The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives

CSE Report 663

Jamal Abedi, Alison Bailey, Frances Butler, Martha Castellon-Wellington, Seth Leon, and James Mirocha

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education and Information Studies (GSEIS)
University of California, Los Angeles

2000/2005

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.4 Assessment of Language Minority Students—OBEMLA
Frances Butler, Jamal Abedi, and Alison Bailey, Project Directors

Copyright © 2005 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement/Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Office of Educational Research and Improvement/ Institute of Education Sciences, or the U.S. Department of Education.

ACKNOWLEDGMENTS

First, we wish to acknowledge the Office of Educational Research and Improvement (OERI)* and the Office of Bilingual Education and Minority Languages Affairs (OBEMLA) in Washington, DC, for recognizing the need to investigate issues of validity in assessment with the K-12 English language learner populations in the United States and for supporting our efforts. David Sweet at OERI and Delia Pompa and Milagros Lanauze at OBEMLA made important contributions to the conceptual framework of this research.

The work reported in this document would not have been possible without the support of colleagues in a number of educational venues. Each contributed in a special way to this effort. Our sincere appreciation to all.

Eva Baker, UCLA
Joan Herman, UCLA
Kris Gutierrez, UCLA
Reynaldo Macias, UCLA
Richard Durán, UCSB
Robin Stevens, UCLA
Roxanne Sylvester, UCLA
Ani Moughamian, UCLA
Katie Hutton, UCLA
Bobby Mendez, UCLA
Richard Maraschiello, Philadelphia
Mitchell Chester, Philadelphia
Sarah Gronna, Hawaii
Selvin Chin-Chance, Hawaii
Alan Ramos, Hawaii
Sandra Storey, Chicago
George Wimberly, Chicago
Arnold Goldstein, NCES
Shari Santapua, ETS

We extend a very special thank you to Heather Larson who provided incredible administrative support at every stage of our work. Her skill and efficiency facilitated our efforts in a major way, and we are grateful.

We also wish to thank Steven Acosta and Fred Moss for producing earlier drafts of this report. Their formatting and word-processing skills, not to mention their patience, are acknowledged with appreciation.

Finally, we thank Katharine Fry for sharing her editorial expertise whenever we asked and for preparing the final draft of this report.

* Now the Institute of Education Sciences.

CONTENTS

INTRODUCTION

The research effort reported here addresses the important national need for determining the validity of large-scale content assessments in English with students who are in the process of acquiring English as a second language. Often these students have been excluded from such assessments, but there have been recent, growing efforts to include them. There is, however, considerable variability nationwide in the inclusion process. The focus of this report is on second language students—English language learners (ELLs)—who have been included in large-scale content assessments regardless of their language ability. Within the context of assuring equal educational access for all students, technical issues around validity are being examined from three perspectives.

First, the potential impact of student background variables such as level of English proficiency and socioeconomic status (SES) on content-based assessment is examined through analyses of extant data from one large city school district (Site 1) and multiple school districts in one large state (Site 2); both sites have substantial ELL populations. Initial results from two other sites—Philadelphia and Hawaii—are reported in Abedi and Leon (1999)

Next, a school district in Southern California made available data from a controlled research environment which allowed comparison of student performance on a standardized achievement test with concurrent student performance on a language proficiency test of reading and writing. The results of the analyses from these data supplement what was learned from the earlier extant data analyses regarding ELL student performance on large-scale content assessments.

Finally, to help characterize the language demands of large-scale content assessments, the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has established evaluation criteria and developed a coding system for identifying language barriers in content tests. Analyses of the language of large-scale content assessments of reading comprehension, science, and math are reported. These data provide information on the potential role of language on test items across content areas.

Each of these three perspectives is covered in a separate chapter. In the final chapter, we discuss overall conclusions and recommendations.

CHAPTER 1

EXAMINING ELL AND NON-ELL STUDENT PERFORMANCE DIFFERENCES AND THEIR RELATIONSHIP TO BACKGROUND FACTORS: CONTINUED ANALYSES OF EXTANT DATA

Jamal Abedi, Seth Leon, and James Mirocha[1]

Summary

Data from a large public school district (referred to as Site 1 from this point on) for Grades 2 through 8 for the 1999 student population were analyzed for all students including English language learners (ELLs). The data included student responses to the reading and mathematics subtests of the Iowa Tests of Basic Skills[2] (ITBS) and student background data such as race, gender, birth date, and number of years of participation in a bilingual education program (number of years of bilingual service). Descriptive statistics and the percent of over-achievement of non-ELL students over ELL students were computed and compared across the different subtest content areas. In multiple regression analyses, student English learning status was related to student test scores and background variables.

A state department of education (referred to as Site 2 from this point on) provided us with student background data and item-level data on the Stanford Achievement Test Series, Ninth Edition (Stanford 9)[3] for all students in Grades 2 through 11 who were enrolled in the public schools statewide for the 1997-1998 academic year. Descriptive statistics compared ELL and non-ELL student performance by subgroup and across the different content areas. In a canonical correlation model the relationship between student language proficiency level, parent education, and family socioeconomic status (SES) (the Set 2 variables) and Stanford 9 performance (the Set 1 variables) was examined.

---

[1] The authors wish to thank Alison Bailey, Frances Butler, Richard Durán, Joan Herman, Milagros Lanauze, and David Sweet for their thoughtful comments and suggestions on earlier versions of this chapter.

[2] Hoover, H. D., Hieronymus, A .N., Dunbar, S. B., & Frisbie, D. A. (1996). *Iowa Tests of Basic Skills, Form M.* Chicago, IL: Riverside Publishing.

[3] Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Series. Ninth edition, Form T.* San Antonio, TX: Harcourt Brace.

The results of our analyses of data from Site 1 and Site 2 were consistent with our earlier analyses from other sites and indicated the following:

1. Student language proficiency level is associated with performance on content-based assessments.

2. There is a gap between the performance of English language learners (ELLs) and their native English speaking peers (non-ELLs).

3. The gap between ELL and non-ELL students increases as the language load of the assessment tools increases.

The term "language load" in this report refers to linguistic complexity of the test items. In her language analysis of standardized achievement tests, Bailey (2000/2005) used the term "language demand" and indicated that the language demand of standardized achievement tests could be a potential threat to the validity of these tests when administered to English language learners. Because of this source of threat, she added, these assessments may not present an accurate picture of ELL student content knowledge. Bailey elaborated on the concept of language demand as uncommon vocabulary, nonliteral usage (idioms), complex or atypical syntactic structure, uncommon genre, or multi-clausal processing. For this part of the study, we did not perform any linguistic analyses of test items. We may do so in our next phases of research. However, test items in some content areas use more language than other content areas. For example, it is obvious that in reading assessments, there is more language involved than in assessments for content-based areas such as math and science.

Perspective

In a previous report (Abedi & Leon, 1999), we discussed the results of analyses that were performed on data from several different locations. The analyses reported earlier included descriptive statistics by English proficiency level, analyses of internal consistency of the test items by English proficiency level, and analyses comparing the structural relationship of the instruments across various English proficiency categories. Results of these analyses indicated that English language learner (ELL) students generally perform lower than non-ELL students in reading, science, math and other content areas—a strong indication of the relationship of English proficiency with achievement assessment. However, the level of impact[4] of

[4] By using the term "impact" we do not mean any causal relationships.

language on assessment performance of ELL students is greater in those content areas with high language load. For example, analyses showed that ELL and non-ELL students have the greatest performance differences in reading. The gap between the performance of ELL and non-ELL students becomes smaller in other content areas where there is less language load. The difference between ELL and non-ELL student performance becomes smallest in math, particularly on math items where language has less impact, such as on math computation items.

The results of our analyses also indicated subtest internal consistency reliabilities were lower among ELL students, particularly in the Limited English Proficient (LEP) group, than among non-ELL students. That is, the language background of students may add another dimension to performance assessment, a language dimension, wherein language might be a source of measurement error.

Analyses of the structural relationships between individual items and between items with the total test scores showed a major difference between ELL and non-ELL students. Structural equation models for ELL students demonstrated lower statistical fit compared with models for non-ELL students. Further, the factor loadings were generally lower for ELL students, and the correlations between the latent content-based variables were weaker for ELL students.

We obtained data from several other locations nationwide. In analyzing the new data sets we have continued our efforts to add to our knowledge and to enable us to respond to the main question of this study: How does student language background impact performance on standardized achievement tests? The following sections are summaries of our analyses of the new data from Site 1 and Site 2.

Public School District, Site 1

Data from Site 1, the public school district, for Grades 2 through 8 for the 1999 student population were analyzed. Similar data for the previous years, 1990 to 1998, will be obtained and analyzed. The 1999 data included student responses to ITBS test items, ITBS subsection scores, and student background data. The background data included student ID number, race, birth date, gender, and the number of years of participation in a bilingual education program (number of years of bilingual service). Other school- or test-related variables such as school unit number, grade, test form and test level were also included in the data files. Three forms of the ITBS were used in the 1999 Site 1 testing, Forms K, L, and M. This report focuses on Form M, which was taken by 98.6% of the students. Data were provided for Levels 7

through 14 of the ITBS. Each test level was given to students from various grades. However, each test level was associated primarily with a particular grade, as follows: Level 8 with Grade 2, Level 9 with Grade 3, Level 10 with Grade 4, Level 11 with Grade 5, Level 12 with Grade 6, Level 13 with Grade 7, and Level 14 with Grade 8. This report follows the primary association just described; for example, ITBS scores from grades other than Grade 8 were not analyzed for Level 14.

Data files from Site 1 did not include student ELL status. However, the files included the number of years of bilingual service. As a proxy for ELL status, we created a Bilingual Status variable from the years of bilingual service as follows: a student with one or more years of bilingual service was designated "Bilingual" and a student with no years of bilingual service was designated "Non-Bilingual." Thus Bilingual Status serves as a proxy for ELL status. We also used another variable as a proxy for ELL status based on the number of years in bilingual education. Since participation in more bilingual classes may increase students' level of language proficiency, students with less than 4 years of bilingual education were categorized as ELL and those with 4 or more years of bilingual education as non-ELL. However, the results of our analyses indicated that the mean score for students with more years in bilingual classes was significantly lower than the mean for students with fewer years in bilingual classes. We therefore decided to use the categorization based on receiving or not receiving bilingual education.

ITBS subsection (subtest) scores were reported in the following forms: (1) raw scores, (2) percentile ranks, (3) normal curve equivalent (NCE) scores, (4) stanine scores, and (5) grade equivalent scores. For Grades 3 through 8, scores were available at the subsection level for math concepts and estimation, math problem solving and data interpretation, math computation, and reading. For Grade 2, the ITBS subsections were math concepts, math problem solving, math computation, and reading (math estimation and data interpretation were not included in the Grade 2 level of the ITBS).

Among the different subsection scores, we decided to analyze and report the normal curve equivalent (NCE) scores.[5] The basis for this decision was consistency with the reports of data from the other sites (see Abedi & Leon, 1999). Some of the

[5] NCE scores are normalized standard scores with a mean of 50 and a standard deviation of 21.06. Because of their distributional properties, for analysis purposes NCEs are preferred over National Percentile ranks or raw scores. NCEs coincide with National Percentile ranks at the values 1, 50, and 99.

math scores were composites of more than one subsection score. For example, the total score of *math concepts and estimation* was a composite of two subtests, the *math concepts* subtest and the *math estimation* subtest. Similarly, the *math problem solving and data interpretation* score was a composite of the *problem solving* and *data interpretation* scores. Thus, there were originally five subsections in the math test. We report the descriptive statistics for the three subsections (*math concepts and estimation, math problem solving and data interpretation,* and *math computation*) but discuss the test item characteristics and internal consistency coefficients for the five math subtests separately.

