What is Predictive Assessment?

TENNESSEE EXECUTIVE SUMMARY

1. Are Discovery Education Predictive Assessments reliable?

These benchmark assessments are highly reliable. For Grades 3 to 8 Reading tests over three time periods (Fall, Winter, Spring), the median reliability was .81 with a median sample size of 18437. The median Mathematics reliability was .81 with a sample size of 18252. The median Science reliability was .77 with a median sample size of 10924.

2. Do Discovery Education Predictive Assessments have content validity?

These benchmark assessments model the objectives and skills of the TCAP Assessment Standards for Reading, Mathematics, and Science.

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

The Grainger County school system participated in a criterion validity study for the 2006-2007 school year. Approximately 1500 students, Grades 3 through 8, completed the Discovery Education Predictive Assessments in Reading and Mathematics. For each grade and subject, the students' percent correct averages were broken down by reporting category and compared to their respective 2007 TCAP averages. The results indicated a close match between predicted and actual percent correct averages.

4. Can Discovery Education Predictive Assessments predict proficiency levels?

Yes, there is a greater than 90% accuracy rate for predicting combined state proficiency percentages. Due to our representativeness throughout the state of Tennessee, direct comparisons of Spring 2007 Test B and actual 2007 TCAP proficiency percentages ("Proficient" and "Advanced" combined) were made for Grades 3 through 8 in Reading and Mathematics. The median Proficiency Prediction Score for Reading was 96%, and the median Proficiency Prediction Score for Mathematics was 92%.

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 TCAP to the 2007 TCAP was conducted for Grainger County and the state of Tennessee. Grainger County had significant improvement (exceeding the state's improvement during the same time period) in Grades 3, 4, 5, 6, and 8 Mathematics and in Grades 3 and 4 Reading.

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?

TENNESSEE

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;

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 .81 with a median sample size of 18,437. The median Mathematics reliability was .81 with a median sample size of 18,252. The median Science reliability was .77 with a median sample size of 10,924.

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

Tennessee Test P – Fall 2007				
	Reading	N	Mathematics	N
Grade 3	.83	18,421	.81	18,428
Grade 4	.82	18,837	.83	18,700
Grade 5	.78	19,439	.83	19,347
Grade 6	.77	18,453	.78	18,252
Grade 7	.81	19,459	.78	19,562
Grade 8	.80	19,020	.80	19,020
Gateway			.78	1,281

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

Tennessee Test A – Winter 2007						
	Reading	N	Mathematics	N	Science	N
Grade 3	.79	15,309	.82	15328	.84	9,415
Grade 4	.76	15,239	.80	15212	.80	9,693
Grade 5	.80	16,561	.78	16497	.77	10,849
Grade 6	.79	15,756	.76	15594	.80	11,832
Grade 7	.82	14,628	.82	15051	.72	10,999
Grade 8	.78	14,654	.81	14266	.77	10,563
Gateway	.86	6,524	.77	9950	.84	3,991

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

	Tennessee Test B – Spring 2008					
	Reading	**N**	**Mathematics**	**N**	**Science**	**N**
Grade 3	.83	14,676	.83	18679	.75	11,991
Grade 4	.82	19,256	.83	19230	.75	12,874
Grade 5	.82	20,286	.82	20692	.76	8,465
Grade 6	.81	18,533	.77	18163	.78	14,211
Grade 7	.74	18,730	.83	18874	.72	14,402
Grade 8	.83	18,703	.85	18398	.74	13,832
Gateway	.89	5,746	.82	7751	.86	3,358

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 reporting categories are used on Discovery Education Predictive Assessments for Tennessee in Reading, Mathematics, and Science. They are based on TCAP Assessment Standards and thus mirror the TCAP reporting categories. We continually update our assessments to reflect the most current version of a state's standards.

TN 0708 Reading Reporting Categories	
Content	Vocabulary
Grammar Conventions	Writing/Organization
Meaning	Writing/Writing Process
Techniques and Skills	

TN 0708 Gateway Reading Reporting Categories	
Writing	Viewing/Representing & Listening/Speaking
Reading	

Discovery Education Assessment
RESEARCH

TN 0708 Mathematics Reporting Categories	
Number Sense/Number Theory	Data Analysis and Probability
Computation	Measurement
Algebraic Thinking	Geometry
Real World Problem Solving	Graphs and Graphing

TN 0708 Gateway Mathematics Reporting Categories	
Number Sense/Theory	Measurement
Algebraic Expressions	Data Analysis and Probability
Geometry	

TN 0708 Science Reporting Categories	
Structure and Function of Organisms	Life Cycles and Biological Changes
Ecology	Space, Weather, and Climate
Interactions Between Living Things and Their Environment	Motion and Forces, Forms of Energy
Diversity and Adaptation Among Living Things	Forces and Motion
Heredity and Reproduction	Interactions of Matter
Earth's Features and Resources	Matter
Biological Change	Earth and Its Place in the Universe
Energy	Food Production and Energy for Life
Cell Structure and Function	Atmospheric Cycles
Structure and Properties of Matter	Earth Features

TN 0708 Gateway Science Reporting Categories	
Cells	Genetics and Biotechnology
Interaction	Diversity
Photosynthesis and Respiration	Biological Change

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 large-scale assessment. Scientifically-based research provided evidence that Discovery Education Predictive Assessments matched TCAP difficulty levels across reporting categories, grades, and subjects.

The Grainger County school system participated in a criterion validity study for the 2006-2007 school year. Approximately 1500 students in grades 3 through 8 completed the Discovery Education Predictive Assessments. For each grade and subject, the students' percent correct averages were broken down by reporting category and compared to their respective 2007 TCAP averages.

Figure 1-18 feature comparisons made between Discovery Education Assessment Test B (Spring 2007) and actual 2007 TCAP reporting category averages. The results featured an average of 264 students per grade level for Reading and Mathematics and an average of 225 students per grade level for Science.

