



*Action Learning Systems, Inc.*

# **MATHEMATICS**

## **Benchmark Tests**

### **2005 Technical Report**

**7<sup>th</sup> Grade Mathematics**  
**Algebra I, Geometry**  
**Algebra II**

**Action Learning Systems, Inc.**

174 W. Foothill Blvd #306

Monrovia, Ca 91016

Telephone/ (626) 357-8041

Fax/ (626) 357-5031

## TABLE OF CONTENTS

INTRODUCTION	2
DEVELOPMENTAL FRAMEWORK OF THE BENCHMARK TESTS	2
Description of the Content Standards Measured	3
Focus Standards	3
Description of the Performance Levels	4
Administration of the Benchmark Tests	4
EVIDENCE OF VALIDITY AND RELIABILITY	4
Item Development	4
Reliability Characteristics	5
Demographics of Participants in 2004-05 Test Administration	6
Relationships between Benchmark Tests and CST	7
Disaggregated Analyses by Language Fluency	8
CONCLUSION	10
Results Summary	10
Future Steps	10
Acknowledgments	11
REFERENCES	11
APPENDIX – Benchmark Test Reports	12

## LIST OF TABLES

1. Average Reliability Coefficients for Mathematics Benchmark Tests	6
2. Demographics based on 2005 CA Dept. of Education Statistics	6
3. Students with at least 1 Benchmark Score and a Valid CST Score (by Grade Level)	7
4. Correlations between ALS' Mathematics Benchmark Tests and 2005 CST Math Scaled Scores	8
5. Breakdown of Students by Language Fluency	9
6. Average Correlations between ALS' Mathematics Benchmark Tests and 2005 CST Math Scaled Scores by Language Fluency Subgroups	9

## **INTRODUCTION**

On Jan. 8, 2002, President Bush signed into law the *No Child Left Behind Act of 2002 (NCLB)*. This law represented the federal government's most extensive restructure of the 1965 Elementary and Secondary Education Act (ESEA). This act incorporated the principles and strategies proposed by President Bush. These included increased accountability for states, school districts, and schools; greater choice for parents and students, particularly those attending low-performing schools; more flexibility for states and local educational agencies (LEA's) in the use of Federal education dollars; and a stronger emphasis on reading, especially for our youngest children. A major component of NCLB is that all students will attain "proficiency" in reading and mathematics by 2014, including students with disabilities and English learners.

The state of California's NCLB accountability plan has embraced high quality tests aligned with state adopted standards. Action Learning Systems Inc. (ALS) addressed NCLB legislation and state standards with its development of curriculum-aligned, formative benchmark tests.

Through its creation of content, and performance standards for English-Language Arts and Mathematics, the state of California has defined what a student should know and at what level of proficiency. Through the adoption of these standards, the state has clearly affirmed what content students need to acquire at each grade level. With these standards in place, student achievement and mastery of these standards are measured with the California Standards Tests (CST), criterion-referenced tests developed specifically for California. As part of the state's accountability system, performance on the CST also constitutes the largest component of a school's API (Academic Performance Index).

Research has consistently shown that the use of formative tests (i.e., benchmark tests) is a strongly recommended method to gauge mastery throughout the school year, provide teachers with diagnostic and prescriptive information, and provide students with test-taking skills. To assist districts, schools, and teachers, ALS has implemented a focus on these standards through formative benchmark testing.

## **DEVELOPMENTAL FRAMEWORK OF THE BENCHMARK TESTS**

ALS has developed Benchmark tests to measure student progress in mastering the California Standards at specific grade levels and content areas. The term benchmark was

adopted to emphasize the concept of on-going assessment throughout the year, at key instructional points, prior to the annual administration of the state's on-demand assessments.

### **Description of the Content Standards Measured**

The State Board adopted the California English-Language Arts content standards in November 1997 and the Mathematics standards in December 1998. These standards designated the content to be taught and what all students should be proficient in by the end of each grade level in the respective content areas. The NCLB legislation requires that all students be at or above "proficient" in these two content areas by the year 2014. In California, "proficient" or above is determined by performance on the CST. In addition, prior to 2014, schools are to have met designated annual measurable objectives (AMO), i.e., percentage of students required to be at "proficient" or above.

At each grade level, there are numerous standards designated to be taught in one school year. With these large numbers, the requirement that all students master all standards and reach the levels of "proficient" or above on these high stakes tests is unrealistic. For example, in English-Language Arts the number of standards to be taught in grade four is 67; at grade seven it is 65; and at grade nine is 107. In response to the impracticality of teaching and mastering these large numbers of prescribed standards, ALS has selected a smaller number of standards that are considered essential for each grade level and discipline. These essential standards, also referred to as "power" or "focus standards," were selected by content area experts after a thorough review of state standards, the determination and weighting of the most tested items on state tests, and instructional sequence.

