Discovery Education Assessment:
An Overview of the Scientifically-Based Evidence
Supporting the Discovery Education's Benchmark Assessments

Since its inception in 2000 by Vanderbilt University, ThinkLink Learning, now named Discovery Education Assessment, has been determinedly 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 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 benchmark tests are aligned to the content assessed by a given state test allowing teachers to track student progress toward the standards and objectives used for accountability purposes.

The goal of our assessment products is to provide a steady stream of reliable and valid data capable of being integrated into instructional practices leading to an increase in the number of students scoring proficient on a state test. To do so, Discovery Education Assessment has carefully adhered to the criteria for "scientifically-based research" put forth in the *No Child Left Behind Act of 2001*. The information below provides consumers with an overview of the evidence in support of Discovery Education's benchmark assessments as it relates to the six No Child Left Behind (NCLB) criteria for scientifically-based research.

(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;*

Systematic and empirical methods are employed at all levels of the benchmark assessments informing careful item and test development, adequate score reliability, accurate scoring, scaling, equating, and standard setting, as well as reporting and interpretation of test results. In the course of eight years, Discovery Education Assessment has accumulated a substantial amount of scientifically-rigorous evidence in support of its benchmark assessments. Data gathered across numerous school districts in over a dozen states were subjected to **psychometric analyses** that

have repeatedly demonstrated **high test reliability** and supported the **validity of test score interpretations** (i.e., proficiency predictions).

Reliability Evidence. The following tables feature the **reliability coefficients** for several states in *Reading and Language* and *Mathematics*. High test reliability indicates that test items are consistently measuring the same concept. Reliabilities of Discovery Education Assessment tests are calculated using Cronbach's alpha.

	2005 Reading and Language			
	Tennessee	***Kentucky***	***Florida***	***Alabama***
Grade 2	.91	.91	.89	.90
Grade 3	.87	.89	.91	.91
Grade 4	.86	.90	.85	.86
Grade 5	.85	.86	.82	.87
Grade 6	.86	.86	.85	.85
Grade 7	.86	.86	.83	.87
Grade 8	.86	.86	.86	.84
Grade 9		.85		
Grade 10		.84		

	2005 Mathematics			
	Tennessee	***Kentucky***	***Florida***	***Alabama***
Grade 2	.88	.90	.87	.88
Grade 3	.86	.89	.88	.88
Grade 4	.85	.90	.86	.88
Grade 5	.88	.89	.87	.88
Grade 6	.85	.88	.88	.87
Grade 7	.85	.85	.86	.85
Grade 8	.85	.89	.87	.82
Grade 9		.89		
Grade 10		.87		

	2006 Reading and Language		***2006 Mathematics***	
	Illinois		***Illinois***	
	Reliability	Sample Size	Reliability	Sample Size
Grade 3	.85	274	.89	258
Grade 4	.84	273	.89	294
Grade 5	.75	310	.87	332
Grade 6	.77	377	.88	364
Grade 7	.81	261	.88	355
Grade 8	.84	402	.84	385

Validity Evidence. To ensure **content validity** of all benchmarks 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 PAS 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 benchmark tests are thus **state specific** and feature **matching reporting categories** of a given state's large-scale assessment used for accountability purposes.

Discovery Education benchmark assessments have been or are being developed for the following states: Tennessee, Alabama, North Carolina, Kentucky, Wisconsin, Florida, Illinois, New York, West Virginia, Mississippi, South Carolina, California, Ohio, Virginia, Missouri, Oklahoma, and New Mexico. The following are several examples of the objectives and reporting categories that appear on the Discovery Education Assessment Student Reports including those of the nationally normed Stanford 10 test:

Tennessee Comprehensive Assessment Program (TCAP)

Reading and Language Reporting Categories
Content, Meaning, Vocabulary, Writing: Organization, Writing: Process, Grammar Conventions, Techniques and Skills

Mathematics Reporting Categories
Number Sense and Operations, Computation, Algebraic Thinking, Real World Problem Solving, Data Analysis and Probability, Measurement, Geometry

Alabama Reading and Math Test (ARMT)

Reading and Language Reporting Categories
Apply advanced phonetic analysis, Demonstrate reading vocabulary knowledge, Use strategies to comprehend literary/recreational materials, Use strategies to comprehend functional and textual/informational materials, Writing Process, Grammar, Study Strategies

Mathematics Reporting Categories
Number and Operations, Algebra, Data Analysis and Probability, Measurement, Geometry

Florida Comprehensive Assessment Test (FCAT)

Reading and Language Reporting Categories
Words, Main Idea, Comparison/Cause and Effect, Reference and Research

Mathematics Reporting Categories
Number and Operations, Algebra, Data Analysis and Probability, Measurement, Geometry

Illinois Standards Achievement Test (ISAT)

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

Mathematics Reporting Categories
Number, Measurement, Algebra, Geometry, Data Analysis and Probability

Stanford Achievement Test (Tenth Edition)

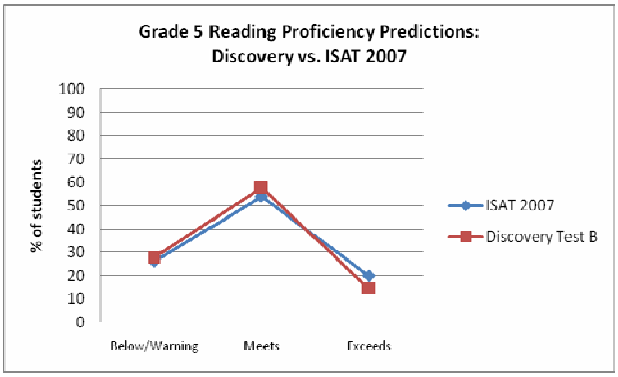
Reading and Language Objectives
Initial Understanding, Interpretation, Analysis, Reading Strategies, Capitalization, Usage, Punctuation, Sentence Structure, Organization

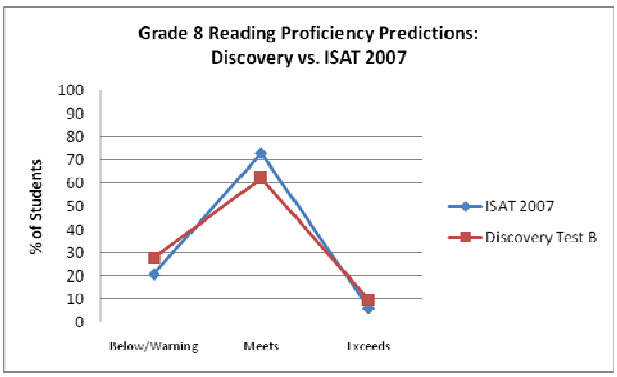
Mathematics Objectives
Number Sense, Geometry and Measurement, Data Analysis, Statistics, and Probability, Algebra, Communications and Representation, Estimation, Connections, Problem Solving

Since content alignment represents a necessary but not sufficient condition for the validity of test score interpretations, consumers are likely to ask for further evidence in the form of **valid proficiency and performance predictions**. Scientifically-based research requires a vendor to publish the percentage of students that attain the various state proficiency levels on its tests compared to the students' percentage of proficiency on the high stakes test of the respective state. Discovery Education Assessment has clear evidence that students' proficiency levels via the benchmark tests consistently match those achieved on the actual high stakes test. The graphs below illustrate several states examples for different grades and subject areas.