Figure 1: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

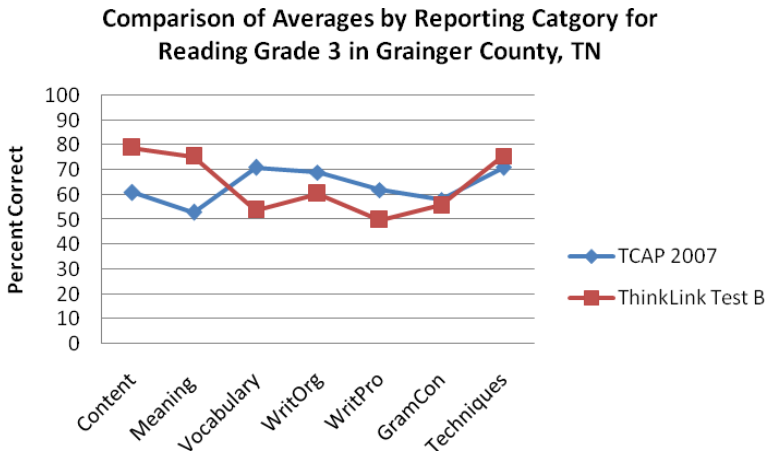


Figure 2: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

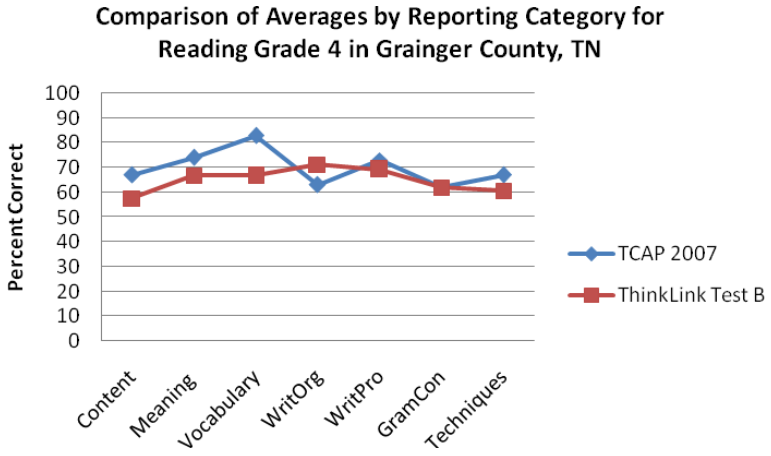


Figure 3: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

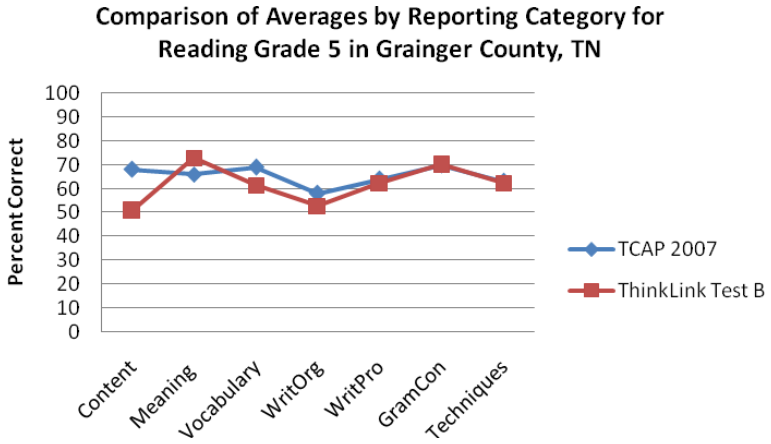


Figure 4: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

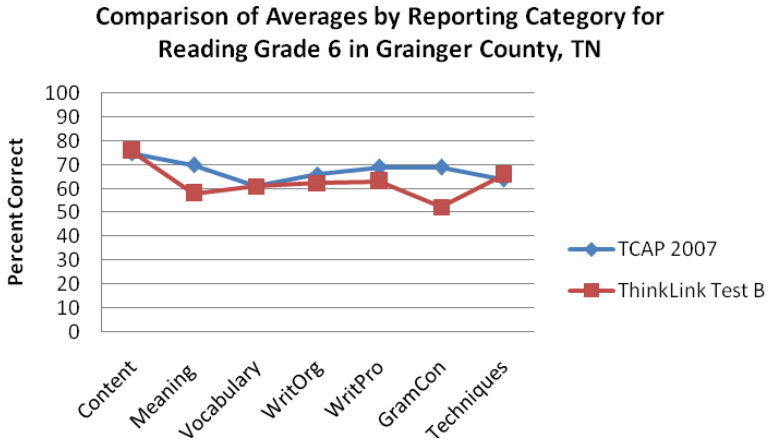


Figure 5: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

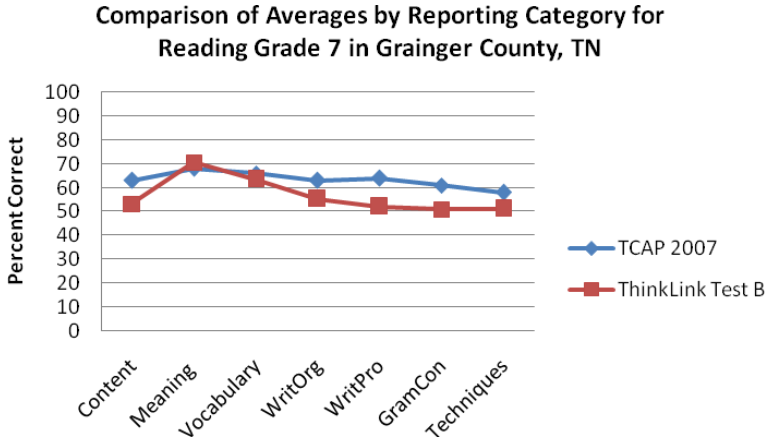


Figure 6: Reading Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

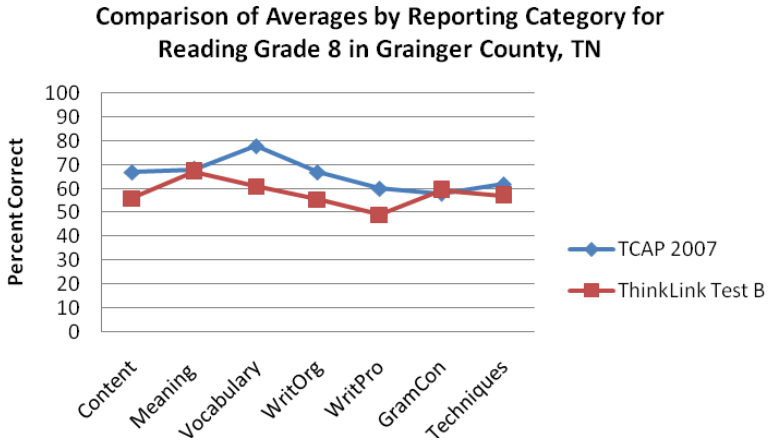


Figure 7: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

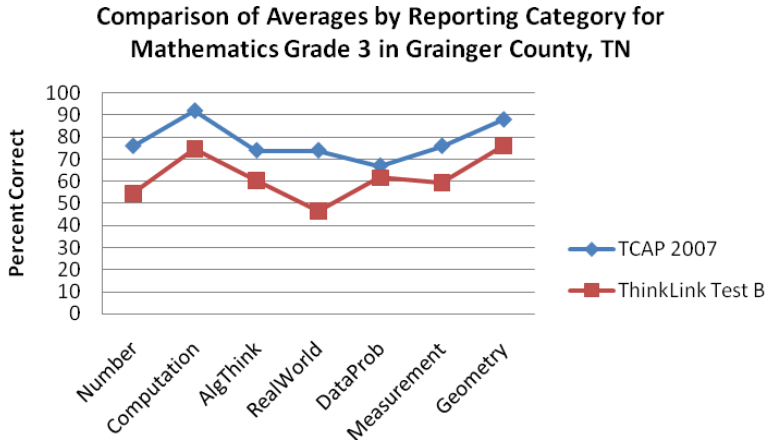


Figure 8: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

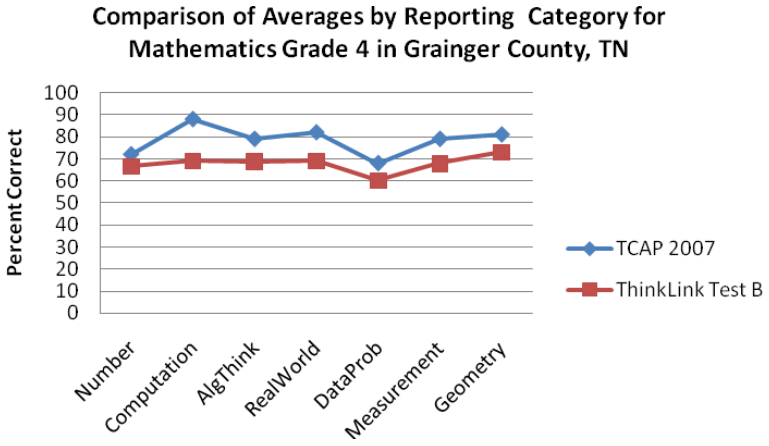


Figure 9: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

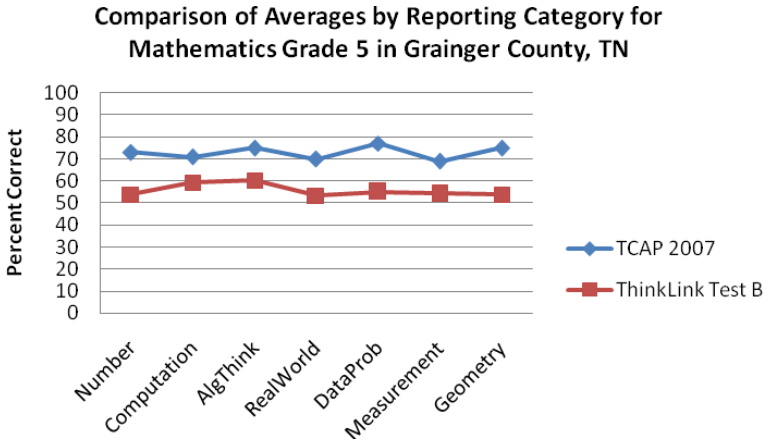


Figure 10: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