Table 1.1 presents the means, standard deviations and number of students with non-missing NCE scores for the ITBS subsections at the various grade and test level combinations. Because of several validity concerns, including issues about the representativeness of the Grade 2 data, results of analyses for Grade 2 were not included in the main part of the report and are provided in the Appendix.

As the results in Table 1.1 show, bilingual students generally performed lower than their non-bilingual peers. For the native English speakers (non-bilinguals) the overall mean NCE subsection score was 46.25 and ranged from 37.84 to 55.99, whereas for the bilingual students the mean score was 37.59 and ranged from 29.65 to 52.37. However, the gap between the test scores of bilingual and non-bilingual students depends on the grade level and the content of the assessment. The difference between the mean NCE scores of bilingual and non-bilingual students was generally small for Grade 3 students, except in reading (where there was about a 7-point difference), and favored the non-bilingual group, except in math computation, where the mean was slightly higher for the bilingual group. Beginning with Grade 4, all the differences favor the non-bilingual group and generally become larger as we move to higher grades. For example, the mean NCE math concepts and estimation score for Grade 3 non-bilingual students was 44.14 versus 41.93 for bilingual students—a small difference (about 2.5 score points higher for the non-bilingual group). In Grade 3 reading, the non-bilingual students obtained a substantially higher mean (M = 37.84, SD = 17.93) than the bilingual students (M = 30.67, SD = 17.07), a gap of approximately one third of a standard deviation. In Grade 4, the reading gap becomes even larger. The mean reading score for Grade 4 non-bilingual students was 45.38 (SD = 15.68), compared with a bilingual student mean of 34.87 (SD = 12.78), a gap of more than two thirds of a standard deviation.

Table 1.1

Mean, Standard Deviation, and Number of Students for ITBS Subsection Scores at the Different Grade/Level Combinations (NCE Scores)

Test level	Grade	Bilingual status	Math concepts & estimation	Math problem solving & data interpretation	Math computation	Reading
9	3	Non-Bilingual				
		M	44.14	40.41	50.10	37. 84
		SD	20.10	21.48	23.86	17.93
		N	29,244	29,206	29,251	29,254
		Bilingual				
		M	41.93	36.37	51.72	30.67
		SD	19.10	20.52	23.22	17.07
		N	7,415	7,421	7,427	7,428
10	4	Non-Bilingual				
		M	44.12	45.42	55.99	45.38
		SD	20.37	17.74	24.11	15.68
		N	25,310	25,303	25,317	25,309
		Bilingual				
		M	34.77	38.07	52.37	34.87
		SD	18.74	15.79	23.83	12.78
		N	5,407	5,401	5,406	5,402
11	5	Non-Bilingual				
		M	44.99	45.81	52.28	46.60
		SD	19.93	17.31	21.33	14.31
		N	22,270	22,256	22,269	22,254
		Bilingual				
		M	32.96	34.52	46.41	33.02
		SD	17.25	15.93	20.28	12.52
		N	3,980	3,978	3,978	3,974
12	6	Non-Bilingual				
		M	45.22	43.90	50.83	42.61
		SD	20.53	18.53	21.02	16.13
		N	25,372	25,352	25,361	25,380
		Bilingual				
		M	35.47	33.54	45.47	29.65
		SD	17.66	14.32	18.42	12.54
		N	3,453	3,450	3,452	3,445
13	7	Non-Bilingual				
		M	41.76	45.07	49.72	46.56
		SD	21.23	17.00	17.58	15.64
		N	23,957	23,941	23,935	23,979
		Bilingual				
		M	29.95	33.94	44.01	33.35
		SD	17.93	15.00	16.15	11.43
		N	2,392	2,391	2,391	2,395
14	8	Non-Bilingual				
		M	48.25	47.41	49.11	46.52
		SD	19.27	15.95	16.39	15.17
		N	23,541	23,539	23,545	23,577
		Bilingual				
		M	36.98	35.99	43.51	32.60
		SD	16.02	13.55	14.77	12.54
		N	2,371	2,371	2,374	2,362

The trend of increasing performance gaps between bilingual and non-bilingual students varies across the content/subsection areas. The largest gap between the two groups was in reading. This result was expected because the reading test items have presumably the highest language load among the four content areas presented in Table 1.1. Among these four content areas, the math computation subsection appears to have the lowest language load. Accordingly, the performance gap between bilingual students and non-bilingual students was the lowest on the math computation subsection. To compare bilingual and non-bilingual group score differences across test level, grade, and content area, the percentage of over-achievement (POA) of non-bilingual students over bilingual students was computed by subtracting the bilingual subtest mean from the non-bilingual subtest mean, dividing the difference by the bilingual subtest mean, and multiplying the result by 100. The result gives the percentage by which the non-bilingual group mean exceeds the bilingual group mean on the particular subtest. A negative POA indicates that the bilingual mean exceeds the non-bilingual mean.

Table 1.2 presents the POAs of the non-bilingual students compared with the bilingual group by test level, grade, and content area. The results in Table 1.2 present several interesting patterns:

1. Except for Grade 3 (Level 9) math computation, the over-achievement percentages are all positive, indicating that on the average, the non-bilingual students outperformed the bilingual students.

2. Major differences between bilingual and non-bilingual students were found for students in Grades 3 and above. The difference between the mean scores of bilingual and non-bilingual students increased sharply by grade, up to Grade 6. Starting with Grade 6, the percent of over-achievement was still positive, but the rate of increase slowed down. For example, in Grade 3 non-bilingual students had over-achievement percentages of 5.3% in math concepts and estimation, 11.1% in math problem solving and data interpretation, –3.1% in math computation (the bilingual group did better than the non-bilingual group on this subtest), and 23.4% in reading. In Grade 4 these percentages increased to 26.9% for math concepts and estimation, 19.3% for math problem solving and data interpretation, 6.9% for math computation, and 30.1% for reading. The percentages further increased in Grade 5 to 36.5% for math concepts and estimation, 32.7% for math problem solving and data interpretation, 12.6% for math computation and 41.1% for reading.

3. As indicated earlier, the largest gap between bilingual students and non-bilingual students is in reading. The next largest gaps are in the content areas that appear to have more language load. For example, the math

Table 1.2

Percentage of Over-Achievement of Non-Bilingual Over Bilingual Students on Reading and Math Subsections

Test level	Primary grade	Math concepts & estimation	Math problem solving & data interpretation	Math computation	Reading
9	3	5.3	11.1	–3.1	23.4
10	4	26.9	19.3	6.9	30.1
11	5	36.5	32.7	12.6	41.1
12	6	27.5	30.9	11.8	43.7
13	7	39.4	32.7	12.9	39.6
14	8	30.5	31.7	12.9	42.7
Average of all levels/ grades		27.7	26.4	9.0	36.8

concepts and estimation and the math problem solving and data interpretation subsections seem to have higher language load than the math computation subsection. Correspondingly, the over-achievement percentages are higher for math concepts and estimation and for problem solving and data interpretation. The average over-achievement percentage for Grades 3 through 8 is 27.7% for math concepts and estimation. That is, the non-bilingual group average in math concepts and estimation was 27.7% higher than the bilingual group average. A similar trend was observed in math problem solving and data interpretation; the average over-achievement for this subsection was 26.4%. The average over-achievement percentage for math computation, however, was 9.0%, which is substantially lower than the corresponding over-achievement percentages for the other two math subsections. The smaller gap between bilingual and non-bilingual students on the math computation subsection might be attributable to the lower language load of the math computation subsection.

Internal Consistency of Test Items by Student Language Status

Earlier in this chapter, based on the analyses of data from other sites, we suggested that the language load of the test items might introduce a bias into the assessment. That is, a language factor may act as a source of measurement error in the assessment of English language learners. To examine the hypothesis of the impact of language on assessment, we performed a principal components analysis on the test item-level data and computed internal consistency coefficients (coefficient alpha) by student bilingual status (for issues concerning factoring phi-coefficient, see Abedi, 1997). Because a different test level was used for each grade, these analyses were performed separately for each grade. Within each grade, we

conducted the internal consistency analyses separately for bilingual and non-bilingual students, so that we could compare the subtest internal consistencies for the bilingual and non-bilingual groups.

Table 1.3 summarizes the results of the principal components and internal consistency analyses for math (problem solving, concepts, estimation, data interpretation, and computation subsections) and reading. For each of the six subsections, more than one component with eigenvalue greater than 1 was extracted. Across the subsections, the number of components (factors) with eigenvalue greater than 1 ranged from two to eight. The percent of common variance explained by the first component was below 26% of the total item variance for each subsection at each grade. If the items in a subtest were all measuring the same construct, then we would have expected a higher proportion of common variance for the first principal component. These results may suggest low internal consistency among the test items in the math and reading subsections, particularly with the bilingual subgroup.

To examine the pattern of item internal consistency among bilingual and non-bilingual students, we computed coefficient alpha separately for the two groups of students. As the results in Table 1.3 show, the item responses of bilingual students in general have lower internal consistency. The gap between the internal consistency coefficients of the two groups varied across grade and subsection. Consistent with our findings reported earlier in this chapter, the differences between the bilingual and non-bilingual groups are small for Grade 3 students. For higher grades, this gap increases. For example, in Grade 3, the average alpha coefficient (across the six subtests) for bilingual students is .74 and for non-bilingual students the average is .76. In Grade 6, the average for bilingual students is .71 and for non-bilingual students is .84. In Grade 8, the average for bilingual students is .74 and for non-bilingual students is .83. This trend may occur because the test items for Grade 3 may be less linguistically complex than the items for the higher grades.

It is also clear from the results in Table 1.3 that the gap between internal consistency (alpha) coefficients for bilingual and non-bilingual students varies across the content areas. Internal consistency coefficients for subsections with more language load are substantially lower for bilingual students. For example, on the reading subsection in Grades 6 and 8 the average alpha for bilingual students is .68, compared with an average alpha of .88 for non-bilingual students. However, on the math computation subsection, where there is possibly less language load, there is a

Table 1.3

Summary Results of Principal Components and Reliability Analyses

Subsection/Grade	Number of components Eigenvalue > 1	Percent of variance of 1st component	Reliability (α) bilingual	Reliability (α) non-bilingual
Math problem solving				
Grade 3	2	22.88	.74	.70
Grade 6	2	20.68	.64	.77
Grade 8	3	16.84	.60	.71
Math concepts				
Grade 3	2	17.43	.72	.74
Grade 6	4	17.01	.66	.82
Grade 8	4	16.49	.75	.83
Math estimation				
Grade 3	2	24.99	.69	.70
Grade 6	3	17.89	.65	.73
Grade 8	5	13.80	.63	.68
Math data interpretation				
Grade 3	2	25.25	.60	.66
Grade 6	2	20.16	.51	.69
Grade 8	3	15.86	.48	.64
Math computation				
Grade 3	5	23.31	.89	.90
Grade 6	7	20.91	.87	.90
Grade 8	7	20.25	.88	.90
Reading				
Grade 3	6	16.77	.82	.85
Grade 6	5	16.64	.65	.88
Grade 8	9	14.67	.72	.87

correspondingly smaller difference between the alphas for bilingual students (.88) and non-bilingual students (.90).

Figure 1.1 compares the internal consistency coefficients for bilingual and non-bilingual students across the six different content areas for Grade 3 students. As Figure 1.1 shows, the differences between the bilingual and non-bilingual alphas are very small and in some cases nonexistent. However, the alpha coefficients for the math subsections are generally lower than the alphas for reading. This may be explained by the differences in the number of items for the different subsections. The reading subsection had the largest number of items.

