and 9 Mathematics (see Figure 13 and 14).

Figure 11: Selective Comparison of Improvement Results for Bell and FL State in Reading.

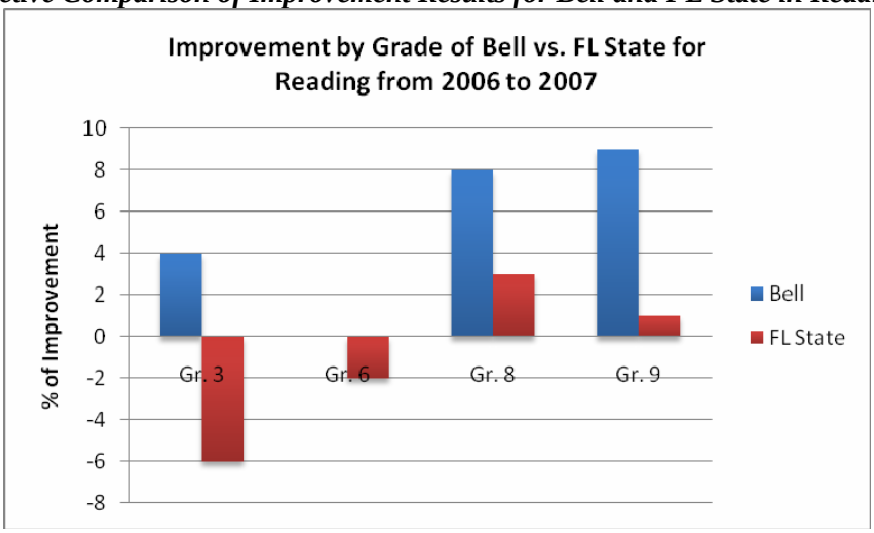


Figure 12: Selective Comparison of Improvement Results for Bell and FL State in Mathematics.

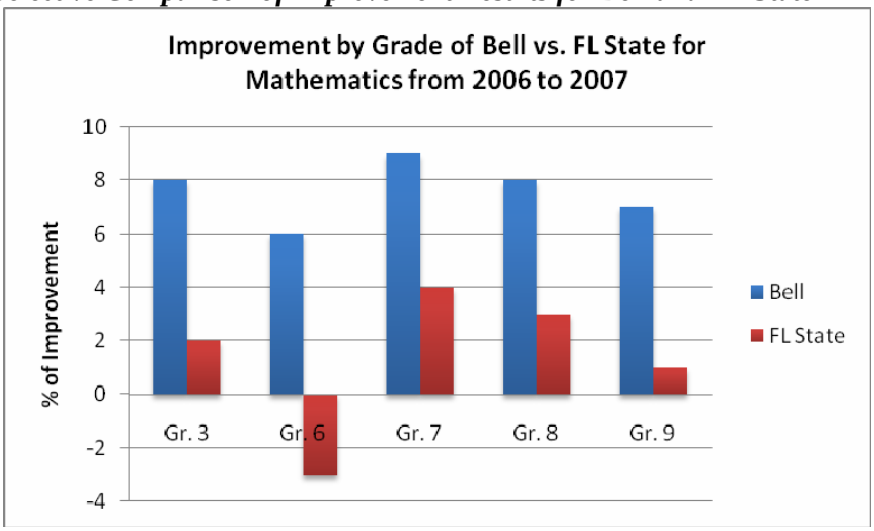


Figure 13: Selective Comparison of Improvement Results for Trenton and FL State in Reading.