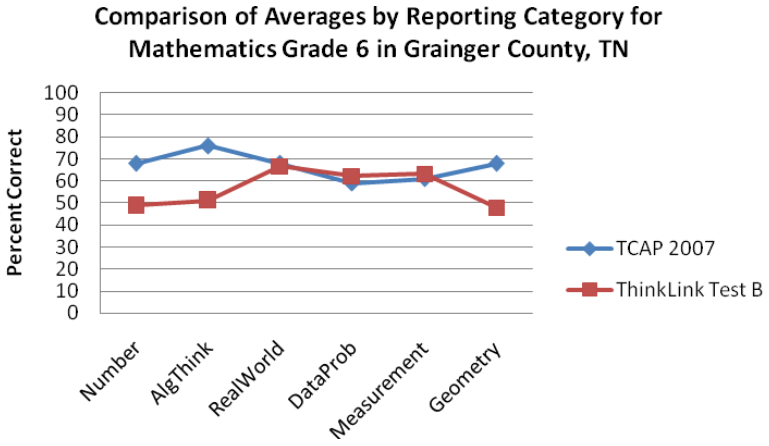


Figure 11: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

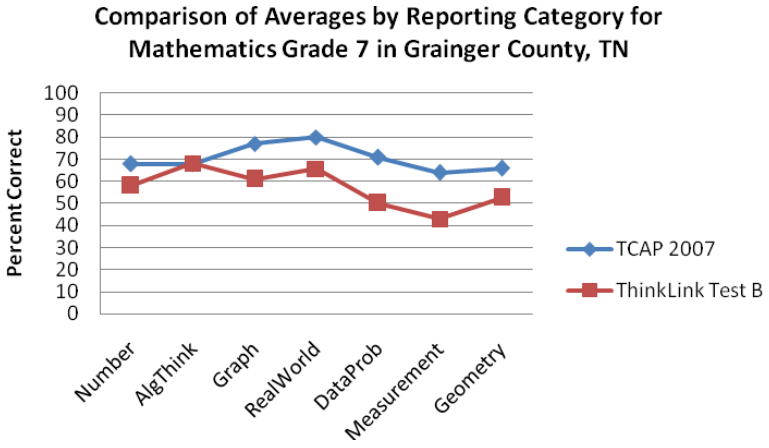


Figure 12: Mathematics Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

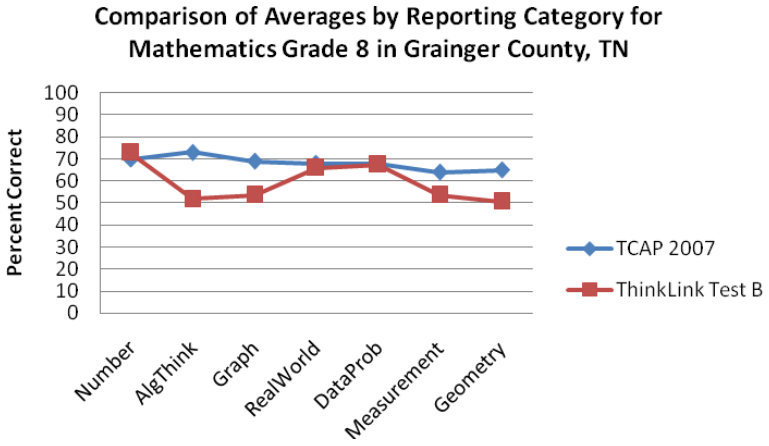


Figure 13: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

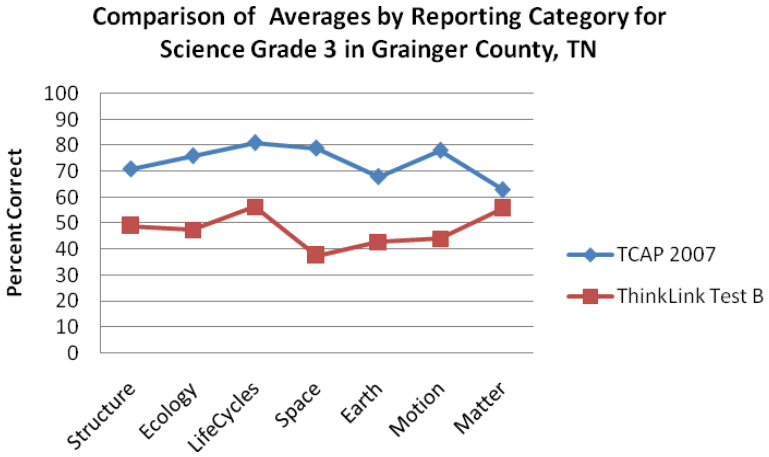


Figure 14: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

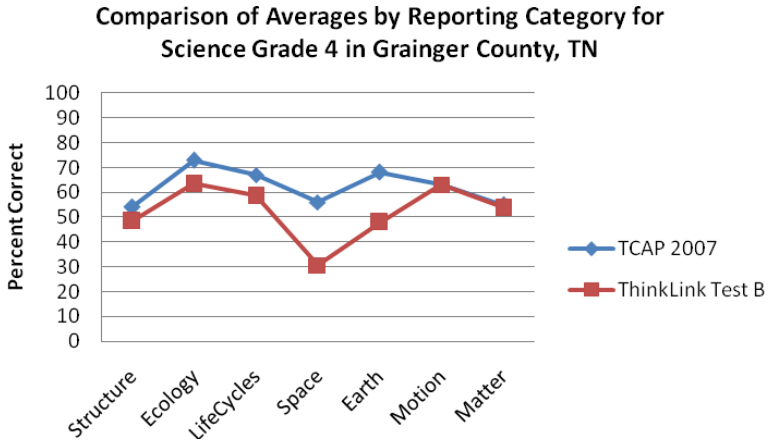


Figure 15: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

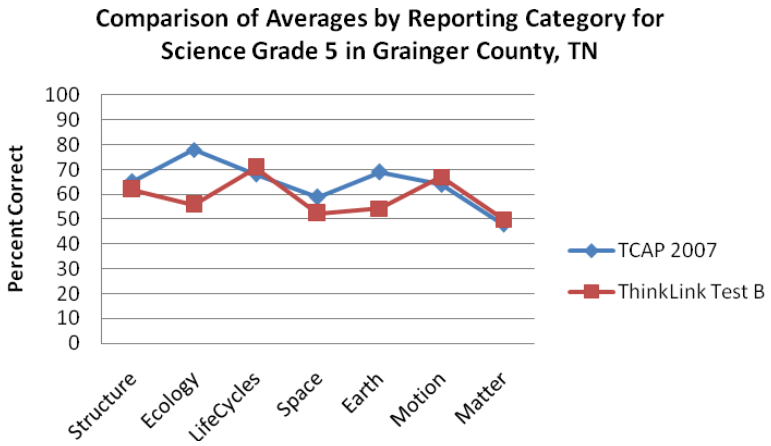


Figure 16: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

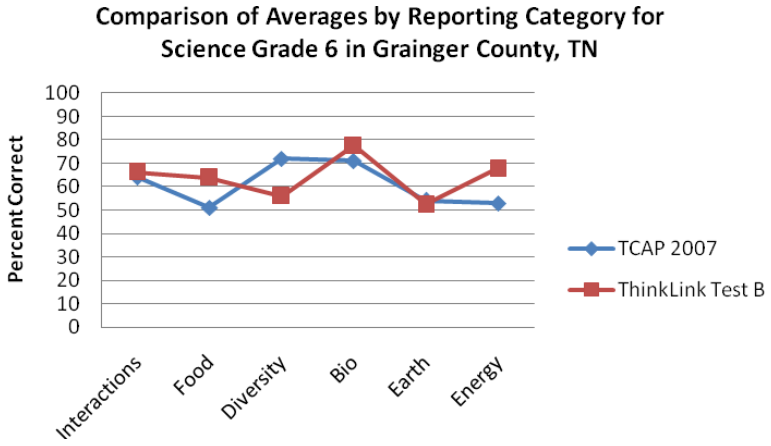


Figure 17: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.

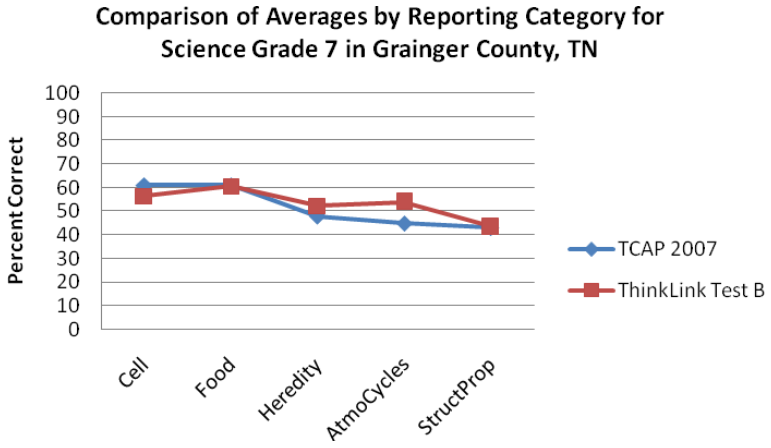
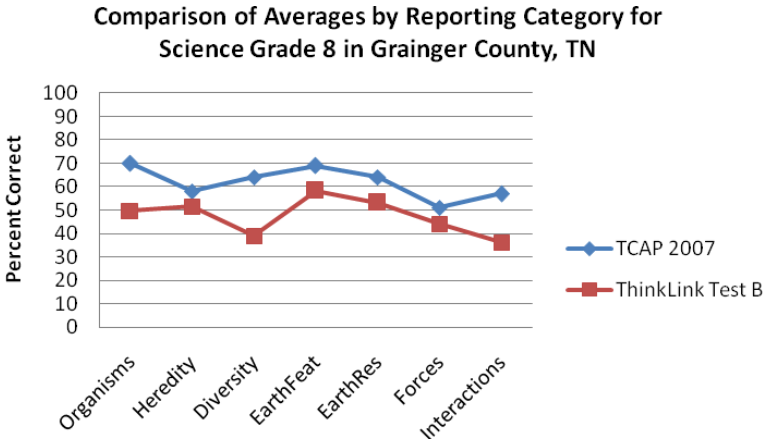


Figure 18: Science Reporting Category Averages for Discovery Education Assessment Test B and 2007 TCAP.



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 Tennessee 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 underpredicted 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.

Discovery Education Assessment Proficiency Predictions vs. TCAP Proficiency Predictions
Due to our representativeness 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 4. The median Proficiency Prediction Score for Reading was 96%, and the median Proficiency Prediction Score for Mathematics was 92%.

Table 4: 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%**

Figure 19 and 20 provide a direct comparison of statewide proficiency percentages for the combined category of "Proficient" and "Advanced" at each grade level. Take a look at Grade 8 Reading. The gap between the TCAP and Discovery Education Assessment proficiency percentages represents the Error Rate (i.e., 14%), which subtracted from 100% yields the Proficiency Prediction Score of 86%.

Figure 19: Comparison of Discovery Education Assessment Test B vs. TCAP Proficiencies in Reading.