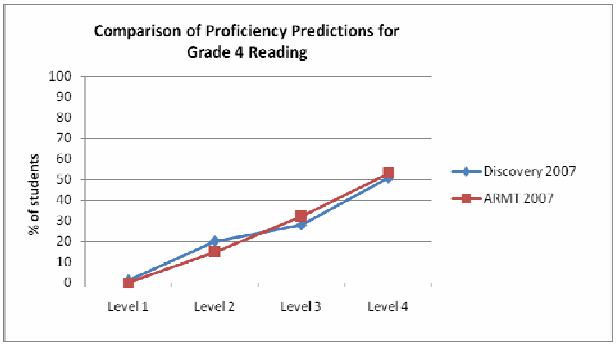
Discovery Education Assessment
RESEARCH

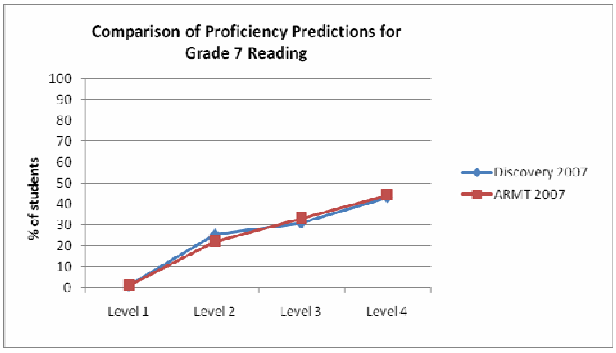
Illinois



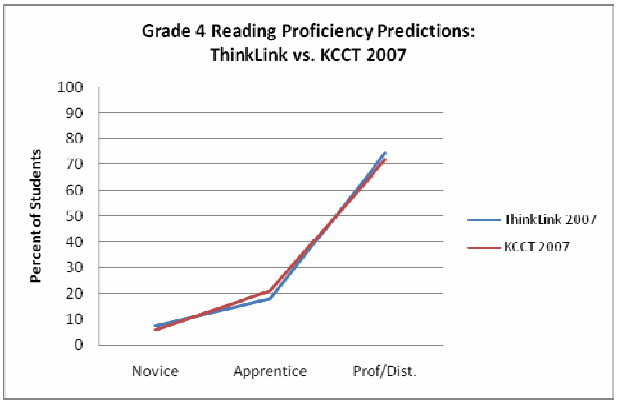


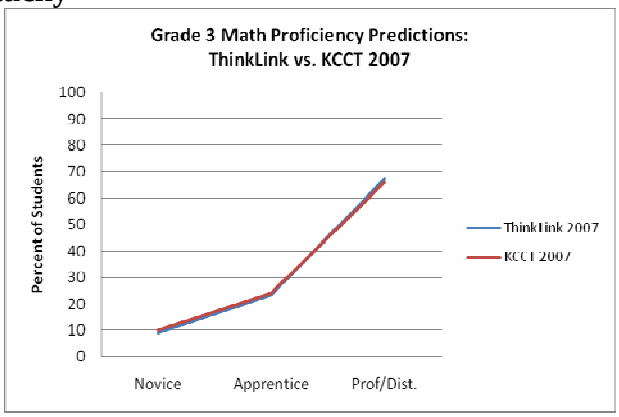
Alabama



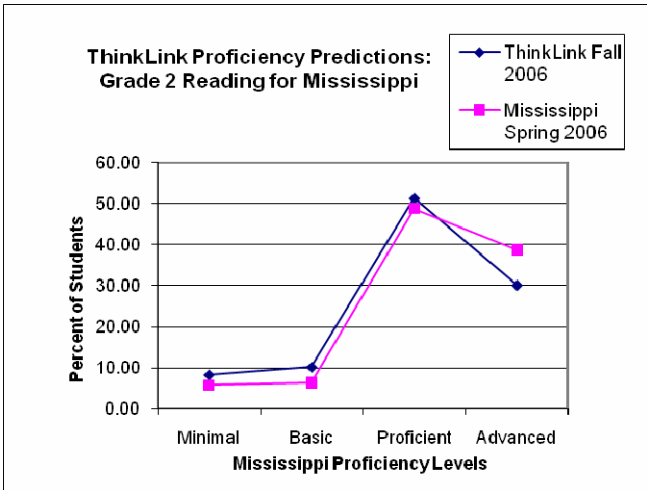


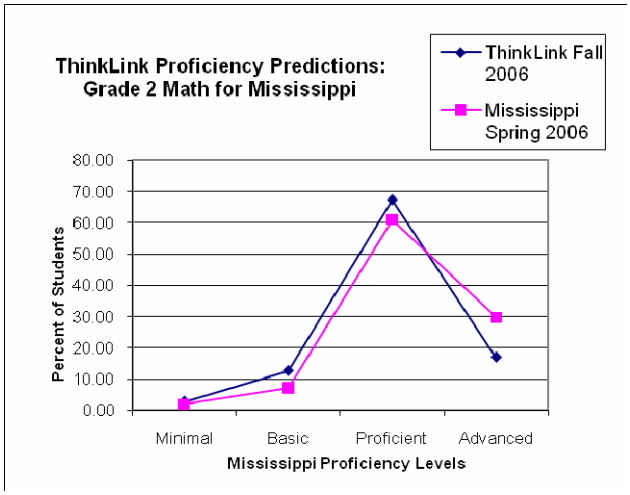
Kentucky



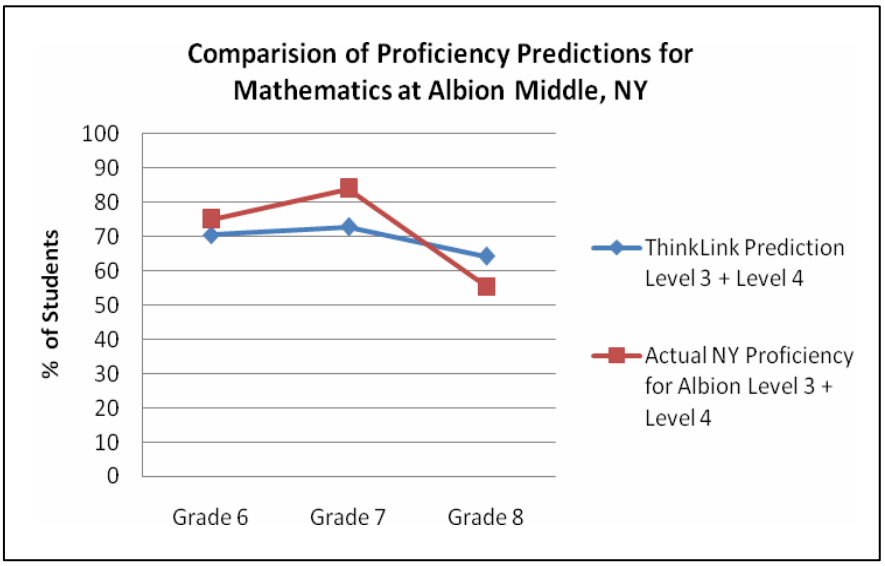


Mississippi



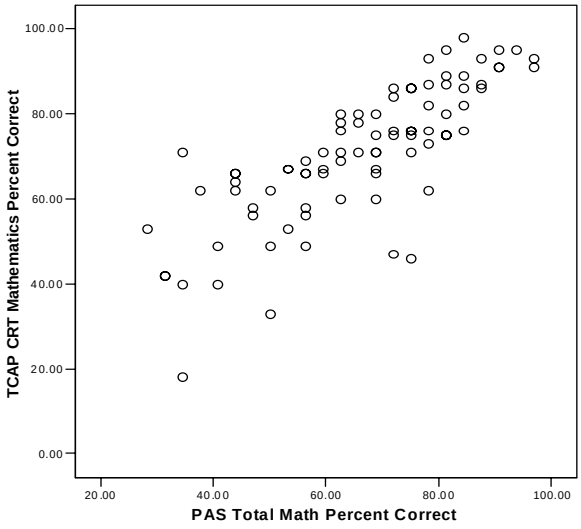


New York

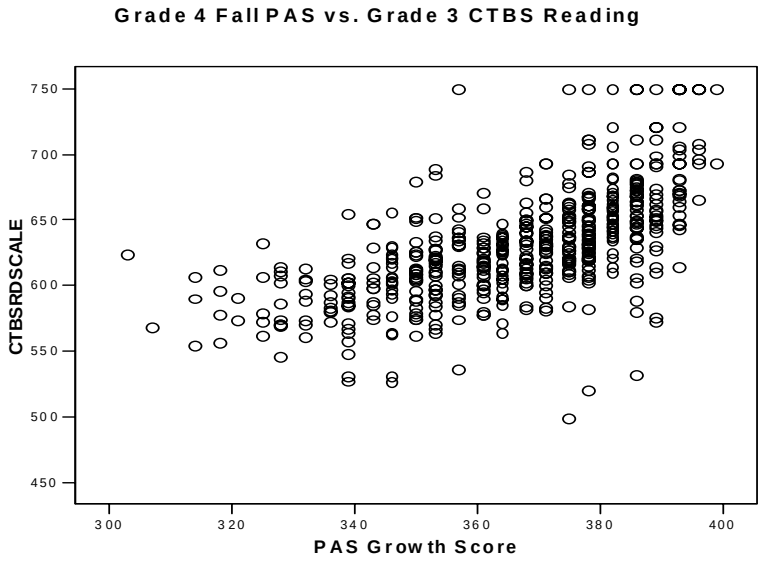


Scientifically-based research further presents evidence that there is a **significant correlation between student performance on a vendor's test and the state's test**. The following examples feature scatter plots that correlate students' percent correct on the benchmark assessment to percent correct on the high stakes test.

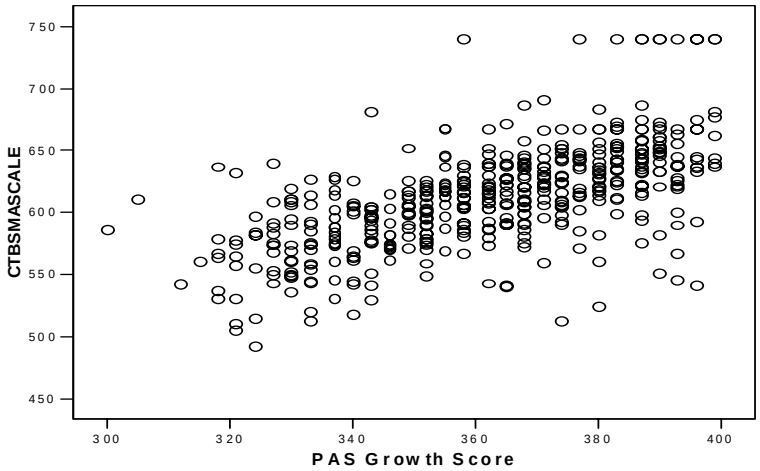
Discovery Education Assessment 4th Grade Mathematics tests predicted scores on the Tennessee CRT for students at Trenton Elementary School. The following figure shows a scatter plot based on 87 students from a Grade 4 Math class at Trenton Elementary School. The correlation between the benchmark assessment Mathematics Percent Correct and the actual percent correct on the Spring 2004 Tennessee CRT was .79 ($p < .01$).



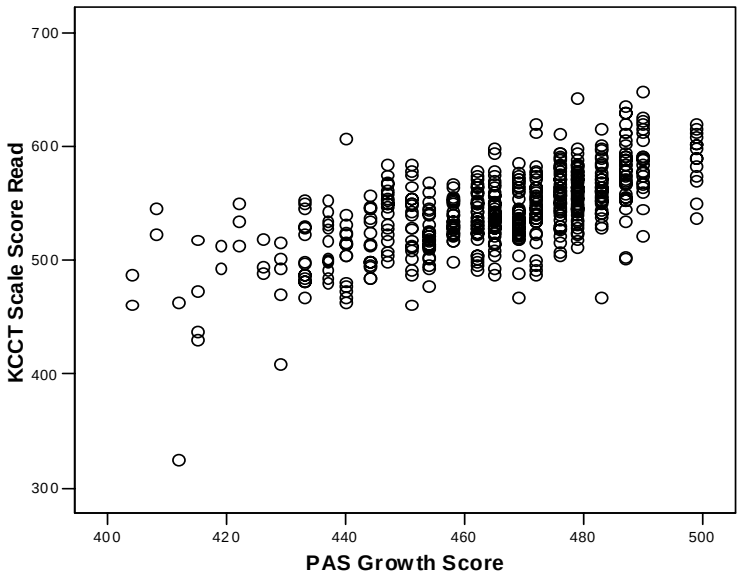