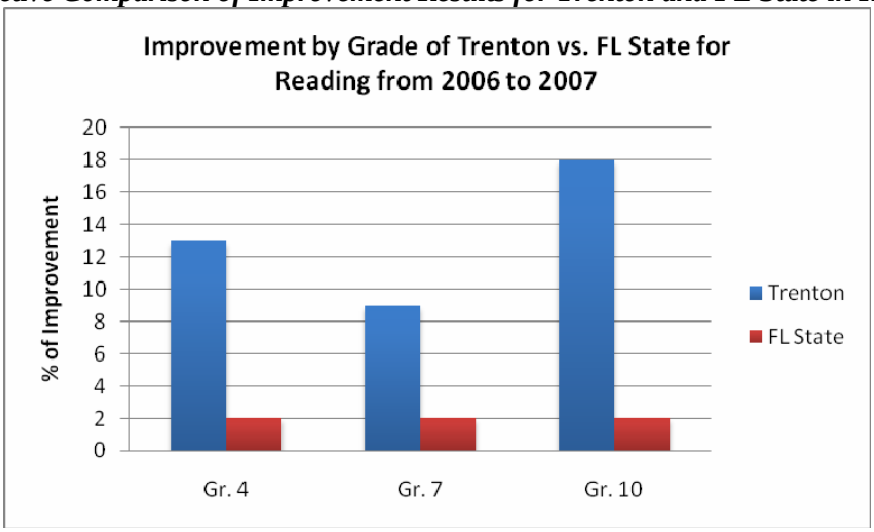
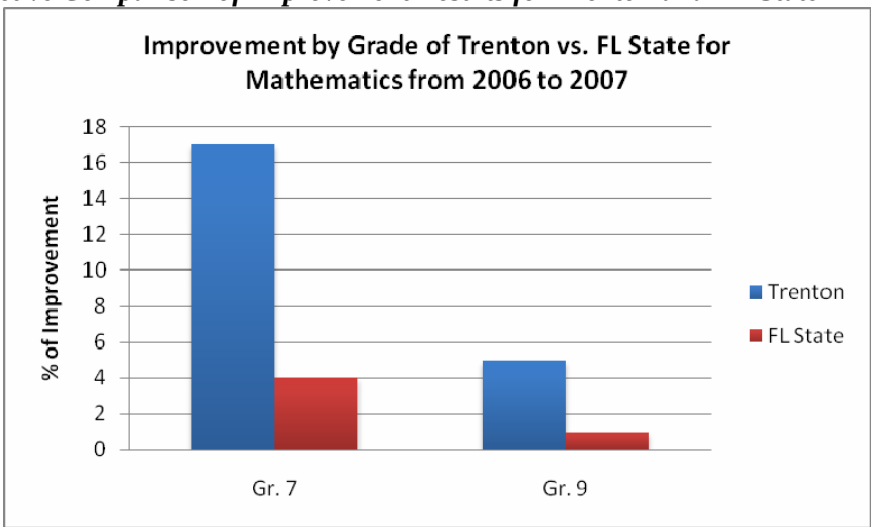


Figure 14: Selective Comparison of Improvement Results for Trenton and FL State in Mathematics.



GROWTH MODELS

6. Can Discovery Education Predictive 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.

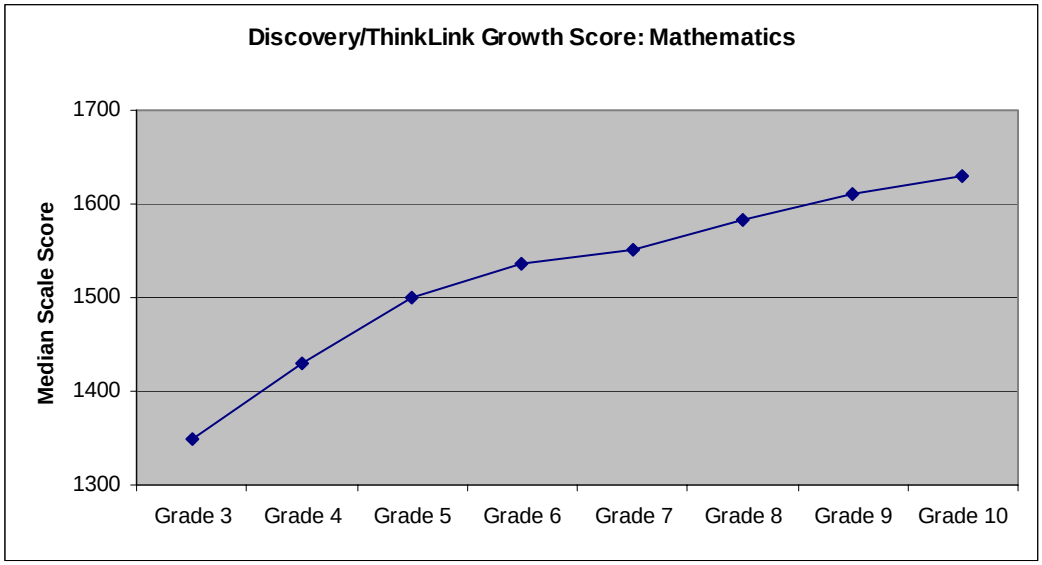
The accurate measurement of student achievement over time is becoming increasingly important to parents, teachers, and school administrators. **Student "growth" within a grade and across grades** has also been sanctioned by the U. S. Department of Education as a reliable way to measure student proficiency 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 Education Assessment Benchmark tests 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).