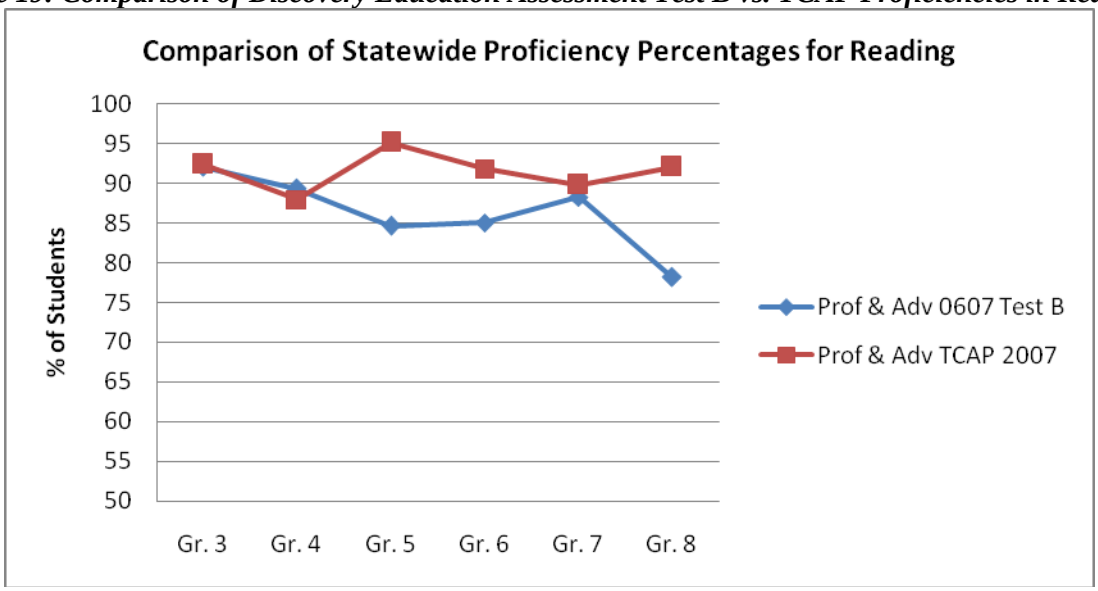
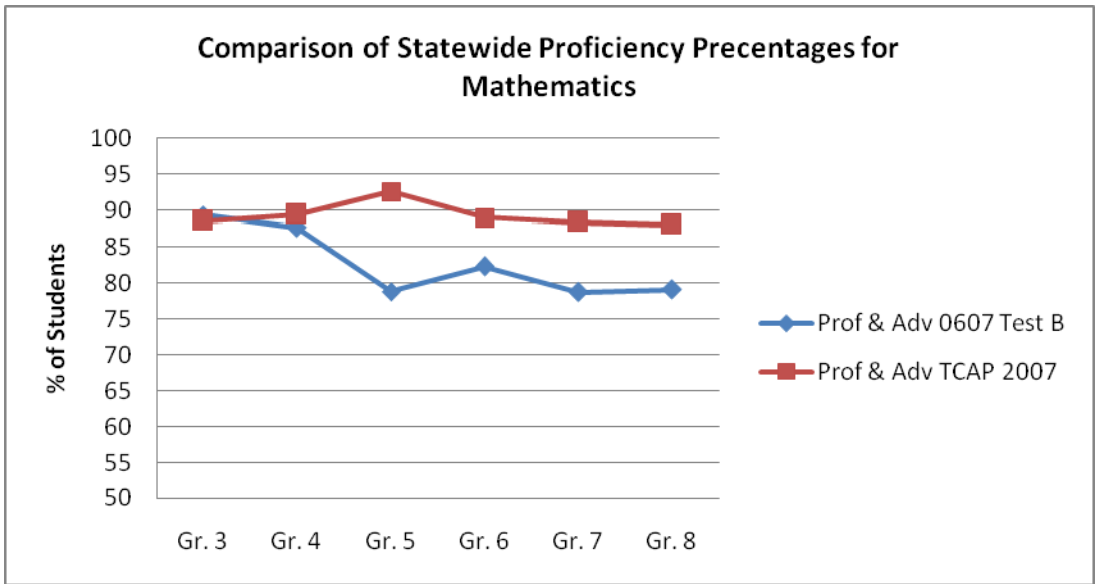


Figure 20: Comparison of Discovery Education Assessment Test B vs. TCAP Proficiencies in Mathematics.



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 Grainger 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" 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 5 and 6.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 5: 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 6: Results of Consequential Validity Study for Grainger County in Reading.

Grainger County, TN Reading				
Grade	**2006**	**2007**	**Difference***	**Grainger ⇔ TN State***
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 Education Predictive 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.

The following figures graphically represent the data of Table 5 and 6. Grainger County had significant improvement (exceeding the state's improvement during the same time period) in Grades 3, 4, 5, 6, and 8 Mathematics and in Grades 3 and 4 Reading (see Figure 21 and 22).

Figure 21: Comparison of Improvement Results for Grainger County and TN State in Mathematics.

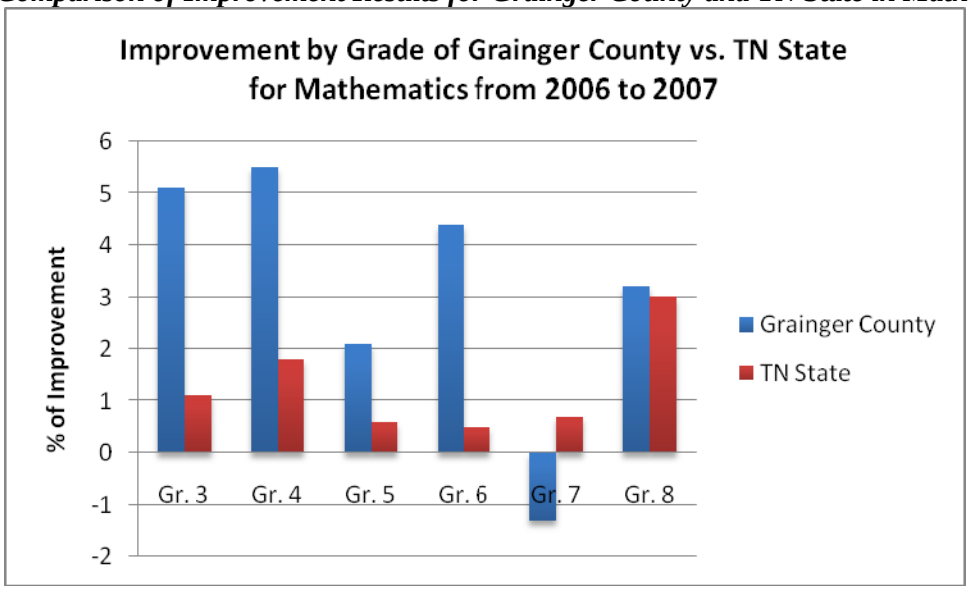
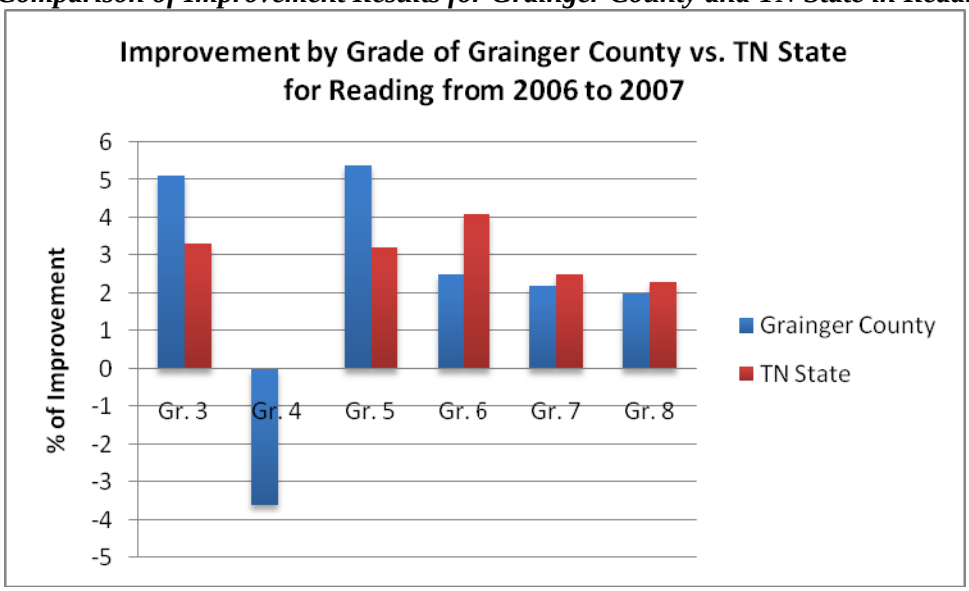


Figure 22: Comparison of Improvement Results for Grainger County and TN State in Reading.