### **Focus Standards**

ALS engaged a representative group of in-service teachers and curriculum specialists to identify by grade level and subject, the most salient and/or important standards for devoting instructional emphasis. These were identified as "focus standards." Once the focus standards were determined, blueprints were developed for each grade and content area. All blueprints were directly aligned with the CST and were reviewed by teams of content-area experts.

## **Description of the Performance Levels**

ALS Benchmark tests and the CST are both criterion-referenced assessments. As such, these tests compare a student's score with a common standard of performance. Percent-correct scores determine whether a student has established minimum acceptable performance. The ALS Benchmark test results are reported using the same performance levels as are used with the CST (i.e., Advanced, Proficient, Basic, Below Basic, and Far Below Basic).

Future analyses will determine the statistical relationships between the performance level designations on the Benchmark tests to the CST designated performance levels. At this time, the Benchmark performance levels should be used as a formative indicator and not an exact predictor of a student's CST performance.

## **Administration of the Benchmark Tests**

As a formative tool, the series of Benchmark tests are sequenced, for the most part, to be administered over the course of the school year, prior to the annual spring administration of the CST. It is intended that the tests be administered in one normally scheduled class period. Students taking those classes for which Benchmark tests have been developed should take the tests in that class. All benchmark tests assess mastery of standards with multiple-choice questions. Each question contains one item stem and four distracters. Students record their response to each question on a separate answer sheet (i.e., scantron sheet). Overall, each Benchmark test should not take more than 40 minutes to administer. To maintain validity and reliability of the assessment results, it is critical that the Benchmark tests remain intact – meaning that the items are maintained as written and assessed in the specific order or format of the particular test.

## **EVIDENCE OF VALIDITY AND RELIABILITY**

### **Item Development**

Teams of professional item writers developed the format and items for the ALS Benchmark tests as prescribed by the CST blueprints. The teams were designated by content area and each member had several years of experience in developing test items for standards-based state assessment programs (e.g. Golden State Exam, CLAS, CST, and

CAHSEE). Lead team members also held lead or supervisory positions in state test development programs.

Team members reviewed grade appropriate textbooks for item development. Prior to developing the math items, the teams reached agreement on the instructional pacing sequence by content by grade. Each focus standard was assessed with a minimum of three items. Each item had four distracters. English-Language Arts reading passages and Mathematics word problems were selected by grade level and length that corresponded to those utilized on the California standards-based assessments.

All Benchmark tests have gone through rigorous field-testing processes. Item analyses were performed, which included meticulous analyses of p-values, pt. biserial coefficients, and other indices of discrimination, after each round of field-testing. The acceptability of difficulty levels included percent correct in the 30 percent to 80 percent range. The discrimination level for each item was at or above 0.3. Items not meeting psychometric criteria were either eliminated and replaced, or modified. New versions were subsequently field tested until the complete test had been determined psychometrically sound.

English-Language Arts and Mathematics subject experts were selected to participate in a validation study. These experts carefully reviewed all items on each benchmark test to determine how well they measured the “focus standards.” Each item was reviewed for alignment to content-grade-specific focus standards, instructional validity (i.e., appropriate grade-level vocabulary and sentence structure, etc.), and bias (i.e., gender, ethnic, offensive language or situations, etc.). This validation review met with the standards as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Individual items from the Mathematics Benchmark tests were developed and tested over a two year span. Item analyses led to refinement of the items included in the 2004-05 administrations of these Benchmark tests.

### **Reliability Characteristics**

Reliability refers to the consistency of measurement or how stable test scores are from one test administration to another. Since one test administration can realistically only furnish a limited indication of performance at a particular point in time, evidence of

reliability can estimate the stability of the scores over several administrations. Reliability is a necessary condition for validity. If the scores are not consistent, they could not provide valid information about what is being measured. To estimate the reliabilities of the Mathematics Benchmark Tests, Cronbach’s alphas were calculated as measures of internal consistency (see Table 1). The alphas presented are considered acceptable levels of internal consistency (Nunnally, 1976).

**Table 1.**

**Average Reliability Coefficients <sup>a</sup> for Mathematics Benchmark Tests**

	<b>Cronbach’s alpha</b>	<b><i>n</i></b>
<b>7<sup>th</sup> grade Math</b>	.70	4154
<b>Algebra I</b>	.72	2560
<b>Geometry</b>	.70	2803
<b>Algebra II</b>	.65	1257

<sup>a</sup>Coefficients were averaged across BM1 – BM3 for each subject area

**Demographics of Participants in 2004-05 Test Administrations**

Students from an urban school district in the Southern California region participated in these test administrations. Demographics of the student population are presented in Table 2.