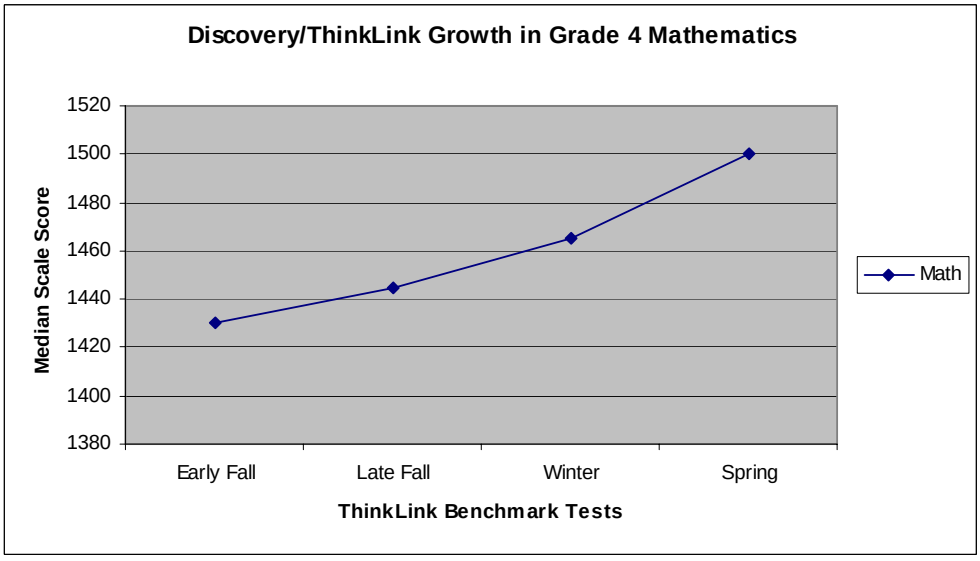
Why use such a rigorous vertical scale?
Isn't student growth similar across grades? Don't students change as much from Grade 3 to Grade 4 as they do from Grade 7 to Grade 8? Previous research on the use of vertical scales has demonstrated that **student growth is not linear**; that is, growth in student achievement is different from grade to grade (see Young 2006). For instance, Figure 15 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 Education Assessment'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 15: 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 Education Assessment benchmark tests 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 Education Assessment benchmark tests provide normative information to assess student growth multiple times each year. Figure 16 illustrates this growth for Grade 4 Mathematics using our benchmark assessments.

Figure 16: Within-Year Growth Results for Discovery Education Assessment Mathematics Tests.



Florida Growth Scale

The following tables and figures 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 15: 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 16: 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

Figure 17: Test Difficulty Comparison on Vertical Scale for 0708 Reading Tests.

