

A Study of the Reliability and Validity of the *Rigby ELL Assessment Kit* Benchmark Tests

Report 226

December 2005

Directors:

Jennifer M. Conner
Ph.D. Indiana University

Beth G. Greene
Ph.D. New York University

Kimberly Munroe
M.Ed. University of Massachusetts

Advisory Board:

Michael Beck, President
Beck Evaluation & Testing Associates, Inc.

Keith Cruse, Former Managing Director
Texas Assessment Program

Joseph A. Fernandez, Former Chancellor
New York City Public Schools

This page intentionally left blank.

A Study of the Reliability and Validity of the *Rigby ELL Assessment Kit* Benchmark Tests

Table of Contents

Executive Summary.....	4
Reliability of the ELL-A Benchmark Test Open-Ended Scoring.....	4
Reliability of the ELL-A Benchmark Test Multiple-Choice Items	4
Validity of the Content Evidenced by Teacher Survey Results.....	4
Full Report	5
Background Information.....	5
Research Questions	5
Design of the Study.....	5
Description of Schools and Teachers Participating in the Study	5
Description of Assessment.....	6
Data Analyses.....	6
Results of the Analysis	7
Conclusions.....	11

A Study of the Reliability and Validity of the *Rigby ELL Assessment Kit* Benchmark Tests in Reading Comprehension and Writing

Executive Summary

This report describes a study of the reliability and validity of Harcourt Achieve’s *Rigby ELL Assessment Kit* Benchmark Tests in Reading Comprehension and Writing. Harcourt Achieve contracted with the Educational Research Institute of America (ERIA) to conduct the study. This executive summary, on pages 2-3, provides an overview of the study’s findings. More detailed information about the study procedures, data analysis, and findings are provided within the text of the full report, on pages 4-13.

In a month-long study, 256 students enrolled in English Language Learner (ELL) programs in six different schools in Indiana school districts were administered the ELL-A Benchmark Tests. It should be noted that the original sample included 276 students, but some students did not complete testing due to absences or ineligibility to participate in the study based on teacher discretion supporting that the content was inappropriate for the student’s level of learning.

Students were administered an ELL-A Benchmark Test at Level C, Level I, or Level O based on teacher determination of the appropriate test based on each student’s level of learning. Although the ELL-A Benchmark Test at each level consists of four sub-tests—Listening, Speaking, Reading, and Writing—for the purposes of this study, students only completed the Reading and Writing sub-tests. The Reading sub-test was comprised of multiple-choice items. The open-ended writing responses that comprised the Writing sub-test were scored and then independently rescored using a rubric that was field tested in an earlier study. The rescore served as a means of gaining inter-rater consistency of scoring data. ERIA completed all scoring and data entry for the study. Teachers involved in the study also completed Teacher Surveys to help determine content validity and results are summarized in the full report.

Researchers at ERIA examined the resulting data in three areas:

- Reliability of the ELL-A Benchmark Test Open-Ended Scoring
- Reliability of the ELL-A Benchmark Test Multiple-Choice Items
- Validity of the ELL-A Benchmark Test Content

Reliability of the ELL-A Benchmark Test Open-Ended Scoring

The test scorer correlations indicate that the scorer reliability was high for the writing test on each of the three Benchmark Tests.

Two scorers independently scored the writing section of each of the three Benchmark Tests. Correlations between scorers were high overall, indicating that the scoring for the ELL-A Writing sub-test is reliable. Detailed tables for Level C, Level I, and Level O are included in the full report.

Reliability of the ELL-A Benchmark Test Multiple-Choice Items

Data suggest that overall the multiple-choice items on each of the three Benchmark Tests are reasonably difficult and discriminate well.

Test item difficulties and discriminations for the multiple-choice reading test items were also examined, and the results indicate that for all three levels the item data indicated reasonable levels of difficulty and discrimination.

Validity of the Content Evidenced by Teacher Survey Results

Data suggest that overall teachers found the content of the ELL-A tests appropriate.