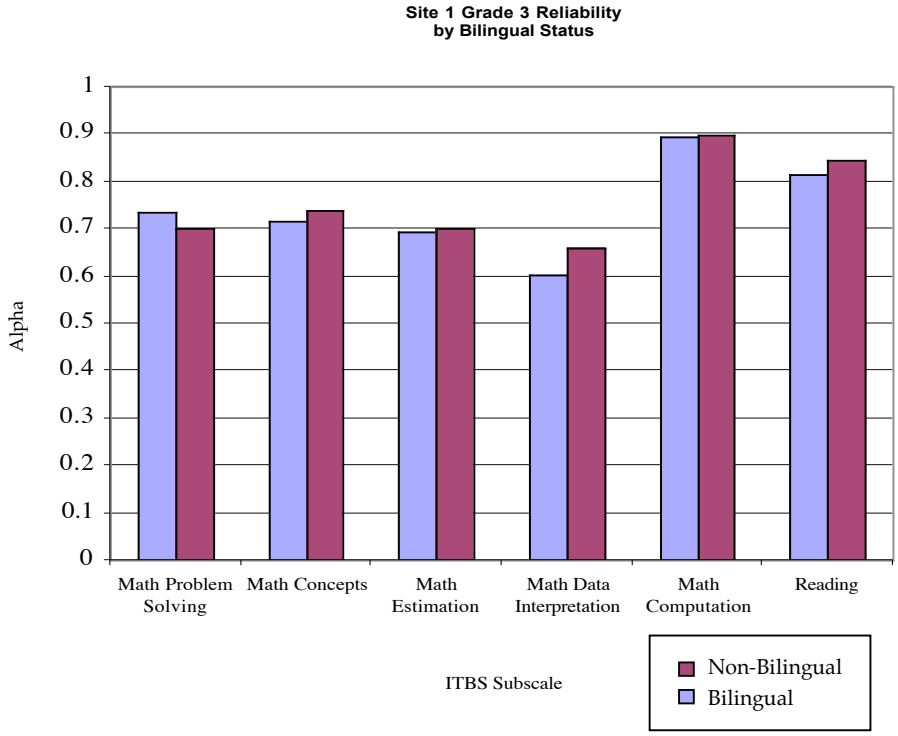


Figure 1.1. Site 1 Grade 3 reliability alpha coefficients.

Figures 1.2 and 1.3 present the same results for students in Grades 6 and 8 respectively. As indicated earlier, the differences in alpha coefficients between bilingual and non-bilingual students in Grades 6 and 8 are substantially larger than the differences in Grade 3. As Figures 1.2 and 1.3 suggest, the largest differences between bilingual and non-bilingual students occur in reading, where the language load is greatest. In math computation, where the language load is smallest, the alpha differences are also the smallest.

The lower reliability (internal consistency) may have been caused by restriction of range in the bilingual population. It is plausible that the restriction of range in the bilingual group is an effect of language and other factors such as family socioeconomic status (SES) and opportunity to learn (OTL). We use the Grade 8 reading and math computation subtests to illustrate the possible impact of restriction of range. In the high language demand reading content area, there is a large difference in the reliabilities for the bilingual and non-bilingual groups, with

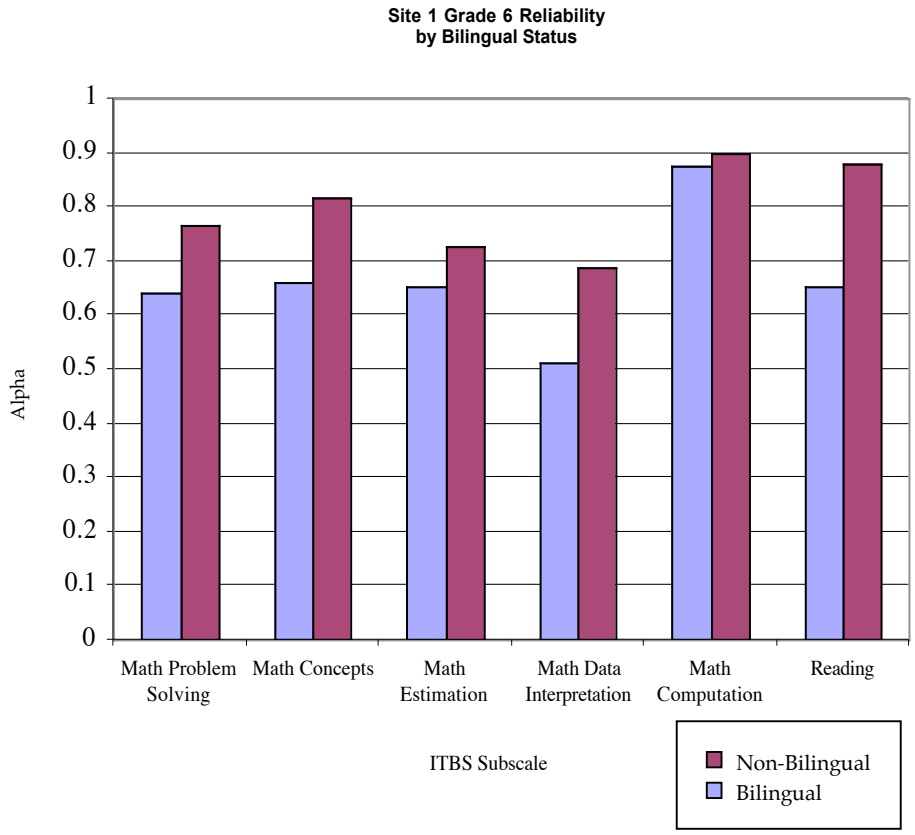


Figure 1.2. Site 1 Grade 6 reliability alpha coefficients.

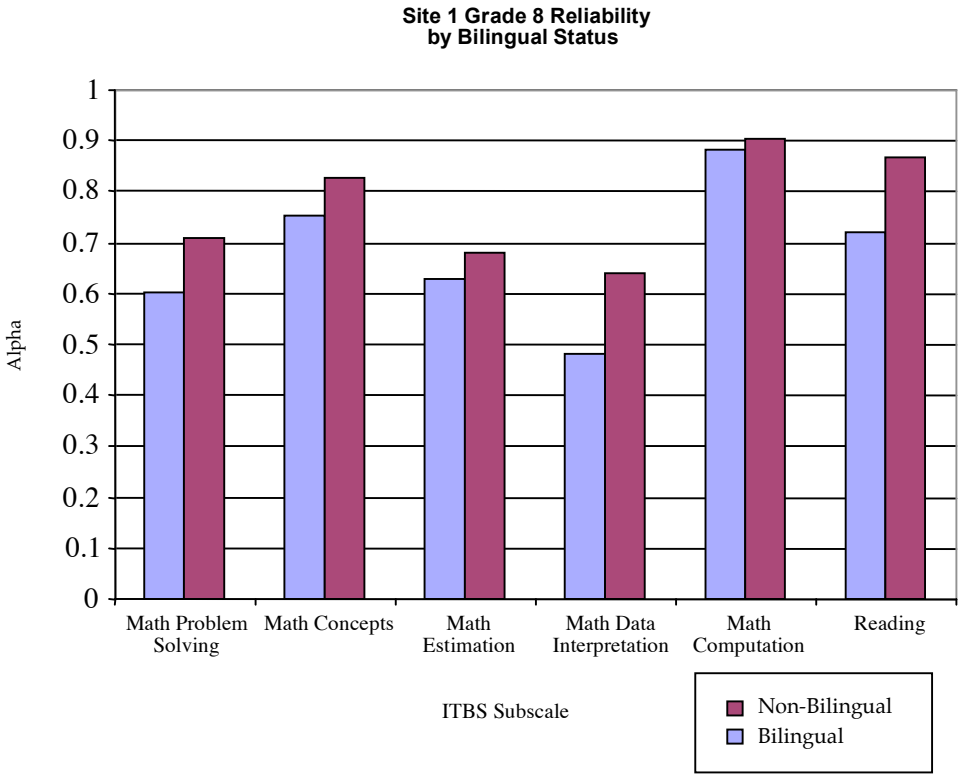
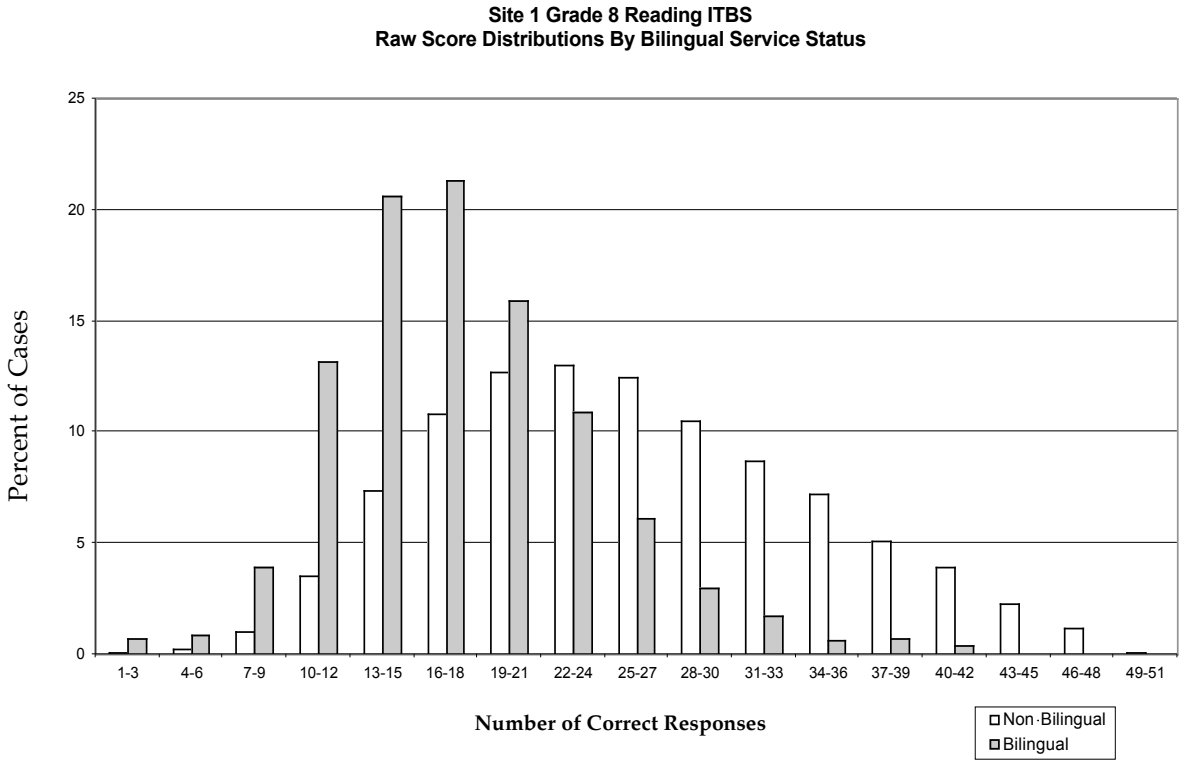


Figure 1.3. Site 1 Grade 8 reliability alpha coefficients.