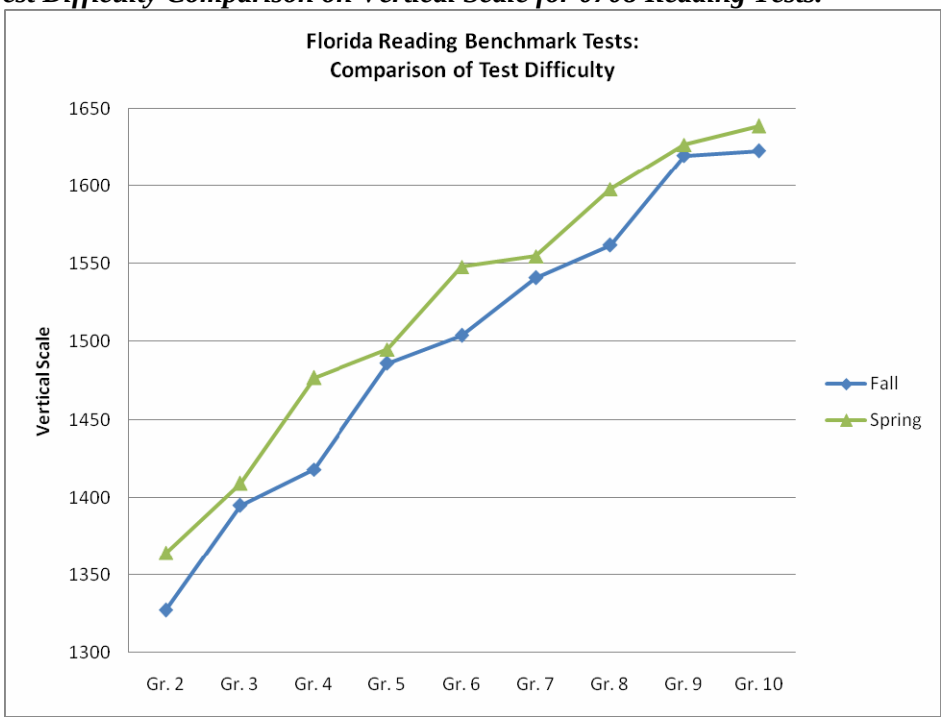
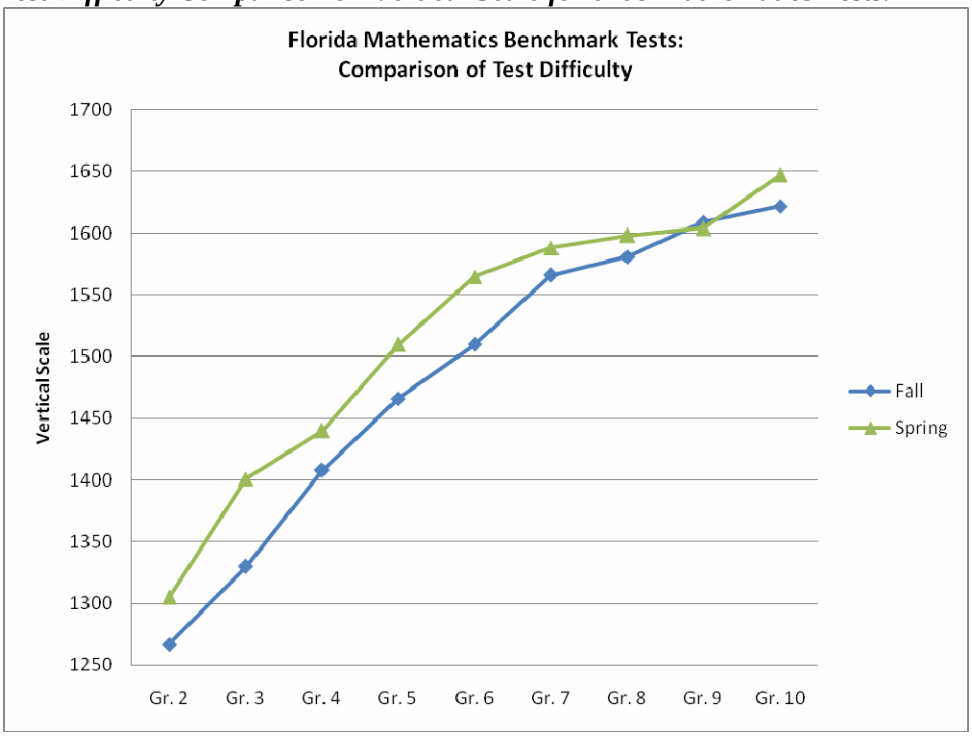


Figure 18: Test Difficulty Comparison on Vertical Scale for 0708 Mathematics Tests.



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, Alabama City Schools

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 schools in Birmingham City used Discovery Education Predictive Assessments whereas the other half did not. 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 the Discovery Education Predictive Assessments 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:

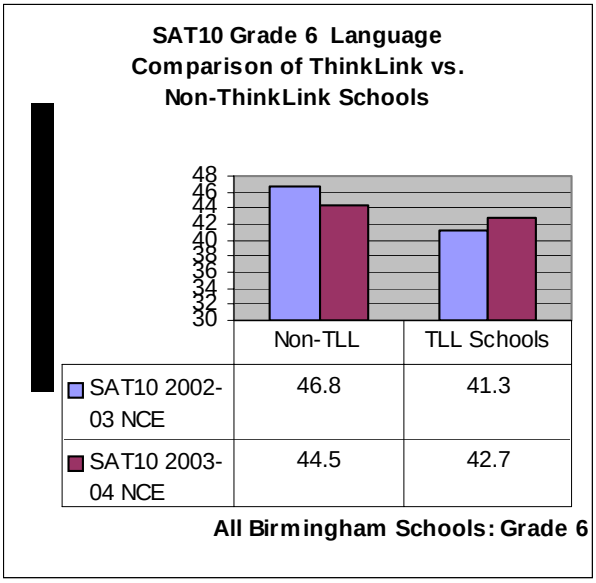
This study is based on data from elementary and middle schools in the City of Birmingham, Alabama. In 2002-03, no Birmingham Schools used Discovery Education's Predictive Assessment Series. Starting in 2003-04, 20 elementary and 9 middle schools used the Discovery Education Assessment program. All Birmingham 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 of Alabama's School Accountability Program. The State of Alabama uses improvement in SAT10 percentiles to gauge school progress and as