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.

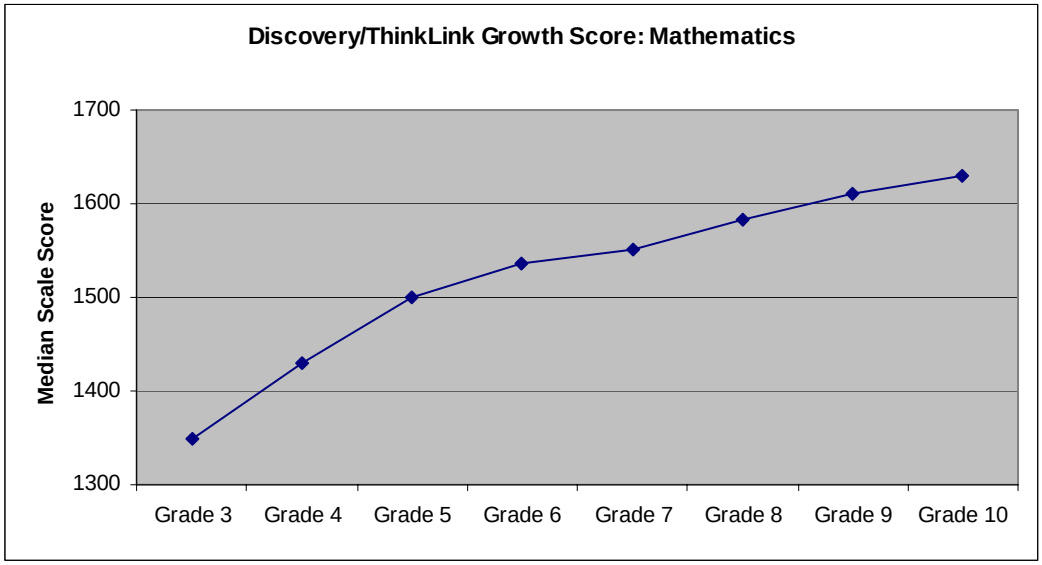
The accurate measurement of student achievement over time is becoming increasingly important to parents, teachers, and school administrators. **Student "growth" within a grade and across grades** has also been sanctioned by the U. S. Department of Education as a reliable way to measure student proficiency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

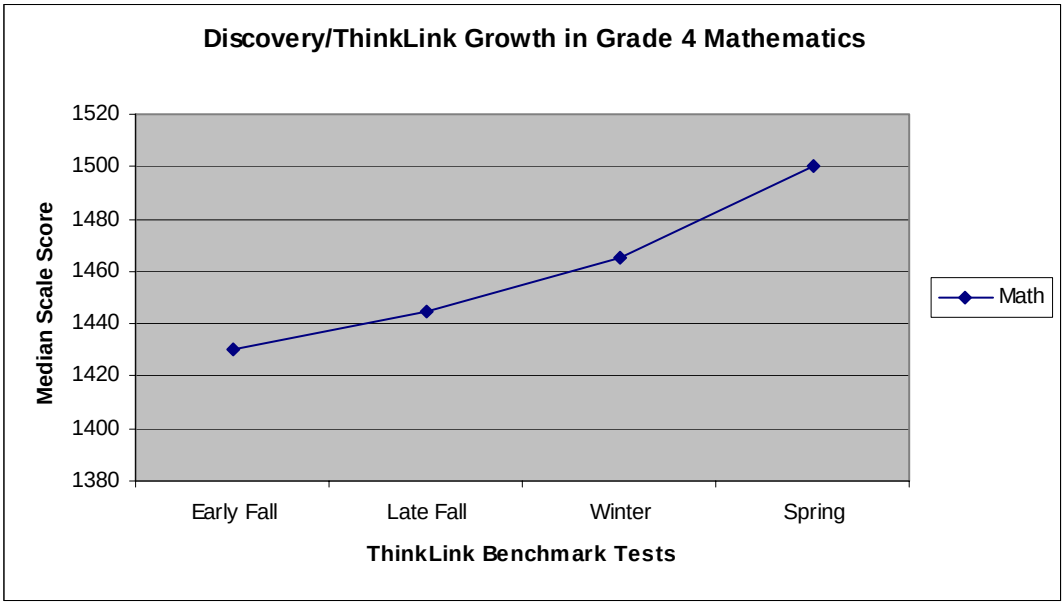
Why use such a rigorous vertical scale?
Isn't student growth similar across grades? Don't students change as much from Grade 3 to Grade 4 as they do from Grade 7 to Grade 8? Previous research on the use of vertical scales has demonstrated that **student growth is not linear**; that is, growth in student achievement is different from grade to grade (see Young 2006). For instance, Figure 23 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 23: 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 24 illustrates this growth for Grade 4 Mathematics using our benchmark assessments.

Figure 24: Within-Year Growth Results for Discovery Education Assessment Mathematics Tests.



Tennessee Growth Scale

The following tables and figures 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 7: 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 8: 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

Figure 25: Test Difficulty Comparison on Vertical Scale for 0708 Reading Tests.