**Table 2.**

**Demographics based on 2005 CA Dept. of Education Statistics**

<b>Characteristic</b>	
ALL STUDENTS	<i>N</i> ≈ 50,000 <sup>a</sup>
Ethnic Subgroups:	
Hispanic	52%
Asian	28%
White	16%
**All other ethnic subgroups were at 3% or less	
Socio Economic Disadvantaged	61%
English Learners	47%

<sup>a</sup>Total student population was rounded to protect confidentiality of district

Demographically, there were three predominant ethnic subgroups: Hispanic (52%), Asian (28%), and White (16%). Sixty-one percent of the students were classified as “Socio Economic Disadvantaged” (as indicated by participation in the free/reduced lunch program), and forty-seven percent were classified as English Learners.

Mathematics Benchmark Test scores of over 8,000 students in grades 7 through 11 were analyzed in relation to their 2005 CST Math scores. Table 3 provides the student numbers by grade level and specific content areas. Only students who had a valid 2005 CST score and at least one Benchmark score were included in these analyses.

**Table 3.**

<b>Students with at least 1 Benchmark score and a Valid CST Score (by Grade Level)</b>						
	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>Total</b>
<b>7<sup>th</sup> grade Math</b>	3292	---	---	---	---	3292
<b>Algebra I</b>	---	1188	719	165	84	2156
<b>Geometry</b>	---	---	870	953	644	2467
<b>Algebra II</b>	---	---	---	350	570	920
<b>Total</b>	3292	1188	1589	1468	1298	8835

### **Relationships between Benchmark Tests and CST**

In describing sources of validity evidence, the *Standards for Educational and Psychological Testing* explained that “relationships between test scores and other measures intended to assess similar constructs provide convergent evidence...” (p. 14). Therefore the correlational relationships between the Mathematics Benchmark scores and the California Standards Test scores provide convergent evidence of validity. This evidence strongly supports the major premise in the development and use of ALS Benchmarks. Positive and statistically significant correlations between the Benchmark Test scores and CST scores indicate that a change in one set of scores is associated with a corresponding change in the other. In other words, if a student scores high on the ALS Benchmark test, that student is more likely to score high on the CST (and vice-versa).

These data show strong positive correlations (ranging from .56 to .75) between the ALS Mathematics Benchmark Tests and the 2005 CST Mathematics tests for the specific math content areas measured. Table 4 displays the breakdown of correlations for each area. The numbers in parentheses represent the numbers of students that had a Benchmark score and a valid CST score. BM1 through BM3 represent the three benchmark components that were administered over the school year prior to the CST administration.

**Table 4.**

**Correlations between ALS' Mathematics Benchmark Tests and 2005 CST Math Scaled Scores**

<b>BM Area</b>	<b>BM1</b>	<b>BM2</b>	<b>BM3</b>	<b>CST Test</b>
<b>7<sup>th</sup> grade Math</b>	.75** ( <i>n</i> =3051)	.72** ( <i>n</i> =3157)	.64** ( <i>n</i> =3200)	CST Math – gr. 7
<b>Algebra I</b>	.59** ( <i>n</i> =2053)	.65** ( <i>n</i> =2094)	.65** ( <i>n</i> =1975)	Algebra I
<b>Geometry</b>	.65** ( <i>n</i> =2398)	.74** ( <i>n</i> =2412)	.72** ( <i>n</i> =2405)	Geometry
<b>Algebra II</b>	.56** ( <i>n</i> =876)	.60** ( <i>n</i> =902)	.56** ( <i>n</i> =840)	Algebra II

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

As may be seen in Table 4, the correlations between the Benchmark test scores and the CST scores for this group of students were strongly correlated. These results provide convergent evidence of the validity of the Benchmark test scores and lend support to the capability of the Benchmark tests in predicting CST performance. Future research and analyses will focus in this area.

### **Disaggregated Analyses by Language Fluency**

Many districts in California are located in geographic areas where the student population may have a significant percentage of English Learners. Therefore it is essential that this type of assessment be able to show similar evidence of validity for these subgroups. For each of the four Mathematics content areas, disaggregated analyses were

performed to provide such evidence. Table 5 displays the overall breakdown of students by language fluency.