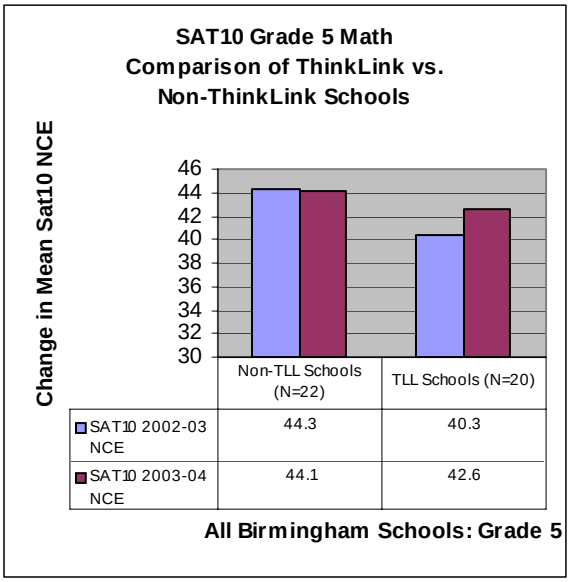
part of its NCLB reporting. National percentiles on the SAT10 are reported by subject and grade level. A single national percentile is reported for all students within a subject and grade level (this analysis is subsequently referred as ALL STUDENTS). Furthermore, national percentiles are disaggregated by various subgroups within a school. For the comparisons that follow, the national percentiles for students classified as utilizing free and reduced lunch (referred to below as POVERTY) were used. All percentiles have been converted to Normal Curve Equivalents (NCE) to allow for averaging of results.

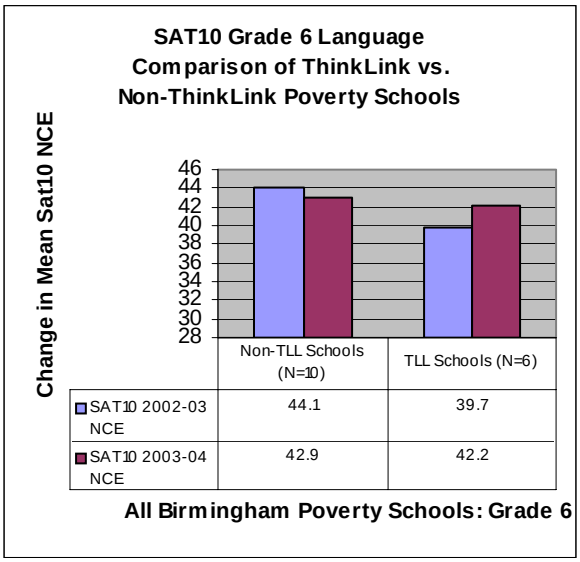
The Discovery Education Assessment schools comprise the experimental group in this study. The Birmingham schools that did not use Discovery Education Assessment comprise the matched comparison group. The following charts show SAT10 National Percentile changes for ThinkLink Schools vs. Non-ThinkLink Schools in two grades levels (Grades 5 and 6) for three subjects (Language, Mathematics, and Reading) for two groups of students (ALL STUDENTS and POVERTY students). In general, there was a significant decline or no improvement in SAT10 scores from 2002-03 to 2003-04 for most non-ThinkLink schools. This trend however did not happen in the schools using Discovery Education Assessment: instead, there was a marked improvement with most grades scoring increases in language, math and reading. In grade levels where there was a decline in Discovery Education Assessment schools, it was a much lower decline in scores when compared to those schools that did not use Discovery Education Assessment.