Discovery Education benchmark assessments predicted scores on the CTBS and the KCCT for Jefferson County Public Schools. The following studies were based on data from the Jefferson County Public School System. This system used the Discovery Education Assessment during the 2004-2005 school year. Results on the benchmark assessment from the Fall 2004 were compared with KCCT data from Spring 2004. In the first graph, Grade 4 Reading Growth Scores from the benchmark assessment were correlated with Grade 3 Reading CTBS scores from a sample of 563 students. The correlation was highly significant ($r = .63$, $p < .01$). In the second graph, Grade 4 Mathematics Growth Scores from the benchmark assessment were correlated with Grade 3 Mathematics CTBS scores from a sample of 492 students. The correlation was highly significant ($r = .63$, $p < .01$). In the third graph, Grade 5 Reading Growth Scores from the benchmark assessment were correlated with Grade 4 Reading KCCT scale scores from a sample of 567 students. The correlation was highly significant ($r = .65$, $p < .01$). The conclusion from these studies is that Discovery Education Assessment is a highly significant predictor of scores on both the Terra Nova/CTBS and the KCCT.



Grade 4 Fall PAS vs. Grade 3 CTBS Mathematics



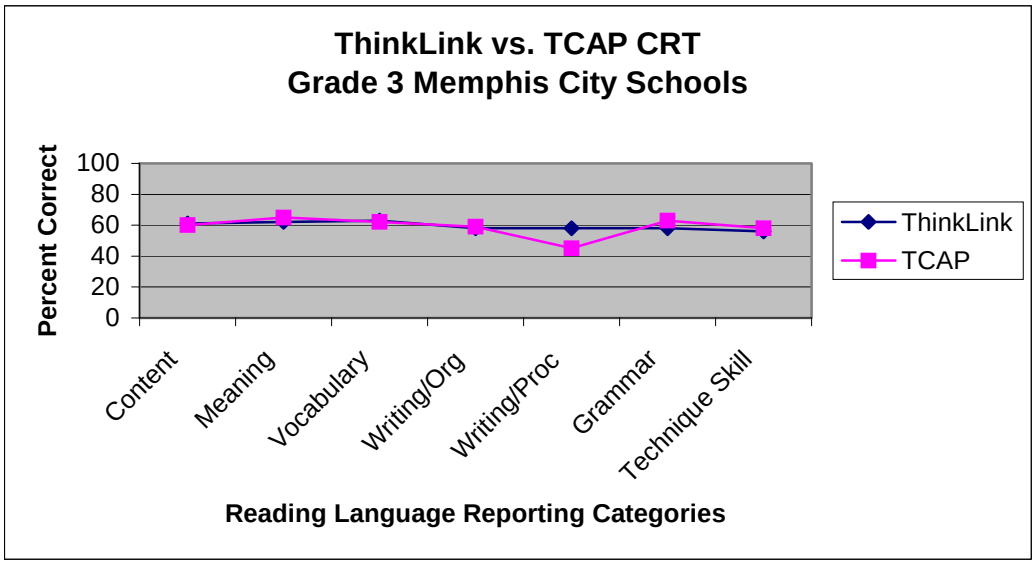
Grade 5 Fall PAS vs. Grade 4 KCCT Reading