**Table 5.**

<b>Breakdown of Students by Language Fluency (<i>n</i> = 8833)</b>	
English Only	34%
English Learners	31%
Re-Designated English Proficient (R-FEP)	27%
Initially Fluent English Proficient (I-FEP)	8%
	100%

Students designated as Initially Fluent English Proficient (I-FEP) were removed from the disaggregated analyses because of this group’s small size. In addition, since R-FEP students are considered English Learners by the state’s accountability system, these two groups were combined. Table 6 displays the average correlations for the English Only and English Learner subgroups.

**Table 6.**

<b>Average Correlations between ALS’ Mathematics Benchmark Tests and 2005 CST Math Scaled Scores by Language Fluency Subgroup</b>		
<b>BM Area</b>	<b>English Only</b>	<b>English Learner</b>
<b>7<sup>th</sup> grade Math</b>	.66	.63
<b>Algebra I</b>	.62	.62
<b>Geometry</b>	.72	.68
<b>Algebra II</b>	.52	.58

\*\*All individual BM correlations are significant at  $p < .01$

As may be seen in the table, the correlations do not differ significantly from one group to the other. These significant correlations provide strong support for the use of Mathematics Benchmark Tests with English Learner students.

## CONCLUSION

ALS' focus on standards and benchmark tests aligned to the standards has emerged as a result of state and federal requirements (i.e., NCLB). It is expected that through the use of benchmark testing students will come to expect and demand meaningful assignments with clear purposes, i.e., standards-based. They will understand the idea of looking at exemplars to help them understand the quality of work expected of them. Teachers will develop units that must be organized around standards. Teachers' activities will be justified in terms of standards. Teachers will use benchmark test results as formative tools as they prepare students to learn how to reason, apply knowledge, and produce quality work. ALS Benchmark Tests carefully aligned to clear instructional objectives can be a means of raising student motivation and achievement. The student test cycle is critical if students are to perform at higher levels.

### Results Summary

The technical information provided in this report strongly supports the use of the Mathematics Benchmark Tests as a formative measure aimed at improving student performance on the California Standards Tests in Mathematics.

- 7<sup>th</sup> grade Mathematics, Algebra I, Geometry, and Algebra II Benchmark Tests administered in 2004-2005 were all significantly, positively correlated to the 2005 CST Math scores.
- Lower, but significant correlations observed for Algebra and Geometry Benchmark Tests may be influenced by the wider span of student grade levels.
- Disaggregated analyses by language fluency strongly support the use of benchmarks among the English Learner student population.
- All four Mathematics Benchmark Tests have yielded acceptable levels of reliability and have demonstrated evidence of convergent validity.

### Future Steps

In the effort to provide the most thorough information regarding test development, reliability, and validity, steps are currently being taken to provide technical information for all Benchmark tests currently developed by Action Learning Systems, Inc. Field-testing and item/assessment refinement are part of ALS' continuous process to

improve the Benchmarks' reliability and validity. We will also continue to update our technical reports in order to provide the most current and accurate information.

### **Acknowledgements**

Action Learning Systems, Inc. would like to recognize the contributions and expertise of all members of the test development team for their past, current, and future efforts. ALS would also like to thank the schools, teachers, and students who have participated in these and all other test administrations.

### **REFERENCES**

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. New Jersey: Prentice-Hall.
- Nunnally, J. C. (1976). *Psychometric theory, 2<sup>nd</sup> ed.* New York: McGraw-Hill.

## APPENDIX

### Benchmark Test Reports

As prescribed by the state performance levels, mastery is considered at “proficient” or above. NCLB requires that a specific percentage of all students meet this level of proficiency each year. Upon each administration of the Benchmark test, results show those objectives either being mastered or not being mastered by the student.

In a joint venture with *Achieve Data Solutions LLC—Data Director*, three distinct Benchmark reports have been developed and may be available for use by students, teachers, and administrators that provide valuable information about performance levels and mastery of standards:

- The **Student Exam Report** includes the response made for each question, and related standard, noting whether the response was correct; the number and percent correct; the student’s performance level; and the number correct for each standard.
- The **Classroom Exam Report** developed for each teacher’s classroom, includes the frequency of response for each multiple-choice item and standard; the correct response for each question; the average number and percent correct for the classroom; the number of students in each performance level; and the number and percent of students answering each specific standard correctly.
- The **School Exam Report** includes, school wide, the percent correct for each classroom’s result by standard; the overall percent correct for each standard; and the number of students in each performance level.

Each report provides information on the performance level attained either by student, classroom or school. With the classroom and school reports, the annual measurable objective (AMO) rate may be calculated. Samples of all three of these *Data Director* reports begin on the following page.