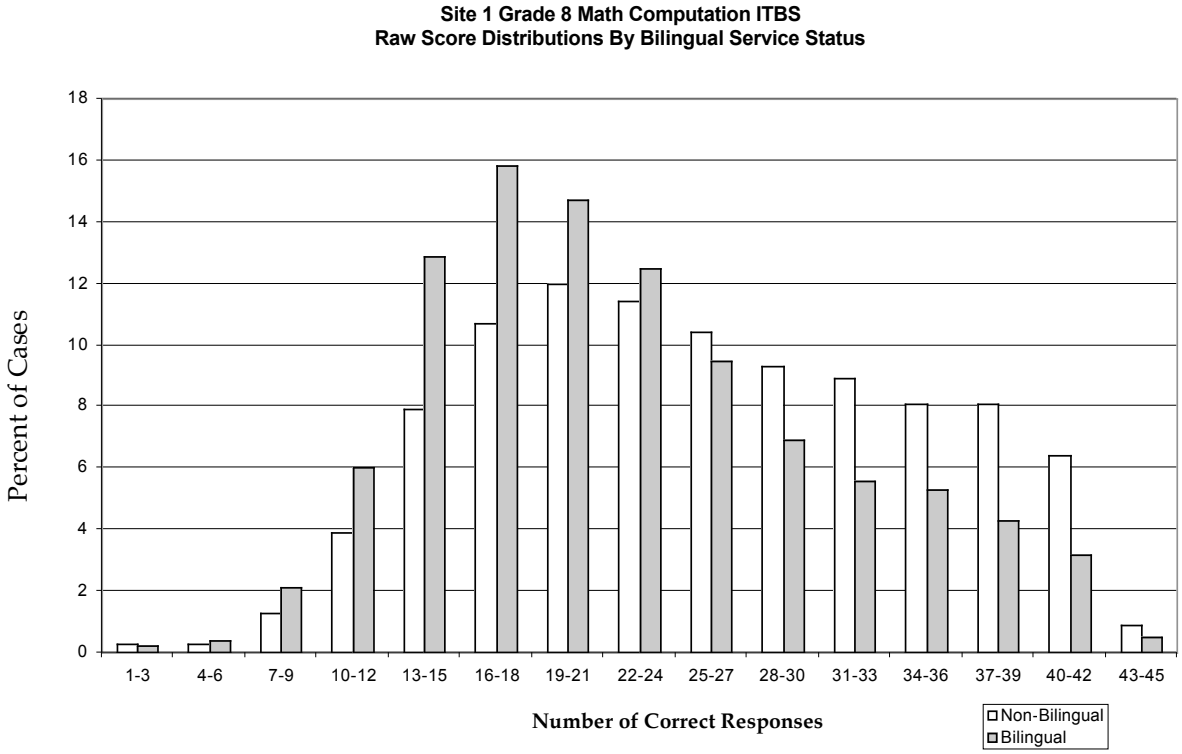
alphas of .722 and .869 respectively. There is also a large difference in the reading raw score[6] variances for the two groups, 32.73 and 62.04, resulting in a significant restriction of range in the bilingual group. Figure 1.4 shows the bilingual and non-bilingual reading raw score distributions for the two groups along with the variances and alpha reliabilities. The bilingual distribution has less spread and is centered lower than the non-bilingual distribution. In stark contrast, in the low language demand math computation area, there is a small difference in the internal consistency reliabilities for the two groups and the raw score variances are similar in magnitude. Figure 1.5 shows the Math Computation distributions, variances, and alphas for the two groups. The distributions are quite similar for the two groups.



ELL/SES status	Mean	Variance	Cronbach alpha
Non-bilingual	25.60	75.14	.869
Bilingual	17.76	36.65	.722

Figure 1.4. Site 1 Grade 8 reading score distributions and reliability.

[6] Here we use raw scores rather than NCEs because Cronbach's alpha utilizes raw score variance.



ELL/SES status	Mean	Variance	Cronbach alpha
Non-bilingual	25.80	79.24	.904
Bilingual	22.49	69.09	.884

Figure 1.5. Site 1 Grade 8 math computation distributions and reliability.

We believe that language (and perhaps other factors such as SES and OTL) causes a restricted range distribution, a distribution of scores with lower variability, and that this in turn causes lower internal consistency.

Because the number of items varied across the subsections, the internal consistency coefficients may have been affected by the number of items. To control for differences in alpha due to differences in the number of items, we adjusted the internal consistency coefficients by the number of items. The subsection with the maximum number of items was the reading subsection for Grade 8, with 49 items. We thus adjusted the alpha coefficients to reflect a constant length of 49 items for each subsection. Table 1.4 presents the unadjusted and adjusted alpha coefficients. As can be seen from the results in Table 1.4, the internal consistency coefficients increased substantially in some cases. However, the general trend of lower internal consistency coefficients for the bilingual students remained.

Table 1.4

Internal Consistency Coefficients Adjusted by the Number of Items

	Unadjusted		Adjusted	
Subsection/Grade	Reliability (α) bilingual	Reliability (α) non-bilingual	Reliability (α) bilingual	Reliability (α) non-bilingual
Math problem solving				
Grade 3 (14 items)	.74	.70	.91	.89
Grade 6 (18 items)	.64	.77	.83	.90
Grade 8 (20 items)	.60	.71	.79	.86
Math concepts				
Grade 3 (20 items)	.72	.74	.86	.87
Grade 6 (28 items)	.66	.82	.77	.89
Grade 8 (32 items)	.75	.83	.82	.88
Math estimation				
Grade 3 (12 items)	.69	.70	.90	.91
Grade 6 (20 items)	.65	.73	.82	.87
Grade 8 (24 items)	.63	.68	.78	.81
Math data interpretation				
Grade 3 (10 items)	.60	.66	.88	.91
Grade 6 (14 items)	.51	.69	.79	.89
Grade 8 (16 items)	.48	.64	.74	.84
Math computation				
Grade 3 (34 items)	.89	.90	.92	.93
Grade 6 (41 items)	.87	.89	.89	.91
Grade 8 (43 items)	.88	.90	.90	.91
Reading				
Grade 3 (36 items)	.82	.85	.86	.88
Grade 6 (44 items)	.65	.88	.68	.89
Grade 8 (49 items)	.72	.87	.72	.87

The results presented so far demonstrate that bilingual students do not perform as well as non-bilingual students, especially in content areas with higher language load. Results of analyses on individual test items are consistent with this general trend. That is, in most of the cases, item scores for the bilingual students are lower than item scores for the non-bilingual students. However, the item-level differences between bilingual and non-bilingual students vary greatly across the items. Some of the test items are more difficult for bilingual students than other items and items may function differently with different groups. We speculated that items with more complex language would be more difficult for bilingual students, regardless of the level of content difficulty.

We computed the difference between the mean score for each individual item across the bilingual categories (bilingual/non-bilingual) by subtracting the mean score for bilingual students from the mean score for non-bilingual students. We called this difference "DBN" (Difference between Bilingual and Non-bilingual student performance). Because all ITBS items were in multiple-choice format, the DBN was the difference between the proportion of correct responses for bilingual and non-bilingual students. A negative DBN indicates that bilingual students had higher performance than their non-bilingual peers on that particular item. Due to space limitations, we did not include the results of this analysis in our report. However, we summarize the results of item-level comparisons in Table 1.5. We rank ordered the items based on the magnitude of DBN. In Table 1.5, we present the minimum, maximum, and average DBN for each ITBS subsection.

Table 1.5

Item-Level Response Differences Between Bilingual and Non-Bilingual Students (DBN)

Subsection/Grade	No. of items	Minimum	Maximum	Average DBN
Math problem solving				
Grade 3	14	.01	.07	.04
Grade 6	18	.01	.19	.12
Grade 8	20	.03	.26	.12
Math concepts				
Grade 3	20	–.02	.08	.01
Grade 6	28	–.01	.25	.09
Grade 8	32	–.01	.21	.12
Math estimation				
Grade 3	12	.00	.03	.01
Grade 6	20	.00	.14	.09
Grade 8	24	–.02	.16	.08
Math data interpretation				
Grade 3	10	–.03	.09	.04
Grade 6	14	.01	.25	.08
Grade 8	16	.05	.28	.11
Math computation				
Grade 3	34	–.05	.01	–.02
Grade 6	41	–.02	.15	.04
Grade 8	43	.01	.17	.07
Reading				
Grade 3	36	–.04	.17	.08
Grade 6	44	.02	.29	.15
Grade 8	49	.03	.38	.15

As the results in Table 1.5 indicate, the range and average of the DBN differ across the grade levels and content areas. For Grade 3 students, the average DBN is small on all subtests except reading; the average DBN is negative in the math computation subtest, indicating that bilingual students performed slightly better than non-bilingual students on math computation. This is consistent with our earlier Grade 3 findings indicating that there is not much of a gap between the performance of bilingual and non-bilingual students. For Grades 6 and 8, the DBN is larger in those content areas with more language load. For example, in Grade 6 reading, there is a maximum difference of .29 between the proportion of correct responses of bilingual and non-bilingual students. This maximum difference increases to .38 for students in Grade 8 reading.

As expected, on items with less language load, the size of the DBN is substantially smaller than the DBNs presented for the reading subsection. For example, on the math computation subsection the maximum DBNs for Grades 6 and 8 are .15 and .17, whereas on the reading subsection the maximum differences for Grades 6 and 8 are .29 and .38 respectively.

Results of Regression Analyses

To investigate the strength of the relationships among bilingual status and test scores, various regression models were explored**.** Student bilingual status (bilingual/non-bilingual) was used as a dependent variable in a regression model in which test scores (math concepts and estimation, math problem solving and data interpretation, math computation, and reading), gender, and ethnicity were used as independent variables. To present a clearer picture of the association of ethnicity (a categorical variable with five categories) and bilingual status, we used criterion-scaling multiple regression methodology (see Pedhazur, 1997, pp. 501-505). Rather than creating k – 1 dummy variables for the ethnic categories (where k is the number of categories), we used the ethnic group averages in one single variable called "ethnicity." Thus, in the criterion-scaling regression model, each individual's value on the variable "ethnicity" is the mean score of the particular ethnic group of which the individual is a member. Because the math subsection NCE scores were highly correlated, to avoid the multi-collinearity problem we used the math total NCE score instead of the math subsection scores.

A separate multiple regression analysis was conducted for each of the three grades (Grades 3, 6, and 8). Table 1.6 summarizes the results of multiple regression analyses for students in these grades.

Table 1.6

Results of Multiple Regression Analyses for Grades 3, 6, and 8

Variable	*B*	*SE B*	ß	*t*	Sig *t*
		Grade 3			
Math total	.0005	.0001	.025	4.479	<.0005
Reading	–.0039	.0001	–.173	–30.851	<.0005
Gender	.0144	.0030	.018	4.431	<.0005
Ethnicity	.9940	.0060	.623	153.350	<.0005
Constant	.1010	.0060			
$R = 0.647$			$R^2 = 0.418$		
		Grade 6			
Math total	–.0006	.0001	–.036	–5.120	<.0005
Reading	–.0047	.0001	–.237	–33.730	<.0005
Gender	.0006	.0030	.001	.175	.8610
Ethnicity	1.0130	.0110	.453	88.150	<.0005
Constant	.2160	.0070			
$R = 0.518$			$R^2 = 0.268$		
		Grade 8			
Math total	–.0008	.0001	–.046	–5.94	<.0005
Reading	–.0043	.0001	–.233	–29.99	<.0005
Gender	.0073	.0030	.013	2.26	.0240
Ethnicity	1.0140	.0160	.365	64.70	<.0005
Constant	.2200	.0070			
$R = 0.447$			$R^2 = 0.200$		

As the data in Table 1.6 suggest, the results of multiple regression analyses are consistent across the three grades and indicate that test scores and ethnicity are powerful predictors of student bilingual status. The multiple *R* for the Grade 3 regression model is .647 and R^2 for this model is .418, indicating that about 42% of the variance of student bilingual status can be explained by math and reading test scores, ethnicity, and gender. In this model, all predictors had a significant contribution to the prediction. Among the predictors, ethnicity (the criterion-scaled variable) had the highest level of contribution to the prediction. The *t* ratios for testing the significance of prediction were significant above and beyond the .01 nominal level for all four predictor variables. Once again, the β coefficients suggest that ethnicity was the strongest predictor of student bilingual status. For the math and reading variables, reading ($\beta = -.173$) had a higher level of contribution to the prediction of bilingual status than math ($\beta = .025$).