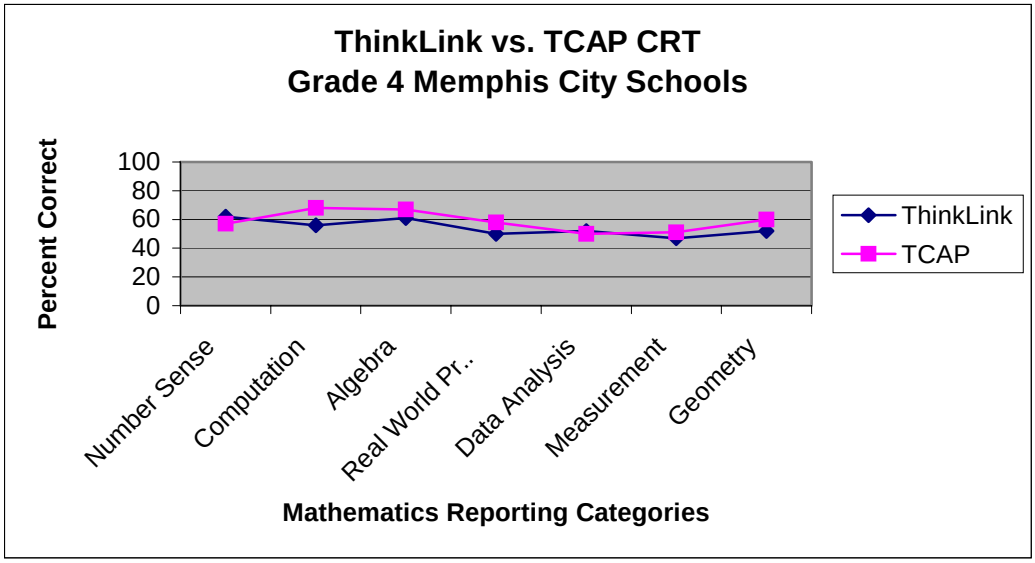


Discovery Education Assessment
RESEARCH

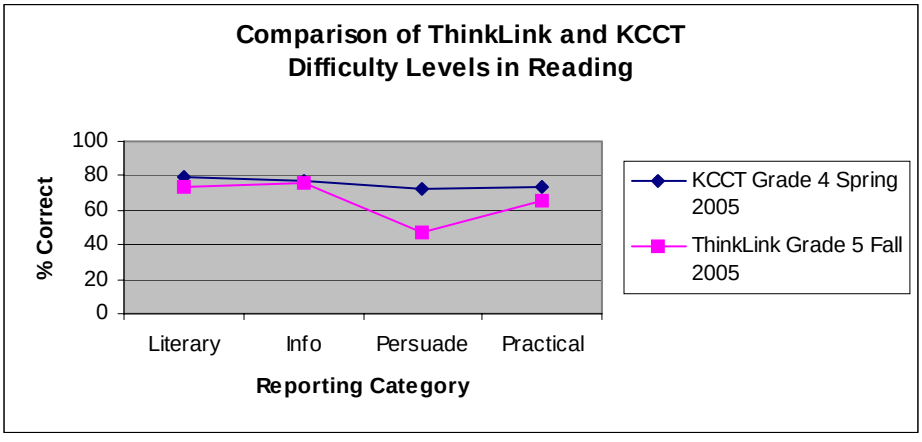
Discovery Education Assessment has further accumulated **validity evidence at the skill level**. That is, averages should be available not only for the entire test but also for specific skills within the test. For instance, in a math test, the overall average is useful, but the average within a skill**,** such as number sense, geometry, or measurement, is even more useful to teachers. These skill averages should be available in every subject area: data must be available on each skill to guide teachers and students on where improvement can occur. Furthermore, test and skill averages should match the average difficulty on state specific tests. In other words, scientifically-based research shows similar difficulty levels *by skill and by grade* when student results of the vendor's test are compared to those on the state's actual test. The following skill-by-skill examples show how closely Discovery Education's benchmark assessment **percent correct for each skill match the respective percent correct on the state test**.

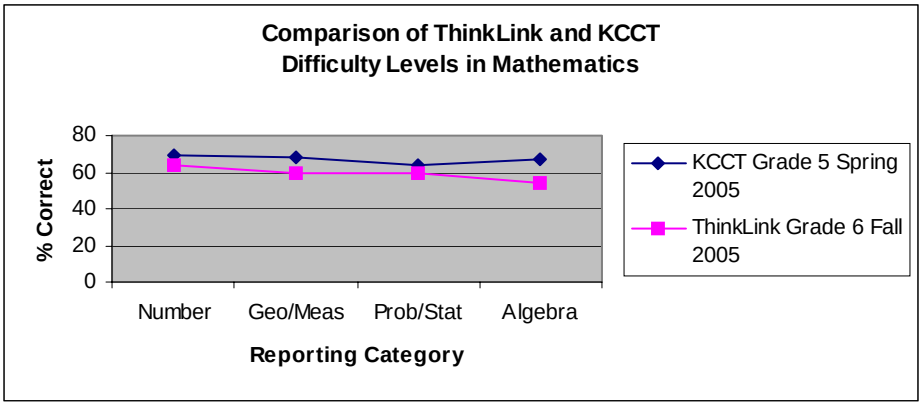
Discovery Education benchmark assessments matched difficulty levels on the Tennessee CRT for Memphis City Schools. Here are two examples from the benchmark assessments that demonstrate that our reporting category averages match those from the Tennessee Achievement CRT. Both examples are based on over 3000 students from Memphis City Schools. The first graph shows a comparison for Grade 3 Reading and Language, and the second graph shows a comparison for Grade 4 Mathematics.