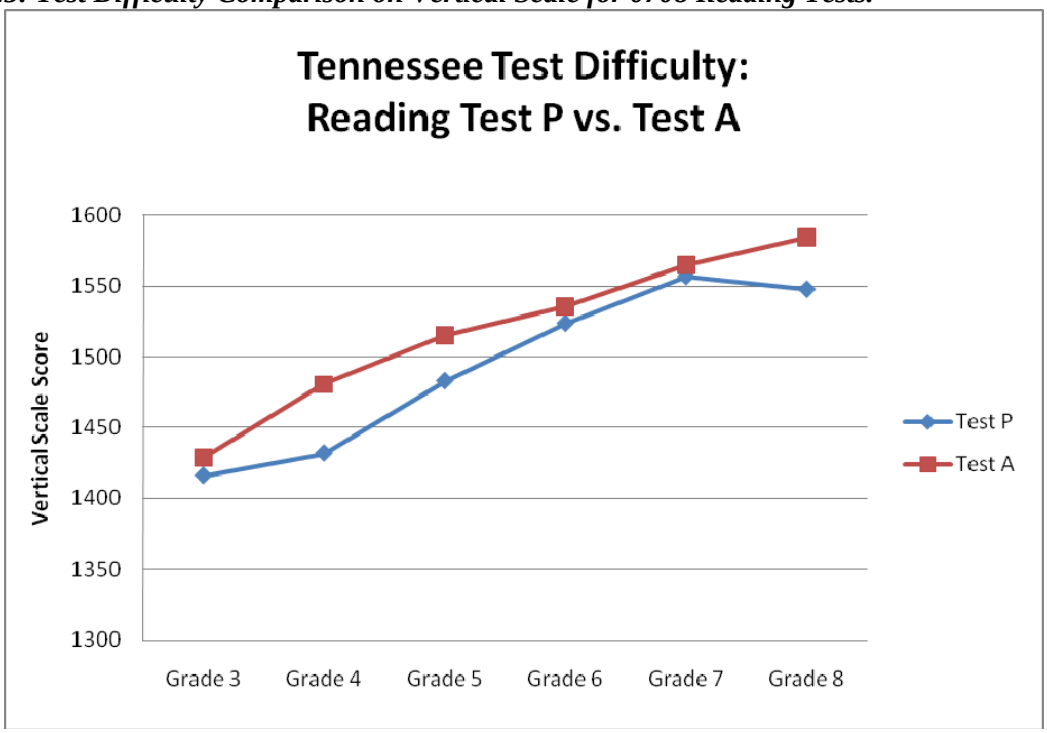
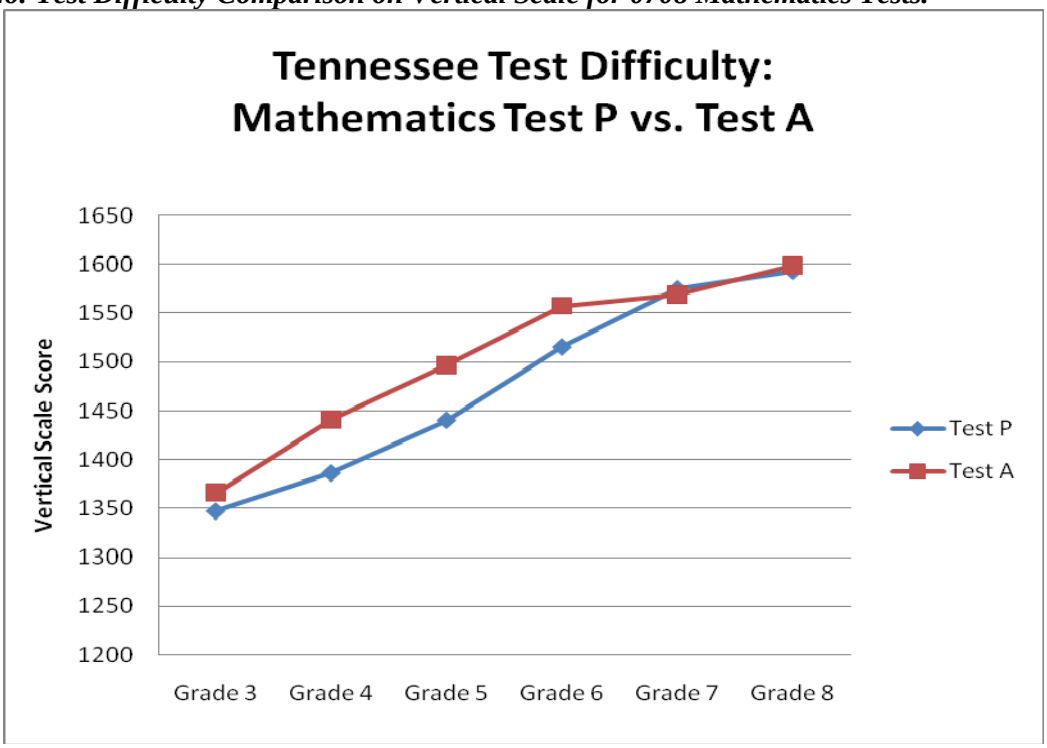


Figure 26: Test Difficulty Comparison on Vertical Scale for 0708 Mathematics Tests.



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:

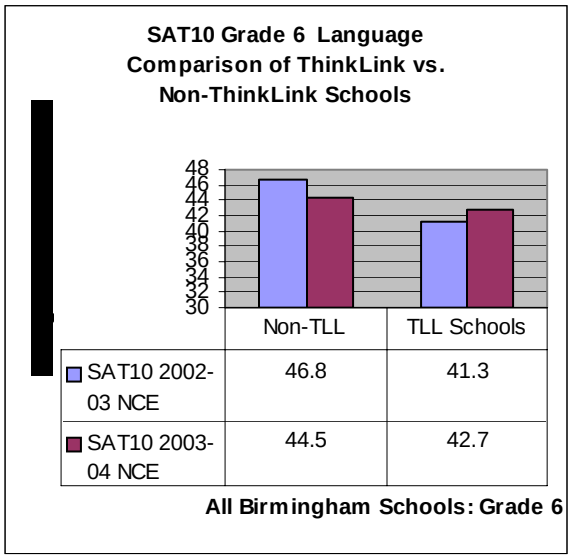
This study is based on data from elementary and middle schools in the City of Birmingham, Alabama. In 2002-03, no Birmingham Schools used Discovery Education's Predictive Assessment Series. Starting in 2003-04, 20 elementary and 9 middle schools used the Discovery Education Assessment program. All Birmingham 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

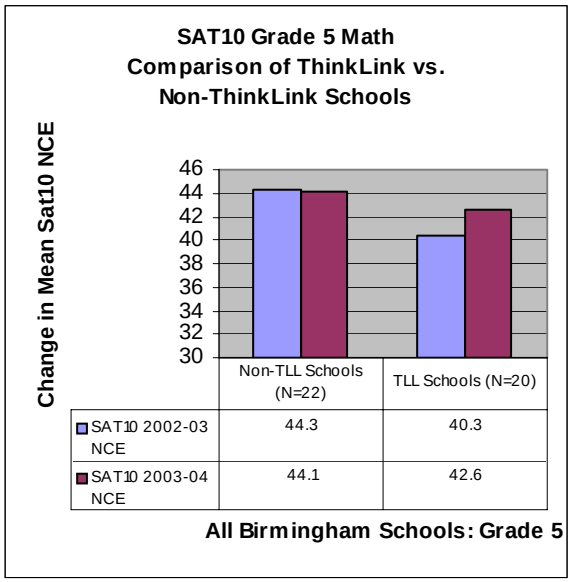
single national percentile is reported for all students within a subject and grade level (this analysis is subsequently referred as ALL STUDENTS). Furthermore, national percentiles are disaggregated by various subgroups within a school. For the comparisons that follow, the national percentiles for students 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.

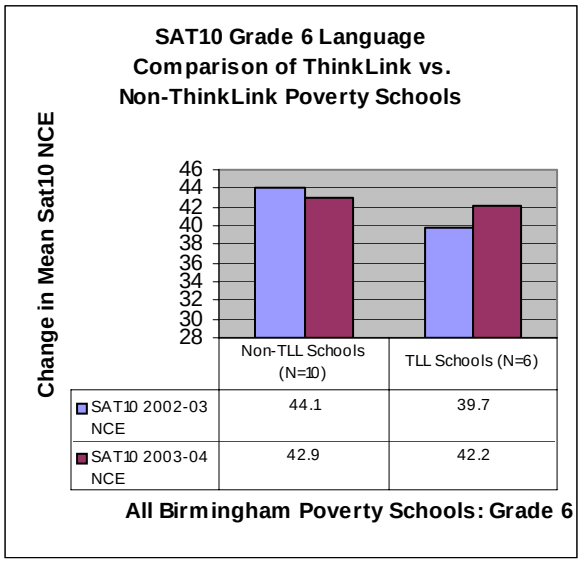
The Discovery Education Assessment schools comprise the experimental group in this study. The Birmingham schools that did not use Discovery Education Assessment comprise the matched comparison group. The following charts show SAT10 National Percentile changes for ThinkLink Schools vs. Non-ThinkLink Schools 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.