As indicated earlier, the results of the multiple regression analyses are consistent across the three grades. All three models suggest that ethnicity is the strongest predictor of bilingual status, with the highest magnitude of β. The next strongest predictor is reading, followed by math. One difference among the results for the different grades is that the strength of association decreases in the higher grades. R^2 for the Grade 3 model is .418 (42% of the variance of bilingual status is explained). For Grade 6, R^2 is .268 (27% of the variance of bilingual status is explained), and for Grade 8, R^2 is .200 (20% of the variance of bilingual status is explained). Another difference is that in Grade 6, gender is not a significant predictor of bilingual status, whereas gender is significant in Grades 3 and 8. However, the gender differences are so small as to be not meaningful. Finally, the directionality of math as a predictor of bilingual status is reversed in Grade 3, where higher math totals are associated with bilingual membership. However, the math "effect" in all three grades is quite small in comparison to the "effects" of reading and ethnicity.

Statewide School Districts, Site 2

The Site 2 Department of Education gave us access to the Stanford 9 test data for all students in Grades 2 through 11 who were enrolled in the public schools statewide for the 1997-1998 academic year. The 1997-98 data included student responses to Stanford 9 test items (item-level data), subsection scores, and student background data. The background data included student ID number, gender, ethnicity, free/reduced-price lunch participation, parent education, student ELL status, Students with Disabilities (SD) status, home language survey results, and district mobility data. Stanford 9 subsection scores were reported as (a) raw scores, (b) percentile ranks, and (c) normal curve equivalent (NCE) scores. Scores were available at the subsection level for reading, math, language, spelling, science, and social science. Some of these subsection scores were not available for all grades. NCE scores were used in our analyses for the purpose of consistency with the other sites (see Abedi & Leon, 1999).

Tables 1.7 and 1.8 present the number of students in Grades 2, 7 and 9 who took the Stanford 9 tests, by student ELL and SD status. Table 1.7 includes information for students with non-missing scores on the Stanford 9 reading, math, and language subsections. Table 1.8 presents similar results for students with non-missing scores on the spelling, science, and social science subsections.

Table 1.7

Stanford 9 Reading, Math, and Language Frequencies for Students in Grades 2, 7, and 9, Site 2 Statewide School Districts

			Students with a normal curve equivalent score					
	All students		Reading		Math		Language	
	No.	%	No.	%	No.	%	No.	%
Grade 2								
SD only	17,506	4.2	15,051	4.1	16,720	4.2	16,076	4.1
LEP only	120,480	29.1	97,862	26.5	114,519	28.4	107861	27.5
LEP and SD	4,629	1.1	3,537	1.0	4,221	1.0	3,891	1.0
Non-LEP/Non-SD	271,554	65.6	252,696	68.5	267,397	66.4	263,955	67.4
All students	414,169	100.0	369,146	100.0	402,857	100.0	391,783	100.0
Grade 7								
SD only	24,683	7.1	22,388	6.7	23,029	6.8	22,264	6.6
LEP only	66,410	19.0	62,273	18.5	64,153	18.9	62,559	18.7
LEP and SD	7,583	2.2	6,801	2.0	7,074	2.1	6,805	2.0
Non-LEP/Non-SD	250,905	71.8	244,847	72.8	245,838	72.3	243,199	72.6
All students	349,581	100.0	336,309	100.0	340,094	100.0	334,827	100.0
Grade 9								
SD only	18,750	6.0	16,732	5.7	17,350	5.8	16,736	5.7
LEP only	53,457	17.2	48,801	16.6	50,666	17.0	48,909	16.7
LEP and SD	4,534	1.5	3,919	1.3	4,149	1.4	3954	1.3
Non-LEP/Non-SD	233,189	75.2	224,215	76.4	226,393	75.8	223,721	76.3
All students	309,930	100.0	293,667	100.0	298,558	100.0	293,320	100.0

Note. LEP = limited English proficient. SD = students with disabilities.

The Site 2 data provide us with a unique opportunity to examine the issues concerning the English language learners. With a very large number of students of limited English proficiency status in the data files, we can study the interaction of language with other background factors. For example, student ELL status and family SES are highly correlated and to some degree are confounded. We need to study large numbers of students in order to understand the unique contributions of language factors above and beyond other background variables such as family SES.

Data from students in Grades 2, 7, and 9 are used for discussion throughout this section of the report. Some analyses also incorporated the data from students in Grades 3 and 11. Tables 1.9, 1.10, and 1.11 present descriptive statistics for student

Table 1.8

Stanford 9 Spelling, Science, and Social Studies Frequencies for Students in Grades 2, 7, and 9, Site 2 Statewide School Districts

			Students with a normal curve equivalent score					
	All students		Spelling		Science		Social science	
	No.	%	No.	%	No.	%	No.	%
Grade 2								
SD only	17,506	4.2	16,489	4.2	NA	NA	NA	NA
LEP only	120,480	29.1	109,198	27.5	NA	NA	NA	NA
LEP & SD	4,629	1.1	4,011	1.0	NA	NA	NA	NA
Non-LEP/Non-SD	271,554	65.6	267,063	67.3	NA	NA	NA	NA
All students	414,169	100.0	396,761	100.0	NA	NA	NA	NA
Grade 7								
SD only	24,683	7.1	23,390	6.8	6,945	6.8	5,998	6.9
LEP only	66,410	19.0	64,359	18.8	22,006	21.4	18,293	21.1
LEP & SD	7,583	2.2	7,178	2.1	2,755	2.7	2,477	2.8
Non-LEP/Non-SD	250,905	71.8	246,818	72.2	70,889	69.1	601,56	69.2
All students	349,581	100.0	341,745	100.0	102,595	100.0	86,894	100.0
Grade 9								
SD only	18,750	6.0	5,417	6.3	17,313	5.8	17,108	5.8
LEP only	53,457	17.2	16,035	18.6	50,179	16.9	49,859	16.9
LEP & SD	4,534	1.5	1,567	1.8	4,108	1.4	4,066	1.4
Non-LEP/Non-SD	233,189	75.2	63,347	73.3	225,457	75.9	223,989	75.9
All students	309,930	100.0	86,366	100.0	297,057	100.0	295,022	100.0

Note. LEP = limited English proficient. SD = students with disabilities.

ELL and SD status, school lunch program participation, and parent education in Grades 2, 7, and 9 respectively. The results of our analyses of the Site 2 data are consistent with our findings from the other sites and suggest that language affects performance in the content areas. The results reported in Tables 1.9, 1.10, and 1.11 indicate that (a) ELL students perform substantially lower than non-ELL students, particularly in content areas with more language load; (b) the gap between the performance of ELL and non-ELL students is smaller in the lower grades; and (c) student ELL status may be confounded with family SES and parent education.

We used the percentage of over-achievement index (POA)[7] to demonstrate the points stated above. In addition to the mean, standard deviation, and number of

[7] Percentage of over-achievement was defined in the Site 1 section.

subjects for each subgroup, Tables 1.9 through 1.11 also include the POA. Through a comparison of the POA for math with the POAs for the language-related subsections (reading, language, and spelling), we can see the impact of language on student performance. The POA for the non-ELL students over the ELL students is lower on the math subtest. For example, for Grade 2 students (Table 1.9), the POA (non-ELL versus ELL) is 55.8% in reading (non-ELL students outperformed ELL students by 55.8%), 60.2% in language, and 42.8% in spelling, as compared with a POA of 33.5% in math. For Grade 7 students (Table 1.10), the POAs are 96.9% for reading and 70.7% for language, in comparison to 50.4% for math. This trend holds also for Grade 9 students.

In Tables 1.9 through 1.11, the means, standard deviations and POA by free/reduced-price lunch (a proxy for SES) and by parent education are also reported. The POA for the free lunch variable suggests that students who did not participate in a free or reduced-price lunch program performed substantially higher than those who did participate. For Grade 2 students (Table 1.9), these percentages are 32.7% in reading (students not receiving free/reduced lunch performed 32.7% higher than those receiving free/reduced lunch), 25.1% in math, 35.2% in language and 25.3% in spelling. The corresponding POAs for Grade 7 (Table 1.10) are 47.2% for reading, 29.5% for math, 32.9% for language, and 31.1% for spelling. For Grade 9 (Table 1.11), the percentages are 33.3% for reading, 19.8% for math, 19.9% for language, 19.3% for science, and 19.4% for social science.

Parent education seems to have a much greater impact on student performance. Percentages of over-achievement for the parent education variable were computed by subtracting the mean score of the lowest education category (Not High School Graduate) from the mean of the highest category (Post Graduate Studies) and dividing the difference by the mean from the lowest category, and multiplying the result by 100. For Grade 2 (Table 1.9) students, the POA is 106.3% in reading (students from parents with post graduate education performed 106.3% higher than those from parents with less than high school education), 84.9% in math, 118.5% in language, and 87.5% in spelling. Similar trends were found for students in Grades 7 and 9 (see Tables 1.10 and 1.11).

Table 1.9

Grade 2, Stanford 9 Subsection Scores and Percent of Over-Achievement (POA) by ELL Status, Free Lunch Program, and Parents' Level of Education

Subgroup	Reading	Math	Language	Spelling
	ELL status			
ELL				
M	31.6	37.7	31.6	33.7
SD	15.9	19.7	18.9	18.4
N	97,862	114,519	107,861	109,198
Non-ELL				
M	49.3	50.4	50.7	48.1
SD	19.7	21.9	23.2	20.1
N	252,696	267,397	263,955	267,063
POA	55.8	33.5	60.2	42.8
	School lunch			
Free/Reduced				
M	35.4	38.8	35.5	36.7
SD	17.5	20.1	20.5	18.7
N	106,999	121,461	116,202	117,482
Not free/Reduced				
M	47.0	48.5	48.0	46.0
SD	20.6	22.4	24.0	20.8
N	304,092	327,409	320,405	324,832
POA	32.7	25.1	35.2	25.3
	Parent education			
Not high school grad				
M	30.1	34.7	29.9	31.4
SD	15.3	19.1	18.2	16.6
N	54,855	63,960	60,466	61,431
High school graduate				
M	40.5	42.6	40.8	40.7
SD	18.1	20.3	21.4	18.8
N	93,031	101,276	98,798	100,142
Some college				
M	48.8	50.3	50.5	47.8
SD	18.6	20.6	22.1	19.2
N	66,530	70,381	69,428	70,149
College graduate				
M	56.5	58.4	59.2	54.9
SD	18.5	20.6	21.8	19.8
N	54,391	56,451	55,803	56,345
Post graduate studies				
M	62.1	64.1	65.3	58.9
SD	18.7	20.4	21.2	20.1
N	25,571	26,367	26,141	26,336
POA	106.3	84.9	118.5	87.5

Table 1.10

Grade 7, Stanford 9 Subsection Scores and Percent of Over-Achievement (POA) by ELL Status, Free Lunch Program, and Parents' Level of Education