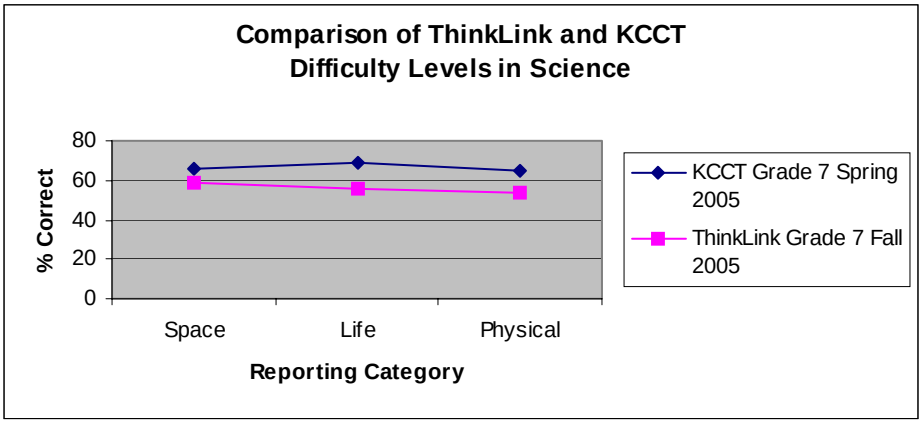


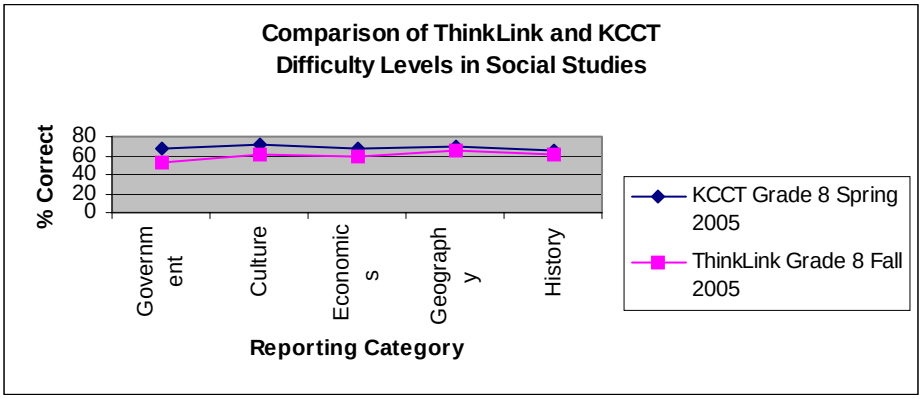


Discovery Education benchmark assessments matched the difficulty levels on KCCT Reading, Mathematics, Science, and Social Studies Tests. The following graphs show how Discovery Education Assessment averages for reporting categories for Reading, Mathematics, Science, and Social Studies match KCCT averages. The first graph compares the benchmark Fall Grade 5 test (equivalent to an end of year Grade 4 test) with the KCCT Spring Grade 4 Reading test from 2005. The second graph compares the benchmark Fall Grade 6 test (equivalent to an end of year Grade 5 test) with the KCCT Spring Grade 5 Mathematics test from 2005. The third graph compares the benchmark Fall Grade 7 Science Test with the KCCT Spring Grade 7 2005 Science Test. The fourth and final graph compares the Grade 8 benchmark Fall Social Studies test with the KCCT Spring Grade 8 Social Studies Test. For Science and Social Studies, benchmark Fall tests are considered end-of-year tests.




Comparison of ThinkLink and KCCT
Difficulty Levels in Mathematics
% Correct
80
60
40
20
0
Number
Geo/Meas
Prob/Stat
Algebra
Reporting Category
KCCT Grade 5 Spring 2005
ThinkLink Grade 6 Fall 2005

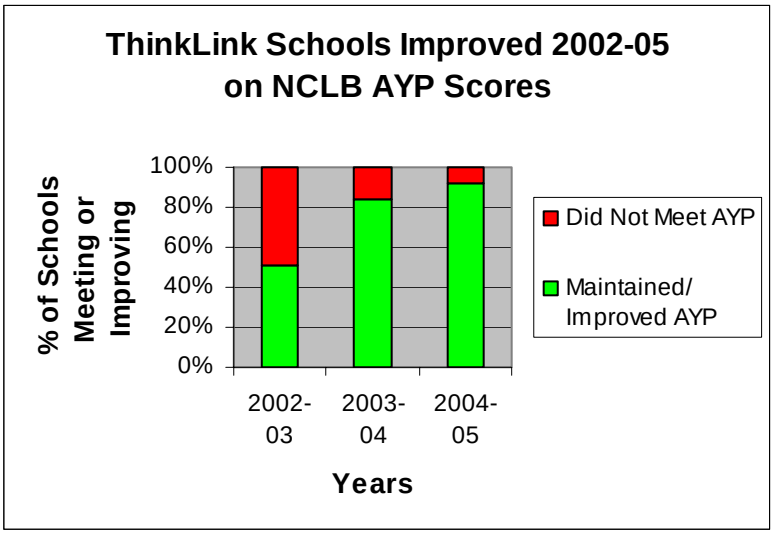

Comparison of ThinkLink and KCCT
Difficulty Levels in Science
% Correct
80
60
40
20
0
Space
Life
Physical
Reporting Category
KCCT Grade 7 Spring 2005
ThinkLink Grade 7 Fall 2005


Comparison of ThinkLink and KCCT
Difficulty Levels in Social Studies
% Correct
80
60
40
20
0
Governm ent
Culture
Economic s
Geograph y
History
Reporting Category
KCCT Grade 8 Spring 2005
ThinkLink Grade 8 Fall 2005

Lastly, Discovery Education Assessment has also accumulated evidence demonstrating the **positive impact of using Discovery Education benchmark assessments on student performance**. A vendor should be able to publish research studies that demonstrate that its assessments help improve student performance and help schools meet NCLB requirements for adequate yearly progress (AYP). Examples of scientific research should include large numbers of students and schools and not simply a single school or district that has achieved test score gains. Below are two examples that underscore the efficacy of PAS to improve student achievement.

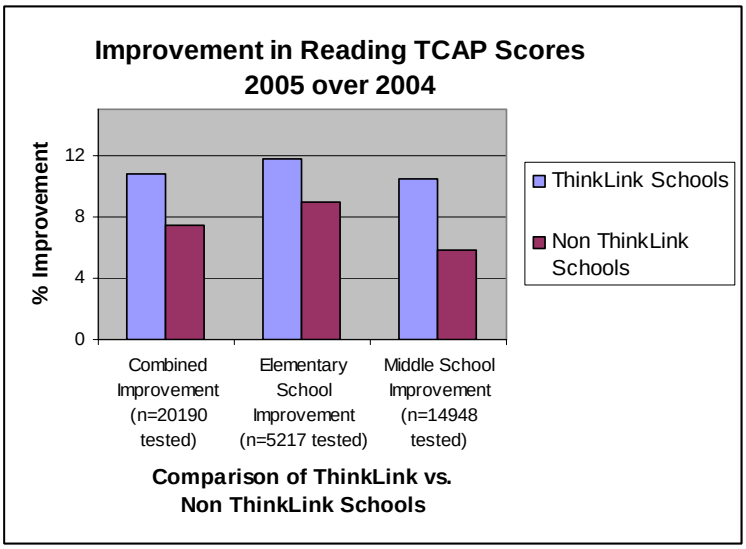
Discovery Education Assessment schools in the State of Tennessee demonstrated significant AYP Improvement from 2002 to 2005. The following results describe improvement in AYP status for Discovery Education Assessment schools from 2002 to 2005. During 2003-2004, three hundred and ninety-nine Tennessee schools used Discovery Education benchmark assessments. These schools represent over 100,000 students in grades 3 to 8. During 2002-2003, these same schools were designated as either "meeting" or "not meeting" AYP requirements for NCLB. For this school year, 51.38% of schools met AYP and 48.62% did not meet AYP. By the end of the 2003-2004 school year, there had been dramatic improvement in AYP status for Discovery Education Assessment schools. By 2004, 84.46% of Discovery Education Assessment schools had met or improved their AYP status and only 15.54% did not meet AYP.

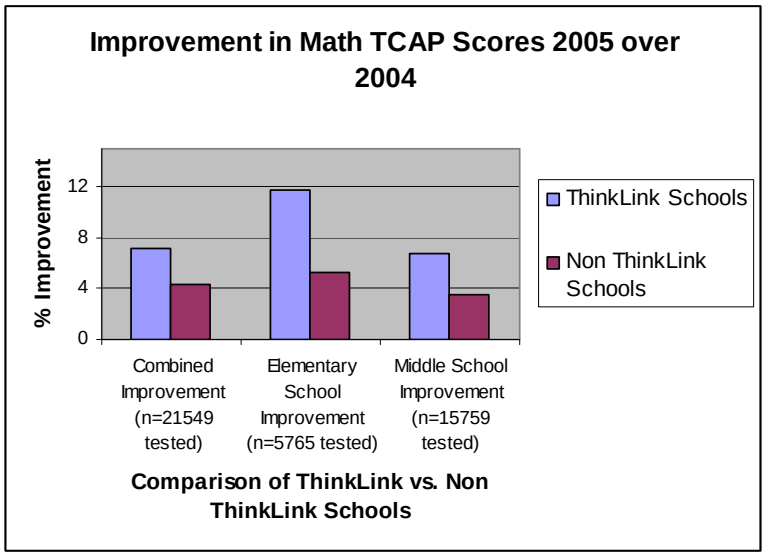
This progress continued through the next school year. During 2004-2005, four hundred and seventy seven schools used Discovery Education benchmark assessments. These schools represent over 150,000 students in grades 3 to 8. By 2005, 92.03% of Discovery Education Assessment schools had met or improved their AYP status and only 7.97% did not meet AYP. This three-year improvement is presented graphically below. These results demonstrate remarkable progress given that most schools that use Discovery Education Assessment are Title 1.