The examination of teacher survey results indicates that teachers found the content of the Reading and Writing sections of the Benchmark Test valid. Teachers were asked how appropriate they considered each sub-test. In Reading, 100% of teachers considered the content very or mostly appropriate, and in Writing, 89% considered the content very or mostly appropriate.

A Study of the Reliability and Validity of the *Rigby ELL Assessment Kit* Benchmark Tests

Full Report

Background Information

Harcourt Achieve contracted with the Educational Research Institute of America (ERIA) to conduct a study of the reliability and validity of the Benchmark Tests within the *Rigby ELL Assessment Kit*, a new program to be published in 2007. Harcourt Achieve wanted to tryout three levels of the ELL-A Benchmark Tests (Level C, Level I, and Level O) to determine scoring and item reliability and validity of content for two of the four sections (Reading and Writing) of these Benchmark Tests.

The study spanned one month in the 2005-2006 academic year. Ten teachers of English Language Learner (ELL) programs from six different schools in Indiana participated in the study. The sample included a total of 256 students.

Research Questions

The following research questions guided the design of the study and the data analysis:

- Are the open-ended written responses that comprise the ELL-A Benchmark Test Writing sub-test reliably and consistently scored, as evidenced by inter-rater agreement on tests at Levels C, I, and O?
- Are the student-selected (multiple-choice) items within the Reading sub-test of the ELL-A Benchmark Test reliable measurements, as evidenced by the Kuder-Richardson reliability index on tests at Levels C, I, and O?
- Is the content of each ELL-A Benchmark Test, represented on tests at Levels C, I, and O, valid and useful to teachers, as reported by teacher survey results?

Design of the Study

A study was conducted in which the ELL-A Benchmark Tests were administered by ELL teachers and data were analyzed by ERIA. The participating teachers were instructed to administer the appropriate Benchmark Test based on their determination of appropriateness of content for varying student levels. Below is the timeline of the study:

Early October 2005	Tests, student guides, and teacher guides sent to schools.
Mid to Late October 2005	Teachers administered Benchmark Tests.
Early November 2005	Tests and materials returned to ERIA.

Teachers were allowed to administer the assessments according to a schedule that met their particular instructional plans and goals. However, teachers were expected to complete testing within the month of October.

Teachers who took part in the study were asked to complete a survey. They were asked questions regarding the content and usefulness of both tests.

Description of Schools and Teachers Participating in the Study

The sample originally included twelve teachers of English Language Learner programs and 276 students from six different elementary schools located in Indiana. Of the twelve teachers, ten ultimately fulfilled the requirements of the study. Of the total students enrolled in the classes that took part, 256 students were administered a Benchmark Test; these students comprised the total sample used in the statistical analyses of scores. Student absences and ineligibility to participate in the study based on teacher discretion resulted in the smaller number that comprised the sample. The teachers who volunteered for the study were all licensed and teaching ELL students at the time of the study.

Table 1 below provides a summary of the six schools included in the study. This school data does not provide a description of the make-up of each of the classes that participated in the study. However, the data does provide a description of each of the schools and, thereby, an estimate of the make-up of the classes that participated in the study. Three of the schools enrolled students in Grades Pre-K to 5, and one of the schools enrolled students in Grades K to 6. Two of the schools enrolled students in Grades 9 to 12.

Table 1 indicates that the average enrollment for the participating schools was 977 and that an average of 36% of the students in these schools were enrolled in a free or reduced price lunch program and an average of 43% of the students were classified as minority students. On average, 17% of the enrolled students were classified as special education, and an average of 7% of students were identified as *Limited English Proficient* (LEP).

Description of Assessment

Although the Benchmark Tests contain four sections—Listening, Speaking, Reading, and Writing—at Levels C, I, and O, for this study, only two sub-tests were administered to assess two areas of student performance: Reading and Writing. A description of the sub-tests at each of the levels is noted in Table 2.

Data Analyses

All of the students' Benchmark Tests and the teacher surveys were returned to ERIA for analysis. Three separate analyses were conducted and reported.