Subgroup	Reading	Math	Language	Spelling
	ELL status			
ELL				
M	26.3	34.6	32.3	28.5
SD	15.2	15.2	16.6	16.7
N	62,273	64,153	62,559	64,359
Non-ELL				
M	51.7	52.0	55.2	51.6
SD	19.5	20.7	20.9	20.0
N	244,847	245,838	243,199	246,818
POA	96.9	50.4	70.7	81.1
	School lunch			
Free/Reduced				
M	34.3	38.1	38.9	36.3
SD	18.9	17.1	19.8	20.0
N	92,302	94,054	92,221	94,505
Not free/Reduced				
M	48.2	49.4	51.7	47.6
SD	21.8	21.6	22.6	22.0
N	307,931	310,684	306,176	312,321
POA	47.2	29.5	32.9	31.1
	Parent education			
Not high school grad				
M	31.2	36.2	36.4	32.8
SD	17.7	15.8	18.8	18.8
N	58,276	59,573	58,237	59,880
High school graduate				
M	39.3	40.9	42.9	40.2
SD	19.3	17.9	20.4	20.2
N	72,383	73,352	72,125	73,729
Some college				
M	49.1	49.0	52.2	48.5
SD	19.3	19.2	20.7	20.3
N	72,589	73,019	72,105	73,304
College graduate				
M	52.8	53.7	56.0	52.1
SD	20.4	21.3	21.6	20.9
N	82,417	82,804	81,855	83,110
Post graduate studies				
M	61.9	63.9	65.2	59.2
SD	20.6	22.2	21.2	20.8
N	39,443	39,609	39,319	39,697
POA	98.4	76.2	79.0	80.5

Table 1.11

Grade 9, Stanford 9 Subsection Scores and Percent of Over-achievement (POA) by ELL Status, Free Lunch Program, and Parents' Level of Education

Subgroup	Reading	Math	Language	Science	Social science
			ELL status		
ELL					
M	24.0	38.1	34.8	34.9	34.5
SD	12.5	15.2	13.7	12.8	13.4
N	48,801	50,666	48,909	50,179	49,859
Non-ELL					
M	46.0	53.5	52.4	49.2	49.3
SD	18.0	19.4	17.7	16.1	17.9
N	224,215	226,393	223,721	225,457	223,989
POA	91.6	40.3	50.5	41.2	34.3
			School lunch		
Free/Reduced					
M	32.0	42.5	41.0	39.4	39.3
SD	16.2	16.4	16.2	14.3	15.3
N	56,499	57,961	56,572	57,553	57,185
Not free/Reduced					
M	42.6	50.7	49.2	47.0	46.9
SD	19.7	20.1	18.9	17.0	18.6
N	338,285	343,480	337,623	341,663	339,445
POA	33.3	19.8	19.9	19.3	19.4
			Parent education		
Not high school grad					
M	29.2	39.6	38.3	37.3	37.2
SD	15.0	15.1	15.3	13.5	14.4
N	69,934	71,697	69,705	71,183	70,801
High school graduate					
M	35.6	44.1	42.9	41.7	41.0
SD	17.0	17.1	16.7	14.9	15.9
N	71,986	73,187	71,722	72,810	72,506
Some college					
M	44.6	51.6	50.5	48.2	47.7
SD	17.2	18.1	17.0	15.4	17.0
N	70,364	70,971	70,089	70,687	70,455
College graduate					
M	48.1	56.3	54.3	51.5	51.4
SD	18.5	19.6	18.1	16.4	18.2
N	87,654	88,241	87,354	87,956	87,746
Post graduate studies					
M	57.6	65.8	62.6	58.8	60.7
SD	19.6	20.7	18.6	17.1	19.7
N	34,978	35,087	34,910	35,022	35,005
POA	97.4	66.4	63.3	57.6	63.0

English language learner students may be more likely to have parents with a lower level of education. Thus, parent education and student ELL status may be confounded. Similarly, student ELL status may be confounded with family SES (measured by free/reduced-price lunch program participation), as ELL students may be more likely to be from families with lower SES. We will examine these hypotheses by applying more complex statistical models such as canonical correlation and regression models.

Comparing Performance of ELL and Non-ELL Students on Each Individual Item

The results of analyses comparing ELL and non-ELL students indicated that ELL students performed substantially lower than non-ELL students. This finding is consistent across grade levels, test levels, and across different sites. The results of item-level analyses are also consistent with the general statement that non-ELL students outperform ELL students. However, individual items may differentially separate ELL from non-ELL students. That is, some test items may show a larger performance difference between ELL and non-ELL students than other items.

To examine the level of differential performance of items when comparing ELL and non-ELL students, we computed the difference between the mean scores for each individual item across the ELL categories (ELL and non-ELL), as discussed in the Site 1 section of this chapter. (In the Site 1 section we compared bilingual and non-bilingual groups.) We computed the DBN (here, this is the difference between ELL and non-ELL student performance) for each individual item. A negative DBN indicates that English language learner students had higher performance than their non-ELL peers for that particular item. Table 1.12 summarizes the results of item-level analyses comparing ELL (bilingual) and non-ELL (non-bilingual) students.

As Table 1.12 shows, there is a large difference between test items in assessing the performance difference between ELL and non-ELL students. For example, the DBN index in math ranges from .03 to .26 for Grade 2 students, from .03 to .39 for Grade 7 students, and from .02 to .32 for Grade 9 students. For language and reading, the range of DBN is even wider than the range for math. For language, the range of DBN is from .05 to .45 in Grade 2, from –.01 to .32 in Grade 7, and from .04 to .31 in Grade 9. For reading the range is from .03 to .24 in Grade 2, from .02 to .50 in Grade 7, and from .03 to .44 in Grade 9.

The large differences between the performance of ELL and non-ELL students suggest that some of the test items could be more linguistically complex than others,

Table 1.12

Item-Level Response Differences Between ELL and Non-ELL Students (DBN)

Subsection/Grade	No. of items	Minimum	Maximum	Average DBN
Math				
Grade 2	72	.03	.26	.12
Grade 7	80	.03	.39	.19
Grade 9	48	.02	.32	.16
Language				
Grade 2	44	.05	.45	.19
Grade 7	48	–.01	.32	.24
Grade 9	48	.04	.31	.19
Reading				
Grade 2	118	.03	.24	.14
Grade 7	84	.02	.50	.25
Grade 9	84	.03	.44	.24

regardless of the item content difficulty. Of course, other factors, such as lack of construct knowledge or opportunity to learn, could contribute to these differences.

Relationship Between Stanford 9 Subsection Scores and Language: A Canonical Correlation Analysis

Literature suggests that student background variables impact students' performance in school (see, for example, Abedi, Lord, & Plummer, 1997; Abedi, Hofstetter, Baker, & Lord, 1998; Abedi, Lord, & Hofstetter, 1998; Alderman & Holland, 1981; Cocking & Chipman, 1988; Garcia, 1991; LaCelle-Peterson & Rivera, 1994). Among these background variables, family SES is one of the strongest predictors of school achievement. To examine the importance of language factors in predicting student performance above and beyond other background variables, a canonical correlation model was created. In this model, student Stanford 9 subsection scores were predicted from a free/reduced-price lunch index (a proxy for SES), parent education, and student ELL status. The purpose of this analysis was to determine how much of the variance of achievement scores can be explained by student ELL status above and beyond the parent education and socioeconomic variables.

We created three canonical correlation models, one for Grade 2, one for Grade 7, and one for Grade 9. The independent (Set 2) variables in all three models were ELL status, parent education, and free/reduced-price lunch status. For students in Grades 2 and 7, the canonical model included Stanford 9 subsection NCE scores in

reading, math, language, and spelling as the dependent (Set 1) variables. For Grade 9, the dependent variables were the reading, math, language, science, and social science NCE scores.

Table 1.13 presents a summary of the results of the canonical analysis for students in Grade 2. The canonical model yielded three functions, of which only the first was statistically significant (Wilks's Lambda = .70, $p < 0.001$) and explained more than 29% of the variance. The canonical correlation for this model was .542. All of the correlations of the Set 1 variables with the canonical variate were high, ranging from .766 (math) to .976 (reading). However, some of the correlations between the Set 2 variables and the canonical variate were not as high as in Set 1. Among the Set 2 variables, parent education had the highest correlation with the canonical variate (.912), ELL status had a moderate correlation with the canonical variate (–.697), and SES had a relatively small correlation with the canonical variate (–.475).[8]

The academic performance (Set 1) canonical variate consists mostly of the reading and language scores, as shown by the standardized canonical coefficients of .684 and .405 respectively. Math and spelling make negligible contributions to the

Table 1.13

Grade 2, Correlations Between Performance and Background Variables and First Canonical Variate, Standardized Canonical Coefficients, Percent of Variance Explained, and Canonical Correlation

Variable	First canonical variate	
	Correlation	Coefficient
Set 1 (dependent) variables		
Reading	.976	.684
Math	.766	–.072
Language	.926	.405
Spelling	.809	.014
Set 2 (independent) variables		
Parent education (ordered categories)	.912	.714
ELL status (categorical)	–.697	–.383
SES (ordered categories)	–.475	–.173
Canonical correlation	.542	
Percent of variance explained by first canonical pair	29.4	

[8] The negative sign of the correlation of a variable with the canonical variate is due to the reverse coding of the variable.

Set 1 canonical variate. The student background (Set 2) canonical variate consists mostly of the parent education variable (standardized coefficient = .714), with smaller contributions from ELL status (–.383) and SES (–.173).

The results of the canonical analysis described above suggest the following: (a) There is a high degree of intercorrelation in student performance among the different subject areas; that is, students who perform high in one of the four subject areas are expected to perform high in other areas. This result suggests that language may be an underlying factor in student achievement. It may also point to an underlying scholastic aptitude factor. (b) Student academic achievement is highly dependent on family and language factors, such as SES, parent education, and ELL status.

Table 1.14 summarizes the results of the canonical analysis for students in Grade 7. As in the Grade 2 model, the Grade 7 model used the four subsection scores (reading, math, language, and spelling) as the Set 1 (dependent) variables and student ELL status, family SES (measured by participation in a free/reduced-price lunch program), and parent education as the Set 2 (independent) variables.

The Grade 7 canonical model also yielded three functions, of which only the first was statistically significant (Wilks's Lambda = .67, $p < 0.001$) and explained over 31% of the variance. The canonical correlation was .558. All of the correlations of the Set 1 variables with the canonical variate were high, ranging from .800 (math) to

Table 1.14

Grade 7, Correlations Between Performance and Background Variables and First Canonical Variate, Standardized Canonical Coefficients, Percent of Variance Explained, and Canonical Correlation

	First canonical variate	
Variable	Correlation	Coefficient
Set 1 (dependent) variables		
Reading	.988	.767
Math	.800	.035
Language	.870	.028
Spelling	.854	.222
Set 2 (independent) variables		
Parent education (ordered categories)	.808	.540
ELL status (categorical)	–.805	–.558
SES (ordered categories)	–.518	–.221
Canonical correlation	.558	
Percent of variance explained by first canonical pair	31.2	

.988 (reading). As in Grade 2, the correlations between the Set 2 variables and the canonical variate were not as high as in Set 1. Among the Set 2 variables, parent education and ELL status strongly correlated with the canonical variate (.808 and –.805 respectively), whereas SES had a smaller correlation with the canonical variate (–.518).

For Grade 7, the reading score (standardized coefficient = .767) dominates in the canonical variate of the academic performance variables, while spelling makes a minor contribution. Surprisingly, the language score makes virtually no contribution (standardized coefficient = .028) to this canonical variate. The math contribution is also essentially nil. The canonical variate of the background variables consists mostly of ELL status and parent education (in roughly equal portions), with a much smaller contribution from the SES index.

Table 1.15 summarizes the results of the canonical analysis for students in Grade 9. The Grade 9 model used five subsection scores (reading, math, language, science, and social science) as the Set 1 (dependent) variables and student ELL status, family SES (free/reduced lunch participation), and parent education as the Set 2 (independent) variables.