Discovery Education Assessment
RESEARCH

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 Discovery Education Assessment schools versus non-Discovery Education Assessment 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.

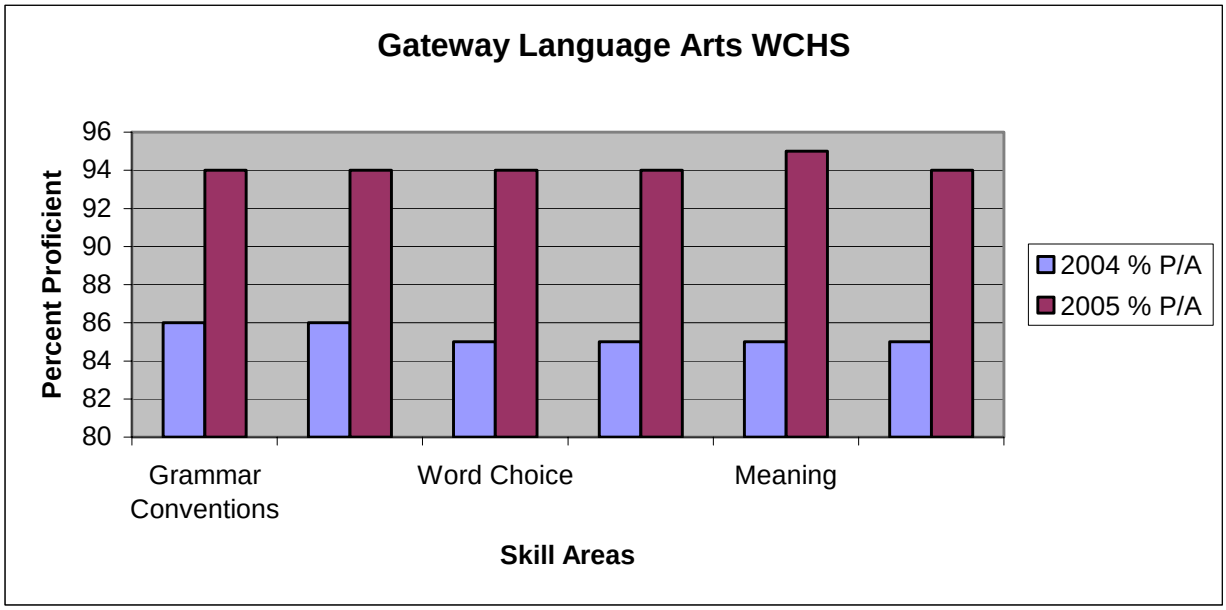




Discovery Education Assessment

RESEARCH

Warren County High School demonstrated significant improvement on its Gateway English II high school exit examination from 2004 to 2005. Students at Warren County High School used Discovery Education Assessment's Gateway practice examinations during the 2004-2005 school year. The percent of students at the Proficient/Advanced level on the English II Gateway improved from 71% in 2004 to 95% in 2005. Significant improvement in the percent of students at the Proficient/Advanced levels was demonstrated in every single reporting category on the English II Gateway as displayed in the following graph.

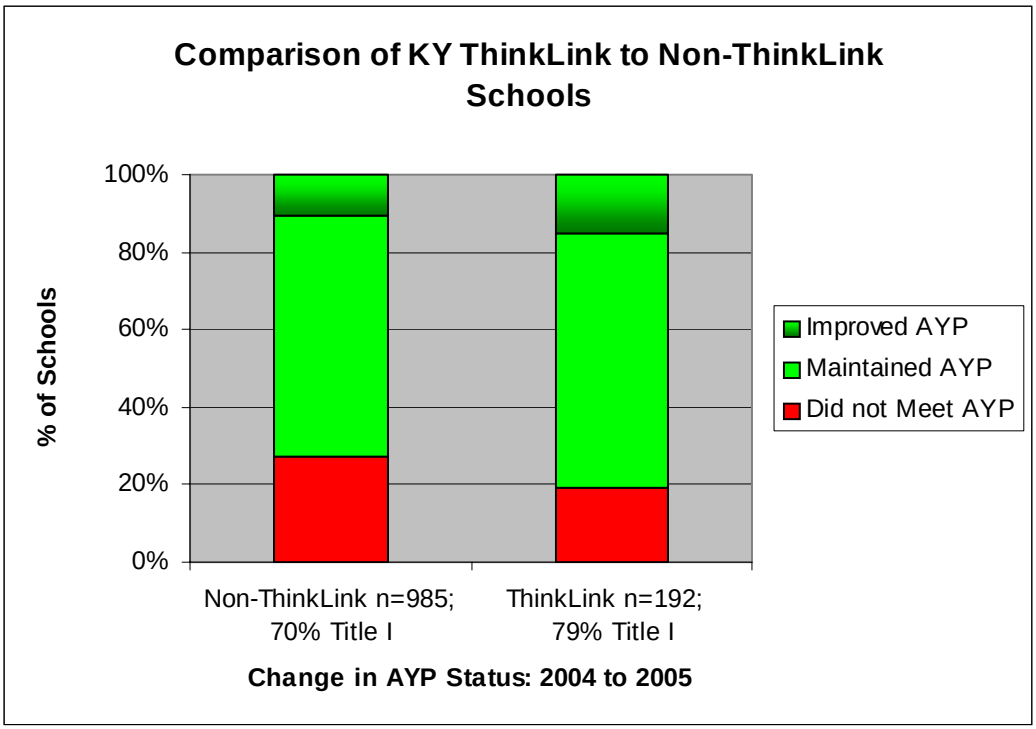


Discovery Education Assessment
RESEARCH

Discovery Education Assessment helped Kentucky schools maintain or improve their AYP status in 2005. During the 2004-2005 school year, 194 schools in Kentucky used Discovery Education benchmark assessments. Approximately 985 schools did not use Discovery Education assessments. Seventy-nine percent of Discovery Education Assessment schools were classified as Title 1; 71% of non-Discovery Education Assessment schools were classified as Title 1. A comparison was made between Discovery Education Assessment schools and non-Discovery Education Assessment schools in terms of Adequate Yearly Progress (AYP) status from Spring 2004 to Spring 2005. Data on AYP status in 2004 and 2005 was maintained from the Kentucky State Department of Education web site. Schools were classified into one of three categories of AYP:

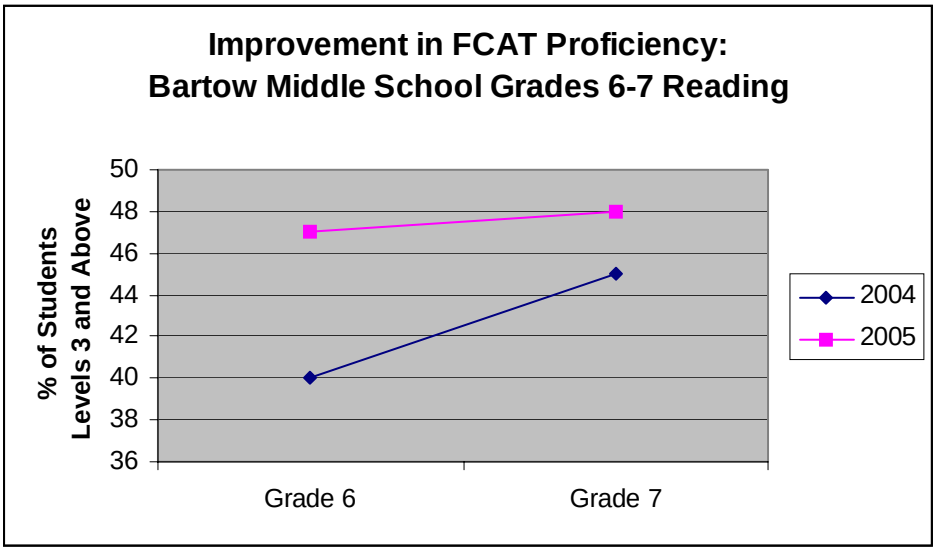
Did not meet AYP:	These schools did not meet AYP in either 2004 or 2005.
Maintained AYP:	These schools obtained AYP in 2004 and maintained the same status in 2005.
Improved AYP:	These schools did not meet AYP in 2004 but obtained AYP in 2005.

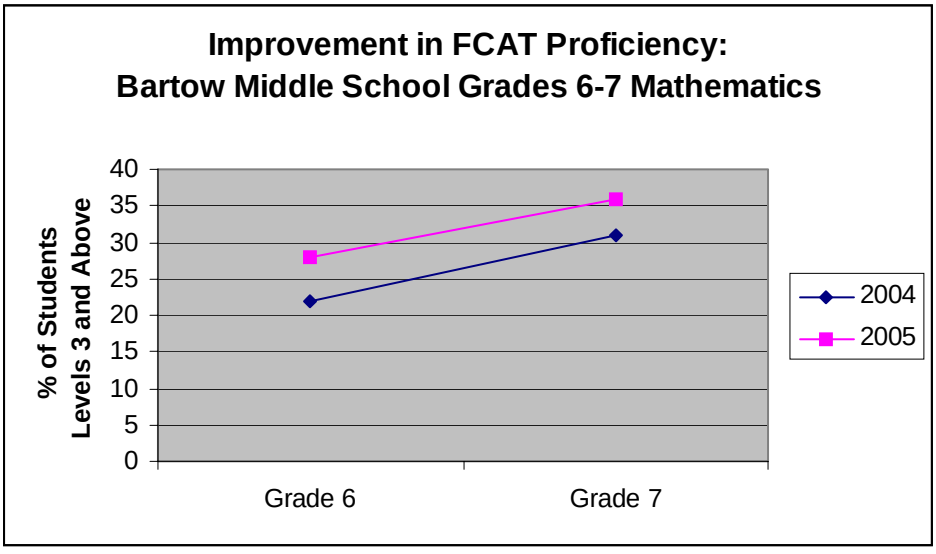
The following graph shows the comparison of these three categories for Discovery Education Assessment and non-Discovery Education Assessment schools.



Discovery Education Assessment
RESEARCH

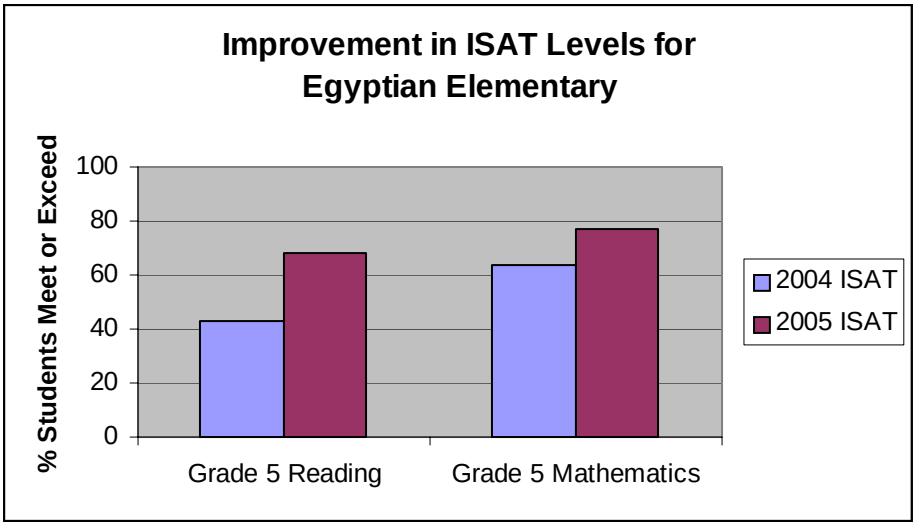
Discovery Education Assessment helped Bartow Middle School increase the number of students at FCAT Reading and Mathematics proficiency levels in Grades 6-7. Bartow Middle School used Discovery Education benchmark assessments during the 2004-2005 school year. These assessments, along with other curricular strategies, helped increase the number of students at Reading and Mathematics Levels 3 and above on the FCAT from the spring of 2004 to the spring of 2005. These increases are displayed in the following graphs.



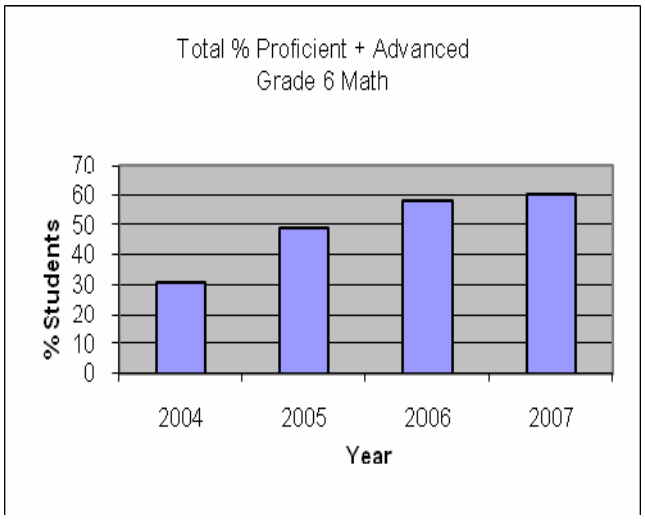


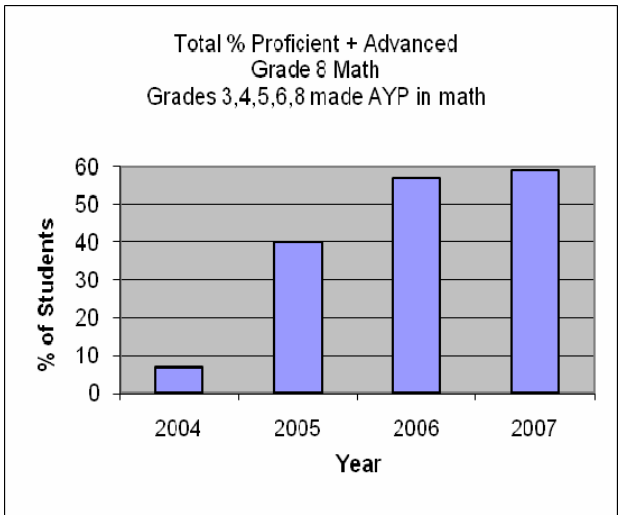
Discovery Education Assessment
RESEARCH

Discovery Education benchmark assessments helped 5th grade students at Egyptian Elementary improve ISAT performance levels from 2004 to 2005. Students at Egyptian Elementary School used Discovery Education benchmark assessments in the school year 2004-2005, particularly in Grade 5. The following graph presents the percent of students who met or exceeded ISAT standards from 2004 to 2005 in the areas of Reading and Mathematics.