The Benchmark Tests were analyzed for reliability of the multiple-choice items and reliability of open-ended scoring using the Kuder-Richardson index. Two independent scorers scored the open-ended items to provide data to analyze the correlation

Table 1
Demographic Data for the Six Schools Included in the ELL-A Reliability and Validity Study

School	State	Locale	Grade Range	Total Enrolled	Percent Free/ Reduced Lunch	Percent Minority	Percent Special Education	Percent LEP Enrolled
1	IN	Mid-size Central City	K-6	508	6%	15%	16%	6%
2	IN	Mid-size Central City	PK-5	628	90%	82%	17%	7%
3	IN	Mid-size Central City	PK-5	386	64%	44%	26%	10%
4	IN	Mid-size City	PK-5	342	9%	70%	15%	6%
5	IN	Urban Fringe of Large City	9-12	2611	7%	10%	11%	4%
6	IN	Mid-size Central City	9-12	1384	41%	34%	17%	7%
Averages				977	36%	43%	17%	7%

Table 2
Description of Benchmark Sub-tests (Level C, Level I, and Level O)

	Level C	Level I	Level O
Reading Comprehension	Twelve multiple-choice items	Eighteen multiple-choice items	Eighteen multiple-choice items
Writing	Graphic organizer and writing piece	Graphic organizer and writing piece	Graphic organizer and writing piece

between Scorer 1 and Scorer 2. The Kuder-Richardson reliability index was computed for each set of test items, and the student averages and the item difficulties and discriminations for each multiple-choice test item were computed. Content validity was also examined based on teacher survey results.

Results of the Analysis

Reliability Data – Benchmark Tests

The reliability of the Benchmark Tests was examined in two parts. The first was to determine the correlation between Scorer 1 and Scorer 2 on the scoring of the writing section. The second was to analyze the multiple-choice items of the reading sections using the Kuder-Richardson reliability index.

LEVEL C BENCHMARK TEST – Reliability Results

The Level C Writing tests were scored by two independent scorers. Pearson Product Correlations for writing as scored by Scorer 1 and writing scored by Scorer 2 was .97. The degree to which the two scorers agreed is indicated by Table 3, which shows the percentage of agreement between the two scorers.

Table 3
Percentage of Scorer Agreement for
Level C Benchmark Test: Writing

Amount of Difference Between the Two Scores	Percentage	Cumulative Percentage
Percentage of Scores Exactly the Same	57%	57%
Percentage of Scores Differing by One Point	31%	88%
Percentage of Scores Differing by Two Points	6%	95%
Percentage of Scores Differing by Three Points	4%	99%
Percentage of Scores Differing by More Than Three Points	1%	100%

The total score possible for the Writing test was 24 points. The degree to which the two scorers agreed was exceptionally high for a test with the large number of points a student could achieve. This

agreement is also supported by the very strong correlation between the two scores of .97.

There were a total of 12 multiple choice items on the Level C Reading test. An item analysis was computed for these 12 items. There were a total of 112 students who took the multiple choice test items. Since the number of items was very small and the sample size was relatively small it was not expected that a very high reliability index would result. The Kuder-Richardson reliability for this short test was .75 which is reasonably high for a test of this length. Table 4 below provides the item difficulty and item discrimination for each of the 12 test items. The difficulties of the items seem very reasonable as one would want the students to perform reasonably well if they were placed at this level by a Screener Test. The average item discrimination was .50 and the average difficulty for the 12 items was .48. This provides additional validity evidence for the overall test.

Table 4
Test Item Difficulties and Discriminations for the
12 Multiple Choice Test Items for Level C

Item Number	Difficulty (p value)	Discrimination Index
1	.51	.55
2	.48	.45
3	.59	.32
4	.64	.66
5	.38	.37
6	.61	.44
7	.52	.52
8	.46	.59
9	.29	.44
10	.46	.58
11	.40	.42
12	.44	.60
Average	.48	.50

LEVEL I BENCHMARK TEST – Reliability Results

The Level I Writing tests were scored by two independent scorers. Pearson Product Correlations for writing as scored by Scorer 1 and writing scored by Scorer 2 was .92. The degree to which the two

scorers agreed is indicated by Table 5 below, which shows the percentage of agreement between the two scorers.