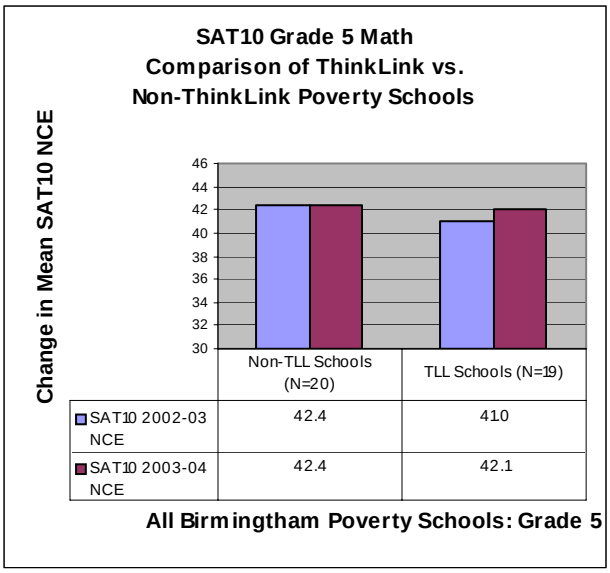
As a result of the improvement that many of these schools made in school year 2003-04, the Birmingham City Schools selected Discovery Education Assessment to be used with *all* of the schools in school year 2004-05. The Birmingham City Schools also chose to provide professional development in each school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

Discovery Education Assessment
RESEARCH



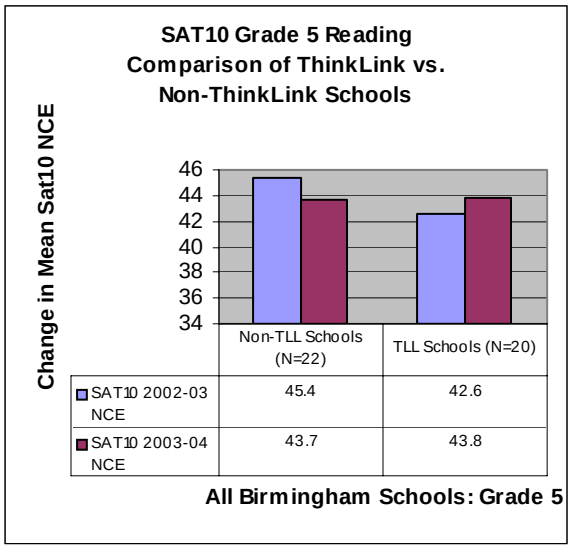


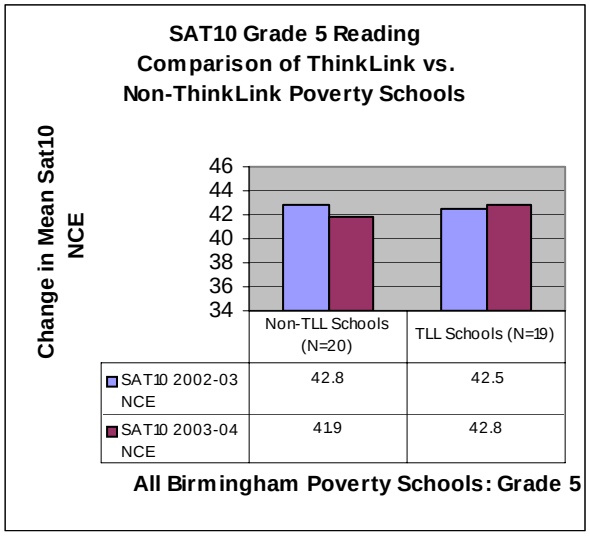




Discovery Education Assessment

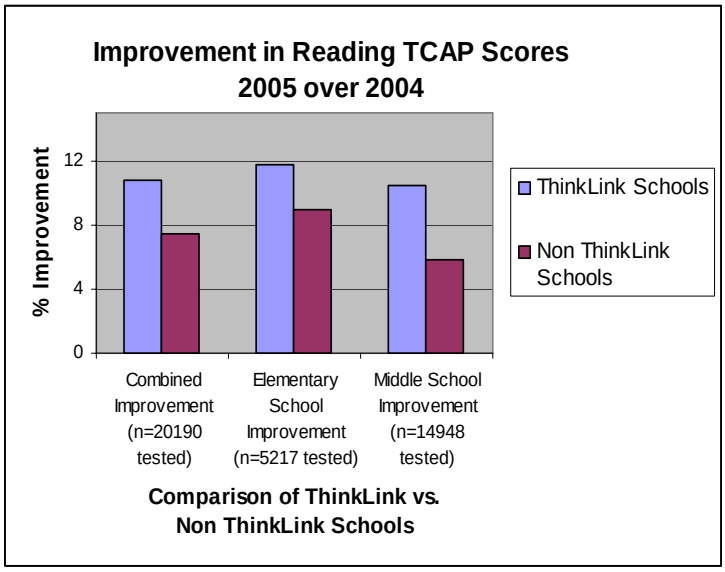
RESEARCH

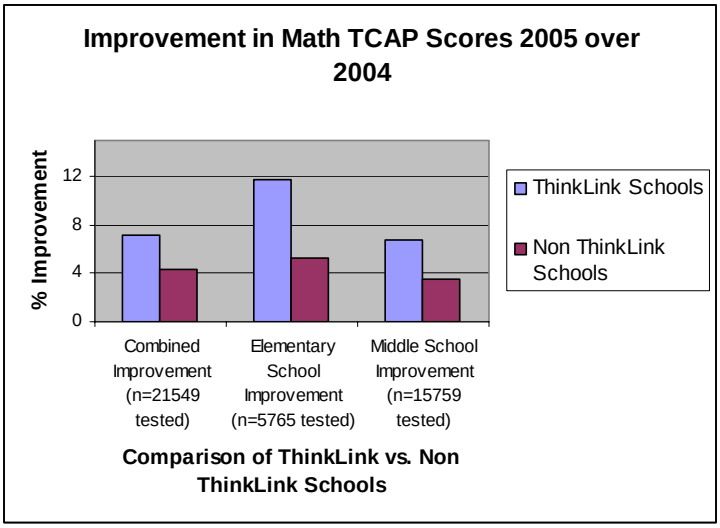




Case Study Two: Metro Nashville, Tennessee City Schools

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 Assessment assessments. Fifty-two elementary and middle schools, representing over 10,000 students, did not use Discovery Education Assessment 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 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 student*. 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.

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

Discovery Education. (2008). *What is predictive assessment: Alabama?* Nashville, TN: Author.

Discovery Education. (2008). *What is predictive assessment: Florida?* Nashville, TN: Author.

Discovery Education. (2008). *What is predictive assessment: Illinois?* Nashville, TN: Author.

Discovery Education. (2008). *What is predictive assessment: Kentucky?* Nashville, TN: Author.

Discovery Education. (2008). *What is predictive assessment: New York?* Nashville, TN: Author.

Discovery Education. (2008). *What is predictive assessment: Tennessee?* Nashville, TN: Author.

5. Tests

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

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