The following graphs are two recent examples that offer proficiency comparisons across several years in Birmingham, Alabama. The graphs document sustained improvement in proficiency levels on the ARMT for students in Birmingham, Alabama.





(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.

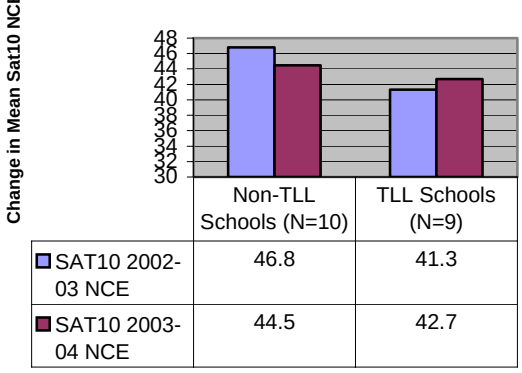
Larger schools and school districts typically do not participate in experimental or quasi-experimental studies due to logistical and ethical concerns. However, a unique situation in Birmingham, Alabama afforded Discovery Education Assessment with the opportunity to investigate the efficacy of the benchmark assessments in respect to a quasi-control group. In 2003-2004, approximately one-half of the schools in Birmingham City used Discovery Education's benchmark 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 PAS as opposed to those that did not. Discovery Education 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 benchmark assessments. 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 benchmark assessments comprise the matched comparison group. The following charts show SAT10 National Percentile changes for Discovery Education Assessment Schools vs. Non-Discovery Education Assessment 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-Discovery Education Assessment 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.

Discovery Education Assessment
RESEARCH

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.

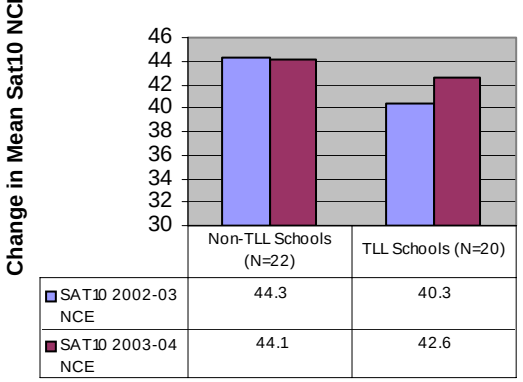
SAT10 Grade 6 Language Comparison of ThinkLink vs. Non-ThinkLink Schools



	Non-TLL Schools (N=10)	TLL Schools (N=9)
SAT10 2002-03 NCE	46.8	41.3
SAT10 2003-04 NCE	44.5	42.7

All Birmingham Schools: Grade 6

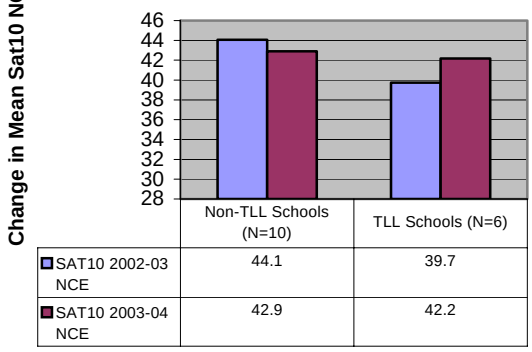
SAT10 Grade 5 Math Comparison of ThinkLink vs. Non-ThinkLink Schools



	Non-TLL Schools (N=22)	TLL Schools (N=20)
SAT10 2002-03 NCE	44.3	40.3
SAT10 2003-04 NCE	44.1	42.6

All Birmingham Schools: Grade 5

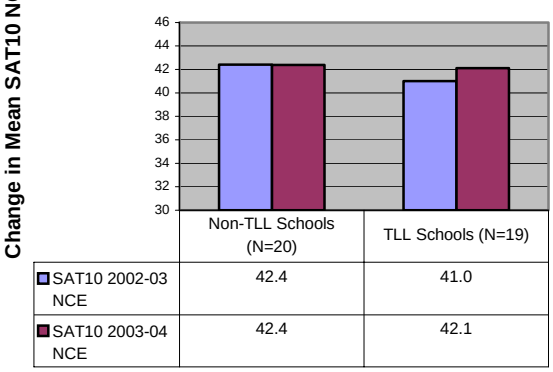
SAT10 Grade 6 Language Comparison of ThinkLink vs. Non-ThinkLink Poverty Schools



	Non-TLL Schools (N=10)	TLL Schools (N=6)
SAT10 2002-03 NCE	44.1	39.7
SAT10 2003-04 NCE	42.9	42.2

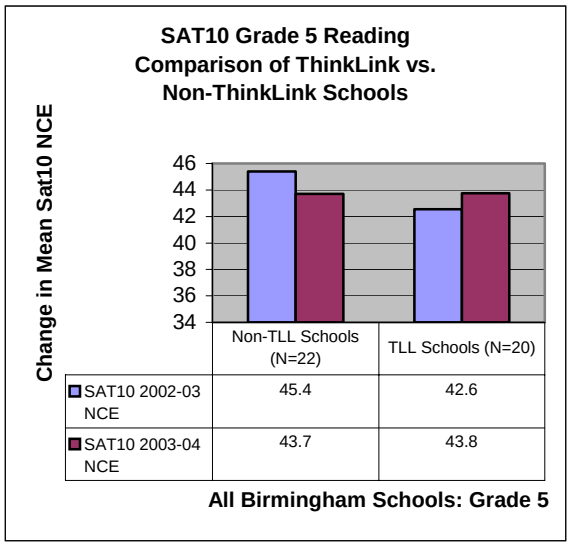
All Birmingham Poverty Schools: Grade 6

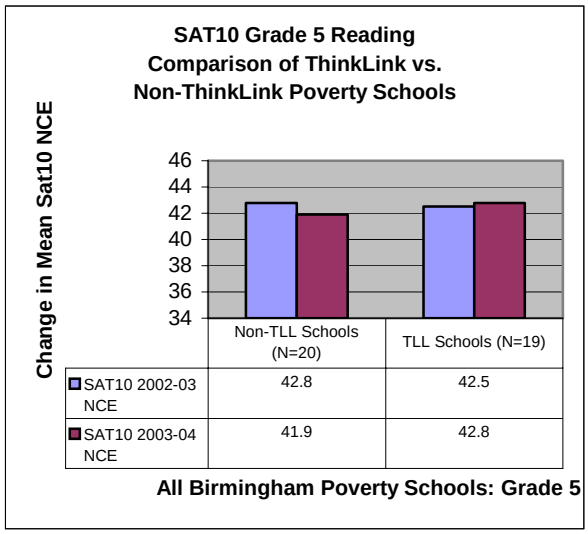
SAT10 Grade 5 Math Comparison of ThinkLink vs. Non-ThinkLink Poverty Schools



	Non-TLL Schools (N=20)	TLL Schools (N=19)
SAT10 2002-03 NCE	42.4	41.0
SAT10 2003-04 NCE	42.4	42.1

All Birmingtham Poverty Schools: Grade 5


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

(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 *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?'."

(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 benchmark tests and results have been incorporated and analyzed in the following peer- reviewed manuscripts and publications:

Jacqueline Shrago and **Dr. Michael Smith** of Discovery Education Assessment 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 ThinkLink Learning 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 (Proposed)
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 benchmark assessments.