The total score possible for the Writing test was 24 points. The degree to which the two scorers agreed was exceptionally high for a test with the large number of points a student could achieve. This agreement is also supported by the very strong correlation between the two scores of .92.

There were a total of 18 multiple choice items on the Level I Reading test. An item analysis was computed for these 18 items. There were a total of 98 students who took the multiple choice test items. Since the number of items was very small and the sample size was relatively small it was not expected that a very high reliability index would result. The Kuder-Richardson reliability for this short test was .64 which is reasonably high for a test of this length. Table 6 provides the item difficulty and item discrimination for each of the 18 test items. The difficulties of the items seem very reasonable as one would want the students to perform reasonably well if they were placed at this level by a Screener Test. The average item discrimination was .39 and the average difficulty for the 18 items was .78. This provides additional validity evidence for the overall test.

Table 5
Percentage of Scorer Agreement for
Level I Benchmark Test: Writing

Amount of Difference Between the Two Scores	Percentage	Cumulative Percentage
Percentage of Scores Exactly the Same	27%	27%
Percentage of Scores Differing by One Point	43%	69%
Percentage of Scores Differing by Two Points	18%	88%
Percentage of Scores Differing by Three Points	10%	99%
Percentage of Scores Differing by More Than Three Points	2%	100%

Table 6
Test Item Difficulties and Discriminations for the 18
Multiple Choice Test Items for Level I

Item Number	Difficulty (p value)	Discrimination Index
1	.91	.24
2	.89	.18
3	.86	.31
4	.85	.21
5	.83	.38
6	.86	.38
7	.98	.03
8	.83	.41
9	.54	.60
10	.85	.52
11	.57	.63
12	.83	.31
13	.85	.35
14	.65	.56
15	.53	.52
16	.63	.46
17	.80	.49
18	.85	.35
Average	.78	.39

LEVEL O BENCHMARK TEST – Reliability Results

The Level O Writing tests were scored by two independent scorers. Pearson Product Correlations for writing as scored by Scorer 1 and writing scored by Scorer 2 was .95. The degree to which the two scorers agreed can be found in Table 7, which shows the percentage of agreement between the two scorers.

The total score possible for the Writing test was 24 points. The degree to which the two scorers agreed was exceptionally high for a test with the large number of points a student could achieve. This agreement is also supported by the very strong correlation between the two scores of .95.

There were a total of 18 multiple choice items on the Level O Reading test. An item analysis was computed for these 18 items. There were a total of 46 students who took the multiple choice test items.

Table 7
Percentage of Scorer Agreement for
Level O Benchmark Test: Writing

Amount of Difference Between the Two Scores	Percentage	Cumulative Percentage
Percentage of Scores Exactly the Same	20%	20%
Percentage of Scores Differing by One Point	44%	63%
Percentage of Scores Differing by Two Points	20%	83%
Percentage of Scores Differing by Three Points	15%	98%
Percentage of Scores Differing by More Than Three Points	2%	100%

Since the number of items was very small and the sample size was relatively small it was not expected that a very high reliability index would result. The Kuder-Richardson reliability for this short test was .70 which is reasonably high for a test of this length. Table 8 provides the item difficulty and item discrimination for each of the 18 test items. The difficulties of the items seem very reasonable as one would want the students to perform reasonably well if they were placed at this level by a Screener Test. The average item discrimination was .41 and the average difficulty for the 18 items was .74. This provides additional validity evidence for the overall test.

Content Validity – Benchmark Tests

The study also examined the content validity of the Benchmark Tests using teacher comments regarding the content of the Benchmark Tests. Teachers were asked via survey to comment on the degree of appropriateness of the Benchmark Tests in relation to their classroom curriculum and practices. A rating of “Very Appropriate” indicates overall satisfaction with the content as it relates to instruction and level appropriateness. A rating of “Mostly Appropriate” suggests there were rare instances where test content seemed unrelated to classroom content or was not level appropriate. A rating of “Inappropriate” suggests that on more than a few occasions, content seemed unrelated to instruction and was not level appropriate. Overall, participating teachers regarded the content of the