Table 1.15

Grade 9, Correlations Between Performance and Background Variables and First Canonical Variate, Standardized Canonical Coefficients, Percent of Variance Explained, and Canonical Correlation

	First canonical variate	
Variable	Correlation	Coefficient
Set 1 (dependent) variables		
Reading	.990	.758
Math	.797	.074
Language	.853	.089
Science	.817	.120
Social science	.776	.022
Set 2 (independent) variables		
Parent education (ordered categories)	.861	.657
ELL status (categorical)	–.753	–.506
SES (ordered categories)	–.397	–.135
Canonical correlation	.544	
Percent of variance explained by first canonical pair	29.6	

The Grade 9 canonical model again yielded three functions, of which only the first was statistically significant (Wilks's Lambda = .69, $p < 0.001$) and explained more than 29% of the variance. The canonical correlation was .544. All of the correlations of the Set 1 variables with the canonical variate were high, ranging from .776 (social science) to .990 (reading). As in Grades 2 and 7, the correlations between the Set 2 variables and the canonical variate were not as high as for Set 1. Among the Set 2 variables, parent education and ELL status strongly correlated with the canonical variate (.861 and –.753 respectively), whereas SES had a smaller correlation with the canonical variate (–.397).

In the Grade 9 model, the academic performance canonical variate is almost exclusively the reading score (standardized coefficient = .758). The other academic variables make very small contributions (each standardized coefficient is at most .120). Parent education and ELL status again dominate in the student background canonical variate.

In all three grades, the academic variable that correlated most highly with the canonical variate was reading (.976 to .990). Among the student background variables, parent education and ELL status correlated most strongly with the canonical variate (magnitudes greater than .69). Taken together, the results of the multivariate canonical correlation analyses confirm our earlier findings which suggest that student language background has significant impact on academic performance.

Relationship Between Stanford 9 Subsection Scores and Language: Regression Analyses

To further examine the contribution of ELL status to predicting student performance, a series of regression models was examined. The dependent variables were the NCE scores on the reading, language, math, science, and social science subtests. For each subtest three models were examined. Model 1 was a simple regression model with the free/reduced-price school lunch index as the predictor variable. Model 2 used the school lunch index and parent education as the predictor variables. Model 3 used three predictor variables: the school lunch index, parent education and ELL status.

Table 1.16 presents a summary of the results of the regression analyses for Grade 9. Because of the large sample sizes, all models were significant with $p < .0005$ and all predictors were also significant with $p < .0005$. All of the Model 1 R^2 values

Table 1.16

Grade 9 Multiple Regression Results for All Subtests Except Spelling

Dependent variable	Model 1 R^2	Model 2 R^2	Model 3 R^2	Betas	
Reading NCE	.044	.212 Δ=.168	.275 Δ=.063	School lunch Parent education ELL	–.073 .339 –.270
Language NCE	.029	.162 Δ=.133	.200 Δ=.038	School lunch Parent education ELL	–.052 .311 –.209
Math NCE	.028	.166 Δ=.138	.185 Δ=.019	School lunch Parent education ELL	–.054 .336 –.149
Science NCE	.030	.157 Δ=.127	.185 Δ=.028	School lunch Parent education ELL	–.061 .311 –.180
Social sciences NCE	.026	.146 Δ=.120	.171 Δ=.025	School lunch Parent education ELL	–.054 .305 –.168

Note. Model 1 predictor: School lunch. Model 2 predictors: School lunch, Parent education. Model 3 predictors: School lunch, Parent education, and ELL status. Δ = change in R^2.

were small, ranging from .026 in social science to .044 in reading. In all content areas, R^2 increased substantially (and significantly) in Model 2 when parent education entered the prediction. The increase in R^2 was largest in reading and smallest in social science. The increases in R^2 when ELL status entered the predictions (from Model 2 to Model 3) were small but statistically significant, ranging from .019 in math to .063 in reading. The standardized regression coefficients (Beta) suggest that in all five content areas, parent education is the most powerful of the three predictors, followed by ELL status. The negative Betas for the ELL status and school lunch variables indicate that higher content NCE values are associated with the non-ELL and no free/reduced-price school lunch categories. As expected, higher NCEs are associated with higher levels of parent education.

Internal Consistency of Test Items by Student Language Status

The results of internal consistency analyses that were reported for Site 1 clearly demonstrated that ELL students' responses to test items suffered from lower internal consistency as compared with responses of non-ELL students. These results may lead us to believe that language factors may be responsible for the lower internal consistency for ELL students. However, the results of multiple regression in

Site 1 and canonical correlation in Site 2 suggested that factors other than language may also contribute to the gap between the internal consistency of the two groups (ELL and non-ELL). For example, the results of multiple regression analyses for Site 1 (reported earlier) showed that ethnicity was the strongest predictor among others (gender, reading and math scores) of students' ELL status. However, ethnicity is a complex construct, and this variable is also confounded with other variables such as student family SES.

The results of canonical analyses on the data from Site 2 also helped us to understand confounding of students' ELL status with other background variables. The results of canonical correlation analyses indicated that parent education was one of the strongest associates of students' ELL status. In this model, SES, which is simply a proxy for family income, also showed a strong level of relationship with students' ELL status. However, the results of multiple regression and canonical analyses suggested that the variability of students' ELL status could not be explained completely by other student background characteristics. To shed light on this issue, we decided to compute and compare internal consistency of test items by SES and ELL categories.

As we indicated earlier, a main factor affecting the internal consistency coefficient (alpha coefficient) is the distribution of scores. Restriction of range in the distribution of scores may have substantial impact on alpha and may cause alpha to be underestimated. To present a clear picture of the restriction of range issue, we also presented the distribution of scores for the subgroups.

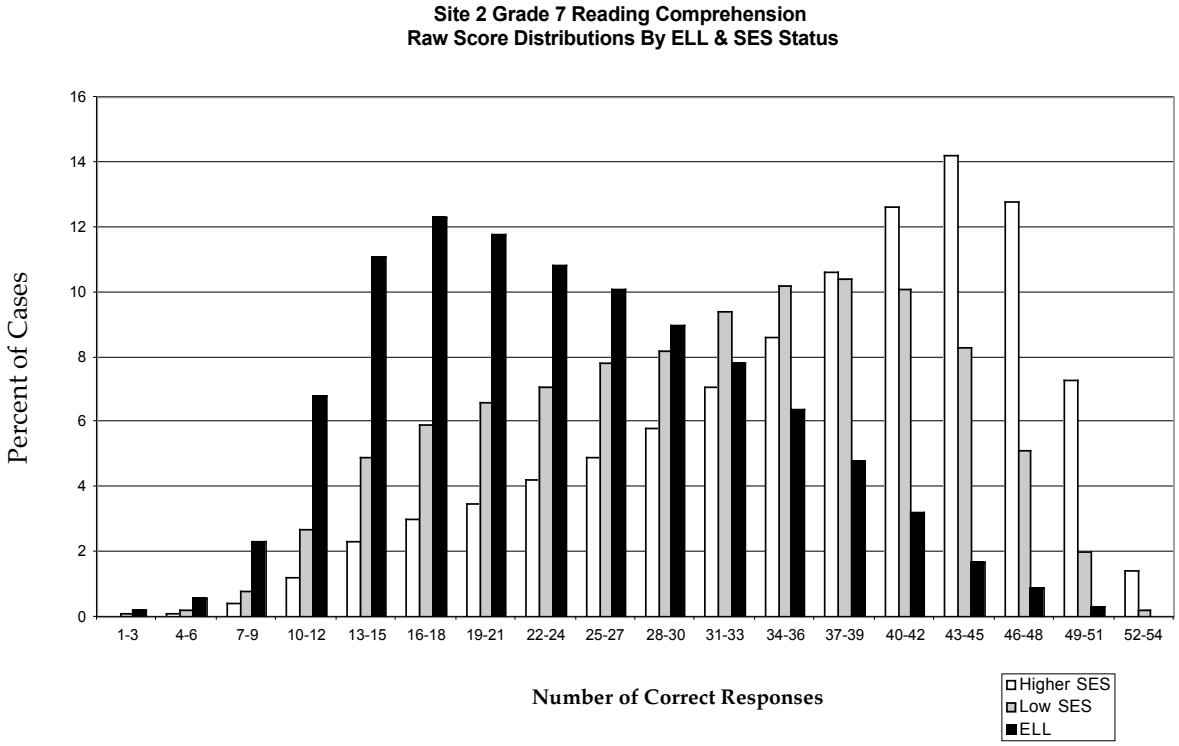
First we discuss the results of our internal consistency analyses, and then we discuss the effect of score distributions on alpha coefficients.

We categorized all students into three mutually exclusive categories. Non-ELL students were categorized as high and low SES based on participation in a free/reduced-price lunch program. The third category was comprised of ELL students. We then computed alpha coefficients for these three subgroups. If students' ELL status is explained mainly by their family SES and if ELL students are mainly from lower SES categories, then alpha coefficients computed for lower SES categories should be similar with those computed for ELL students.

As indicated earlier, we computed alpha coefficients for students in Grades 2, 7, 9, and 11 in Site 2. However, the trend of results is very similar across the different grades. Therefore, we report the results for Grade 7 only.

The table at the bottom of Figure 1.6 presents alpha coefficients for reading comprehension for Grade 7 students. As the table shows, the alpha coefficient for the high SES group is .906 as compared with the alpha of .902 for the low SES group, a minor difference. The coefficient for the ELL group, however, is lower (α = .870) than the coefficient for the low SES group (α = .902). Variance for the high SES group (104.49) and for the low SES group (109.19) is similar, but the ELL group has a smaller variance (86.40). Thus, the lower reliability for the ELL group may be due to restriction of range. However, as indicated earlier in this report, restriction of range may have been the result of language factors because language may have limited students' level of ability in responding to the test items.

Figure 1.6 presents the distribution of reading comprehension scores for the three groups (high SES, low SES and ELL). ELL students have a positively skewed

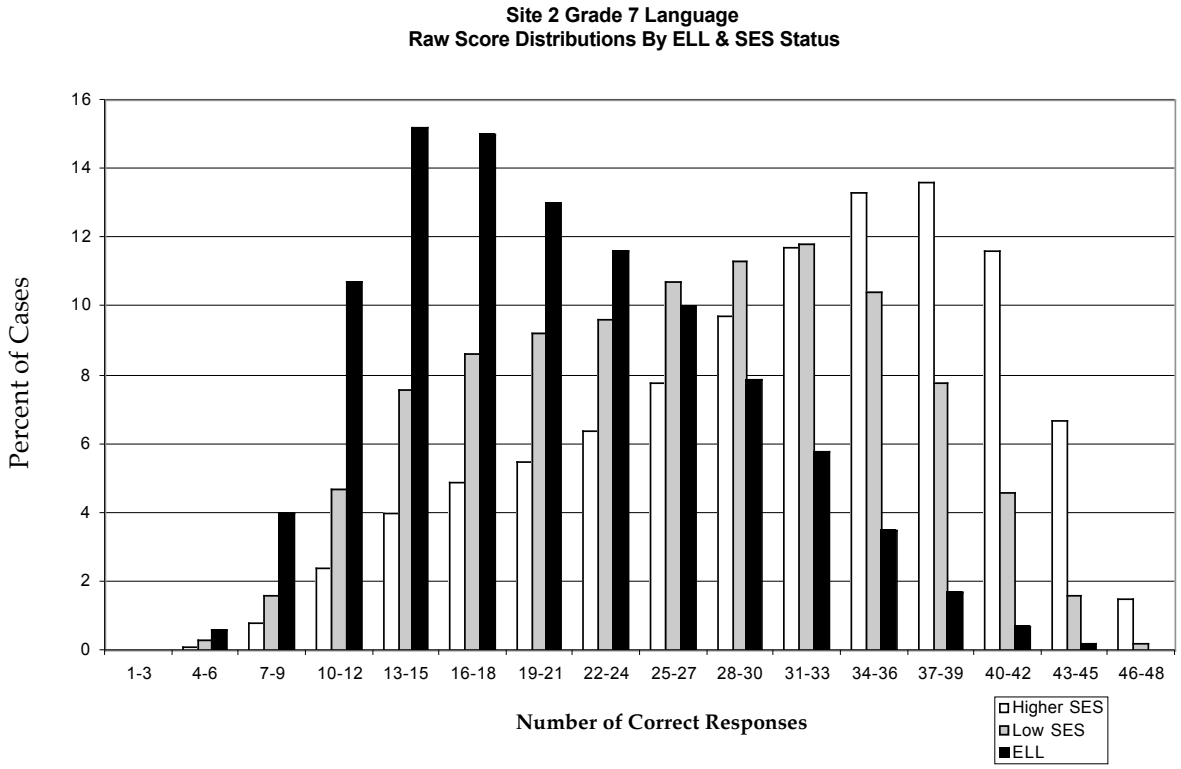


ELL/SES Status	Mean	Variance	Cronbach alpha
High SES	36.74	104.49	.906
Low SES	31.26	109.19	.902
ELL	23.85	86.40	.870

Figure 1.6. Site 2 Grade 7 reading comprehension score distribution and reliability.

distribution. Distributions for high and low SES students, on the other hand, are negatively skewed. As Figure 1.6 shows, the distributions for the high SES and ELL groups have relatively similar degrees of skewness but in different directions. These results may confirm our earlier statement that a portion of variance in the students' ELL status may be unique and may not be explained by other background characteristics.