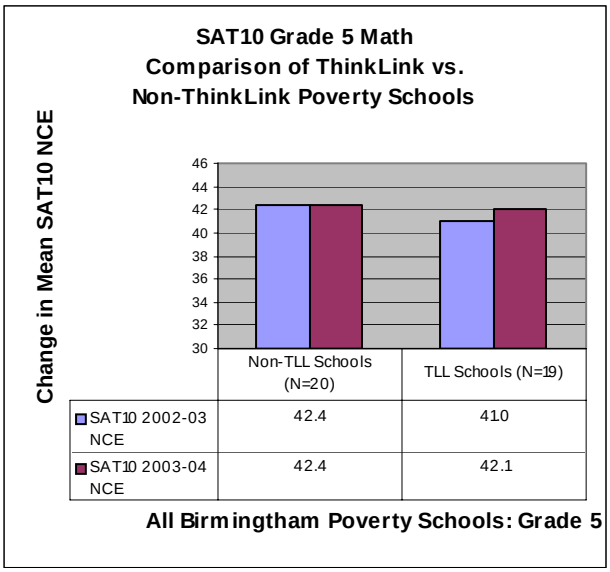
As a result of the improvement that many of these schools made in school year 2003-04, the Birmingham City Schools selected Discovery Education Assessment to be used with *all* of the schools in school year 2004-05. The Birmingham City Schools also chose to provide professional development in each school to help all teachers become more familiar with the concepts of standardized assessment and better utilize data to focus instruction.