Table 8
Test Item Difficulties and Discriminations for the 18
Multiple Choice Test Items for Level O

Item Number	Difficulty (p value)	Discrimination Index
1	.46	.50
2	.89	.39
3	.80	.42
4	.74	.48
5	.72	.28
6	.80	.55
7	.59	.39
8	.98	.12
9	.93	.53
10	.43	.30
11	.85	.37
12	.61	.49
13	.91	.45
14	.50	.51
15	.83	.51
16	.87	.37
17	.57	.39
18	.83	.30
Average	.74	.41

ELL-A Benchmark Tests as appropriate. It is likely that a single teacher administered more than one Level; thus, results are summarized among the three Levels in Table 9.

Table 9
Teacher Ratings of the Content of the
BENCHMARK TESTS

	Very Appropriate	Mostly Appropriate	Inappropriate
1. How appropriate was the content of the Reading section of the Benchmark Test?	67%	33%	
2. How appropriate was the content of the Writing section of the Benchmark Test?	56%	33%	11%

Additional Survey Results

Participating teachers were asked to respond to survey questions after administering the Benchmark Tests. This section provides a summary of those responses.

Surveys were returned by all but one of the participating teachers. Table 10 details the teachers' responses to questions regarding the Benchmark Tests. Table 11 summarizes teachers' written responses on both tests.

Table 10
Teacher Ratings of the BENCHMARK TESTS

Questions	Very Clear	Somewhat Clear	Confusing
1. How clear were the Benchmark Test directions for the students?	5	3	1
2. How clear were the Benchmark Test directions for the teachers?	6	3	
	15-20 minutes	20-25 minutes	More than 25 minutes
3. On average, how long did it take to administer the Benchmark Test?		2	7
	Very Comfortable	Somewhat Comfortable	Uncomfortable
4. Overall, how comfortable were you in administering the Benchmark Test?	6	3	
	Every Assignment	Some Assignments	Never/ A Different Tool
5. How often do your students use graphic organizers for writing?	3	5	
	Very Interested	Somewhat Interested	Uninterested
6. How interested did your students appear to be in the essay writing prompt?	2	5	2
	Agree	Undecided	Disagree
7. Describe the level to which you agree with the following statement: The Benchmark Test can be valuable to my classroom instruction.	5	2	2

Table 11
Teacher Comments on the BENCHMARK TESTS

Questions	Teacher Comments:
1. Please provide any additional feedback you have regarding the Benchmark Tests.	• More blends and diagraphs for initial, middle, and ending would make a better assessment.
	• The students did not know when to stop reading and what questions went with what story.
	• Younger students do better when they illustrate and discuss ideas before writing.
	• Put in teacher notes to write name and date on paper.
	• We have not used drawing in a procedure chart before. I like it though, and I am going to use it in the future.
	• In reading, too many wh_ questions and not enough for measuring other skills. I did personally like the writing assignment and the examples used for those. I like the use of the three-box chart, using it for three-step procedures, and using it for cause-and-effect. I think it helped my students do a better job of writing.
	• Picture content was useful and beneficial for student understanding.
	• Level O students really "got into" their prompt.
	• I would use the term "graphic organizer" because it is more common.
	• Younger children were lost. I ended up administering everything one-on-one.
	• Younger children are more likely to write based on experience.
	• The writing portion was confusing and frustrating for students.

Conclusions

The results of the various analyses demonstrated the reliability and validity of the ELL-A Benchmark Tests at Levels C, I, and O. Overall results suggest that the tests are working very well. Summarized results for each measurement used in the study are as follows:

- Data suggest that the level of agreement between scorers of the Writing subtest on Levels C, I, and O were well correlated.
- Data suggest that the multiple-choice items that comprise the Reading sub-test at Levels C, I, and O are of reasonable difficulty and discriminate very well.
- Teacher surveys of the content appropriateness of the benchmark tests across the three levels studied imply that the assessment has a high degree of content validity.

©2006 Harcourt Achieve
All rights reserved. Printed in the U.S.A.
5M/NY/SPEC/3-06
9994213253 IND
9994213261 10-PK