



Action Learning Systems, Inc.

ENGLISH/LANGUAGE ARTS Benchmark Tests (Form 4) 2005 Technical Report

(GRADES SEVEN THROUGH ELEVEN)

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	3
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	5
Item Development	5
Reliability Characteristics	6
Demographics of Participants in 2004-05 Test Administration	6
Relationships between Benchmark Tests and CST	8
Disaggregated Analyses by Language Fluency	9
CONCLUSION	10
Results Summary	11
Future Steps	11
Acknowledgments	11
REFERENCES	12
APPENDIX – Benchmark Reports	13

LIST OF TABLES

1. Average Reliability Coefficients for English/Language Arts Benchmark Tests	6
2. District Demographics based on 2005 CA Dept. of Education Statistics	7
3. Students with at least 1 Benchmark Score and a Valid CST ELA Score (by Grade Level)	7
4. Correlations between ALS' English/Language Arts Benchmark Tests and 2005 CST ELA Scaled Scores	8
5. Breakdown of Students by Language Fluency and Percentages By Grade Level	9
6. Average Correlations between ALS' English/Language Arts Benchmark Tests and 2005 CST ELA Scaled Scores by Language Fluency Subgroup	10

INTRODUCTION

On Jan. 8, 2002, President Bush signed into law the *No Child Left Behind Act of 2002 (NCLB)*. This law represents the federal government's most extensive restructure of the 1965 Elementary and Secondary Education Act (ESEA). This act incorporates the principles and strategies proposed by President Bush. These include 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 embraces high quality tests aligned with state adopted standards. Action Learning Systems Inc. (ALS) addresses 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 is clearly affirming 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 in specific grades and/or subjects. 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 of 1997 and the Mathematics standards in December of 1998. These standards designate 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/or mathematics 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/or course. 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 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, Benchmark tests should not take more than 40 minutes each to administer. To maintain validity and reliability of the assessment results, it is critical that the Benchmark tests remain intact – meaning that the items written and assessed in a specific order or format.

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 include 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 that 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 are either eliminated and replaced, or modified. New versions are subsequently field tested until the complete test has 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 standard, 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 English/Language Arts Benchmark tests were developed and tested over a span of several years. This version, Form 4, is the most current form of

the English/Language Arts tests. 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 English/Language Arts Benchmark Tests, Cronbach’s alphas were calculated as measures of internal consistency (see Table 1). The alphas presented are considered strong levels of internal consistency (Nunnally, 1976).

Table 1.
Average Reliability Coefficients ^a for English/Language Arts Benchmark Tests

Grade Level	Cronbach’s alpha	<i>n</i>
7	.82	4004
8	.81	3964
9	.84	4768
10	.85	3881
11	.82	3165

^aCoefficients were averaged across BM1 – BM3 for each subject area

Demographics of Participants in 2004-05 Test Administrations

Students from two school districts in urban regions within California participated in these test administrations. Demographics of the student populations are presented in Table 2.

Table 2.

District Demographics based on 2005 CA Dept. of Education Statistics	
Characteristic	TOTAL
ALL STUDENTS	≈ 89,000 ^a
Ethnic Subgroups:	
Hispanic	53%
Asian	21%
White	14%
African-American	6%
**All other ethnic subgroups were at 5% or less	
Socio Economic Disadvantaged	65%
English Learners	38%

^aThis total represents a combination of both districts and has been rounded to protect anonymity.

Demographically, there were three predominant ethnic subgroups: Hispanic (53%), Asian (21%), and White (14%). Sixty-five percent of the students were classified as “Socio Economic Disadvantaged” (as indicated by participation in the free/reduced lunch program), and thirty-eight percent were classified as English Learners.

Of this group, the English/Language Arts Benchmark Test scores of over 20,000 students in grades 7 through 11 were analyzed in relation to their 2005 CST ELA scores. Table 3 provides the student numbers by grade level. Only students who had a valid 2005 CST ELA score and at least one Benchmark score were included in these analyses.