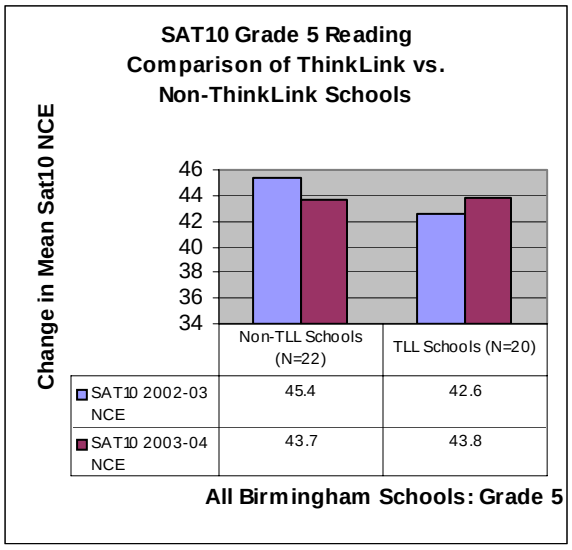
Discovery Education Assessment
RESEARCH

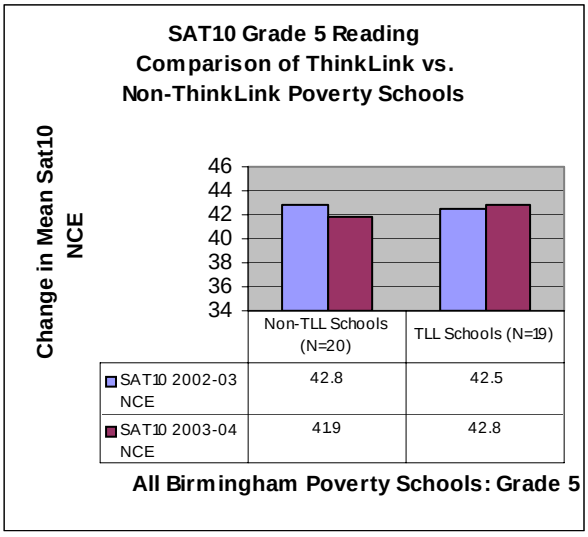






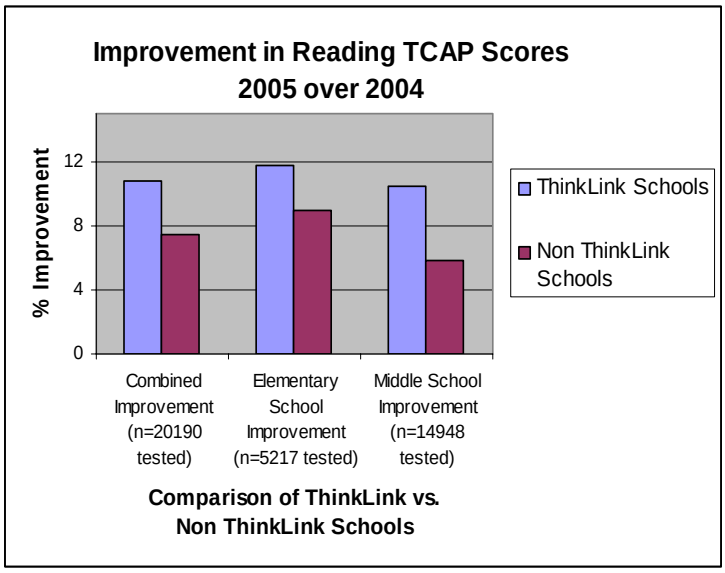


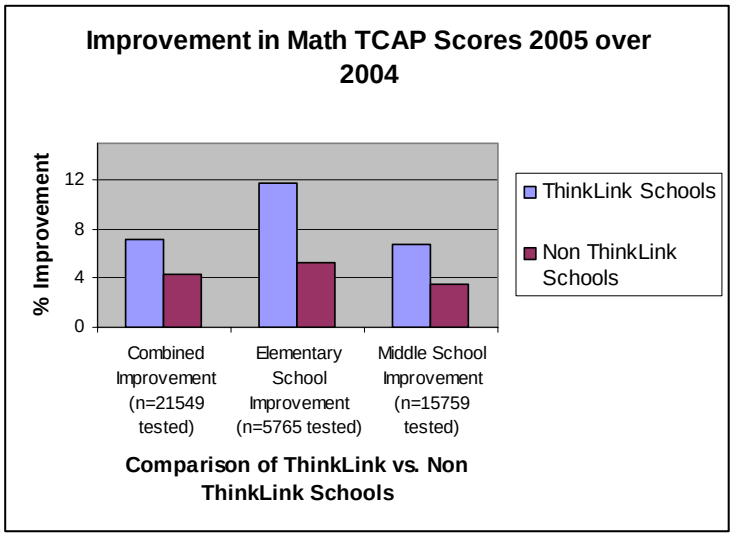

SAT10 Grade 5 Reading
Comparison of ThinkLink vs.
Non-ThinkLink Schools
Change in Mean Sat10 NCE
46
44
42
40
38
36
34
Non-TLL Schools (N=22)
TLL Schools (N=20)
SAT10 2002-03 NCE
45.4
42.6
SAT10 2003-04 NCE
43.7
43.8
All Birmingham Schools: Grade 5


SAT10 Grade 5 Reading
Comparison of ThinkLink vs.
Non-ThinkLink Poverty Schools
Change in Mean Sat10 NCE
46
44
42
40
38
36
34
Non-TLL Schools (N=20)
TLL Schools (N=19)
SAT10 2002-03 NCE
42.8
42.5
SAT10 2003-04 NCE
41.9
42.8
All Birmingham Poverty Schools: Grade 5

Case Study Two: Metro Nashville, Tennessee City Schools

Metro Nashville schools that used Discovery Education Assessment made greater improvements in AYP than Metro Nashville schools that didn't use Discovery Education Assessment. During the 2004-2005 school year, sixty-five elementary and middle schools in Metro Nashville, representing over 20,000 students, used Discovery Education assessments. Fifty-two elementary and middle schools, representing over 10,000 students, did not use Discovery Education assessments. The improvement in the percent of students at the Proficient/Advanced level from 2004 to 2005 is presented in the graph below. The results compare ThinkLink schools versus non-ThinkLink schools in Metro Nashville. Discovery Education Assessment 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.





(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 ThinkLink website www.thinklinklearning.com 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;

Discovery Education Assessment tests and results have been incorporated and analyzed in the following peer- reviewed manuscripts and publications:

Jacqueline Shrago and **Dr. Michael Smith** of ThinkLink Learning contributed a chapter on formative assessment to "Online Assessment and Measurement: Case Studies from Higher Education, K-12, and Corporate" by Scott Howell and Mary Hicko in 2006.

Dr. Elizabeth Vaughn-Neely, Associate Professor, Dept. of Leadership & Counselor Education, Ole Miss University (eiv@olemiss.edu) and **Dr. Marjorie Reed**, Associate Professor, Dept. of Psychology, Oregon State University presented their peer-reviewed findings based on their joint research and work with schools using Discovery Education Assessment at:

Society for Research & Child Development, Atlanta, GA. April 2005
Kappa Delta Pi Conference, Orlando, FL November 2005
Society on Scientific Study of Reading, July 2006

Two dissertations for Ed.S studies have also been published:
Dr. Juanita Johnson, Union University (johnsonj4@k12tn.net)
Dr. Monica Eversole, Richmond KY (meversol@madison.k12.ky.us)

Please contact us for other specific information requests. We welcome your interest in the evidence supporting the efficacy of our Discovery Education Assessment tests.