Table 3.

Students with at least 1 ELA Benchmark score and a Valid 2005 CST Score (by Grade Level)	
Grade Level	Number of Students
7	4566
8	4578
9	4473
10	3686
11	3059
Total	20362

Relationships between Benchmark Tests and CST

In describing sources of validity evidence, the *Standards* explain that “relationships between test scores and other measures intended to assess similar constructs provide convergent evidence...” (p. 14). Therefore the correlational relationships between the English/Language Arts Benchmark Test 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).

The data show, in fact, there are strong positive correlations (ranging from .58 to .83) between the ALS English/Language Arts Benchmark Tests and the 2005 CST ELA tests for each of the five grade levels. Table 4 displays the breakdown of correlations for each. 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.

Summary of Correlations between ALS’ English/Language Arts Benchmark Tests and 2005 CST Scaled Scores

Grade Level	BM1	BM2	BM3
7	.78** (n=4207)	.80** (n=4217)	.81** (n=3916)
8	.79** (n=4303)	.81** (n=4273)	.81** (n=3845)
9	.80** (n=4076)	.81** (n=4129)	.82** (n=3408)
10	.72** (n=3485)	.81** (n=3428)	.83** (n=2915)
11	.58** (n=2761)	.66** (n=2922)	.77** (n=2478)

**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. At each grade level, disaggregated analyses were performed to provide such evidence. Table 5 displays the overall breakdown of students by language fluency and grade level.

Table 5.
Breakdown of Students by Language Fluency and Percentages by Grade Level
(N=20358)

Classification	7	8	9	10	11	TOTAL
English Only	36%	36%	36%	37%	36%	36%
English Learners	32%	31%	31%	26%	25%	30%
Re-Designated English Proficient (R-FEP)	23%	24%	24%	27%	31%	25%
Initially Fluent English Proficient (I-FEP)	9%	9%	9%	10%	8%	9%
Totals	4565	4576	4473	3685	3059	20358

Students designated as I-FEP were removed from the disaggregated analyses because of the group’s overall small size. Students classified as R-FEP are considered English Learners by the state’s accountability system and could be combined with the English Learner subgroup. However, since the R-FEP group’s size was comparable to the other groups, R-FEP student scores were analyzed as an intact group. English language development (ELD) levels for the English Learner students were not known and therefore

analyses could not be further disaggregated. Table 6 displays the average correlations for the language fluency subgroups.

Table 6.

Average Correlations between ALS' English/Language Arts Benchmark Tests and 2005 CST ELA Scaled Scores by Language Fluency Subgroup

Grade Level	English Only	English Learners	R-FEP
7	.79**	.71**	.73**
8	.81**	.67**	.74**
9	.81**	.66**	.76**
10	.78**	.71**	.72**
11	.66**	.63**	.58**

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

As may be seen in the table, all correlations for all language subgroups were statistically significant. Correlations for the English Learner subgroup were lower at some grade levels; this may be due to the influence of various ELD levels per grade level. However, all correlations were significantly positive providing strong support for the use of English/Language Arts Benchmark Tests with English Learner students and other language subgroups.

CONCLUSION

ALS' focus on standards and benchmark tests aligned with 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 English/Language Arts Benchmark Tests as a formative measure aimed at improving student performance on the California Standards Tests.

- English/Language Arts Benchmark Tests administered in 2004-2005 for grade levels 7 through 11 were all significantly, positively correlated to the corresponding 2005 CST ELA scores.
- Disaggregated analyses by language fluency strongly support the use of benchmarks among the English Learner and R-FEP student populations.
- Lower correlations among test scores for English Learner students were likely influenced by the ELD levels of the students. However, all correlations for this subgroup were statistically significant.
- All English/Language Arts Benchmark Tests (Form 4) 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, 2nd 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.