Figure 1.7 presents the results for language scores for Grade 7 students. The trend of results for the language subsection is similar across the three content areas to the results just described for the reading comprehension subsection. Alpha coefficients for the high and low SES groups are relatively similar to each other and are different from the alpha coefficient for the ELL group.

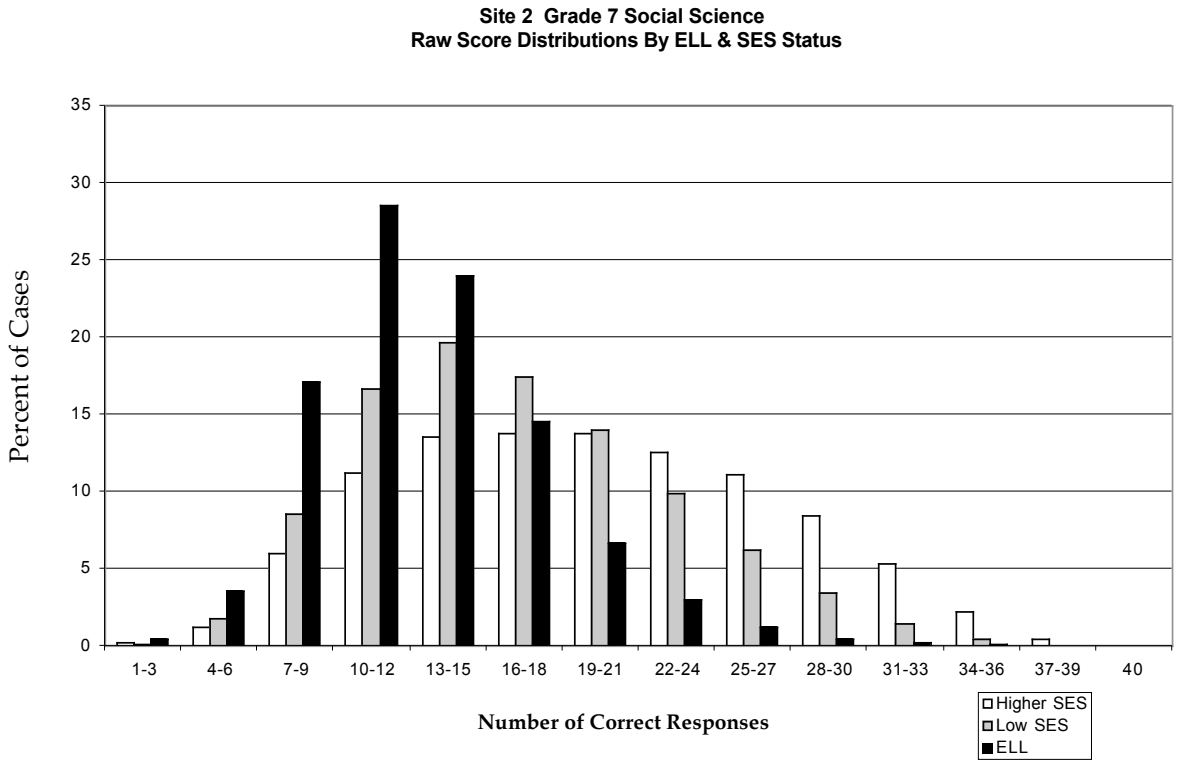


ELL/SES status	Mean	Variance	Cronbach alpha
High SES	31.15	83.69	.868
Low SES	26.35	79.69	.847
ELL	20.49	59.56	.803

Figure 1.7. Site 2 Grade 7 language score distribution and reliability.

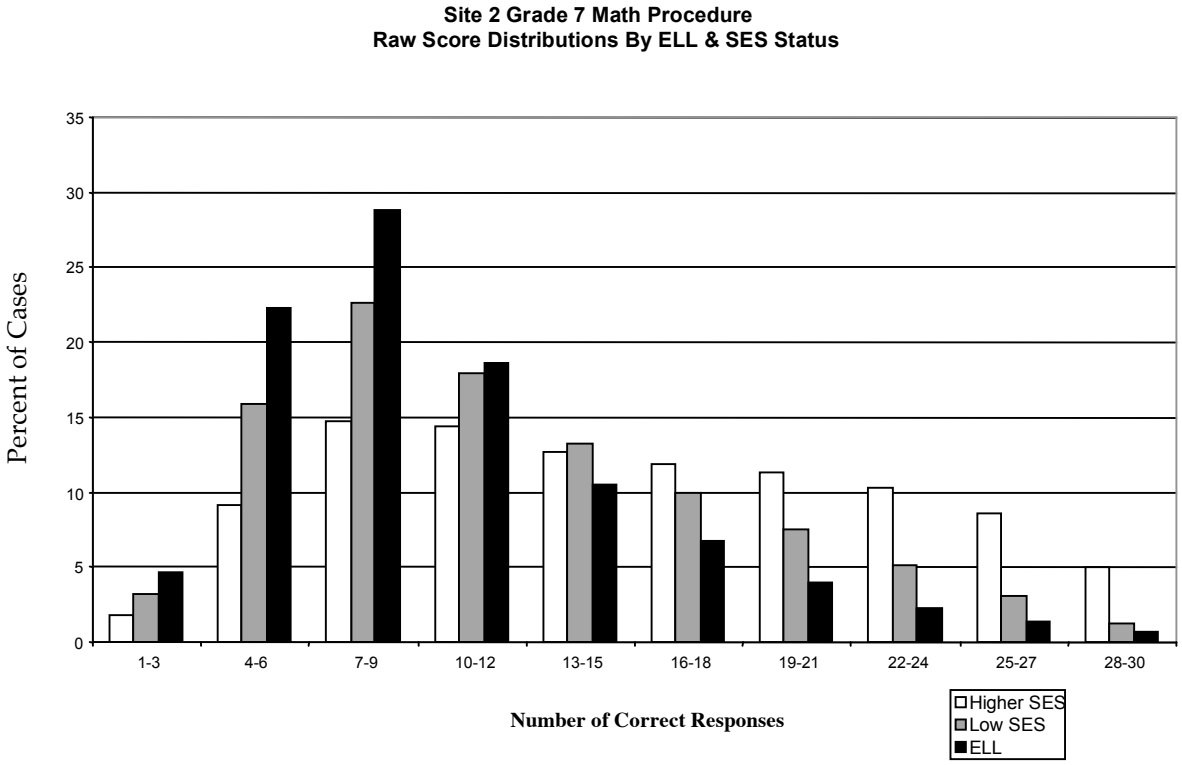
Figure 1.8 presents results for social science scores for Grade 7 students. For social science, more difference can be seen between the alpha coefficients for the high (.837) and low (.767) SES groups than had been seen in the reading and language subsections, but there is also a much larger difference between the ELL group (.605) and the non-ELL groups. On the reading and language subsections the distributions for the high SES and ELL groups showed a relatively similar degree of skewness but in different directions. The distribution across SES and ELL categories for the social science subsection, on the other hand, shows a large difference in the degree of skewness in the same direction.

Figure 1.9 presents results for math procedures scores for Grade 7 students. The distribution in this subsection more closely resembles the distribution of the social science subsection than the distributions seen in the reading and language



ELL/SES status	Mean	Variance	Cronbach alpha
High SES	19.65	52.47	.837
Low SES	16.70	36.82	.767
ELL	13.12	20.90	.605

Figure 1.8. Site 2 Grade 7 social science score distribution and reliability.



ELL/SES status	Mean	Variance	Cronbach alpha
Higher SES	15.38	51.38	.892
Low SES	12.07	38.03	.852
ELL	10.06	28.19	.803

Figure 1.9. Site 2 Grade 7 math score distribution and reliability.

subsections. The difference between the alpha coefficients for the high SES (.892), low SES (.852), and ELL (.803) groups is smaller than the results described in the social science subsection.

The results of these analyses suggest, once again, that even though students' ELL status may be confounded with their SES and other background characteristics, it may not be explained mainly by those characteristics.

Discussion

The purpose of the analyses of the existing data was to shed light on the issue of language and performance for English language learners. Specifically, by analyzing the existing data, we tried to answer the main research question in this

study: How can we determine whether content assessments in English are valid measures of ELL students' competence in subject areas?

For our extant data analyses, we have been fortunate to have access to several large school districts nationwide. Complete item-level data on standardized achievement tests along with student background variables, including language background variables, were obtained from different sites across the nation. Among the student background variables were family SES, ethnicity, gender, and parent education. However, it must be noted that the data files from the various sites were different in many aspects. Different standardized tests were used by the different sites. For example, the Stanford 9 was used by most of the sites, but different tests, such as the ITBS, were used by other sites. The student background variables also varied from site to site. Some sites provided data on student free/reduced-price lunch program participation as an index of family SES. At some sites we had access to other SES variables such as Aid to Families with Dependent Children (AFDC); at other sites we did not have any data on student SES. The main difference among the data from the different sites was the nature of student ELL status. Some sites provided student ESL status, some provided ELL status, and others provided bilingual program participation status. However, in spite of the differences in the data from the different sites, the existing data provided an excellent opportunity for examination of information relating to our main research questions.

The results of the analyses of the existing data were consistent within and across the sites. In a previous report (Abedi & Leon, 1999), we discussed the results of analyses that were performed on the data from Philadelphia and Hawaii. Results of these analyses indicated that ELL students generally performed lower than non-ELL students in all subject areas, and particularly so in those areas with more language load. For example, in our previous report we demonstrated that the gap between ELL and non-ELL students was smallest (and in some cases nonexistent) in content areas with a low level of language load, such as math computation, and was largest in content areas with a high level of language load, such as reading and writing. The fact that the gap between the performance of ELL students and native English speakers increases as the language load of the items increases provides strong evidence of the impact of language load on content area performance, particularly for ELL students.

A major finding in our study of extant data was lower reliability/internal consistency for the ELL students. The results of our analyses indicated that test items

for ELL students, particularly ELL students at the lower end of the English proficiency spectrum, suffered from lower internal consistency. Structural relationships between test scores for English language learners and native English speakers are different. For ELL students, the structural relationships were weaker. We speculated that this is due to language. That is, language factors introduce another source of measurement error into the structural models for ELL students.

In this chapter, we presented a summary description of the analyses that we performed on the data from Site 1 and Site 2. We tried to conduct analyses similar to those we discussed earlier in the previous reports. Similar analyses with different data sets enable us to examine the consistency of our findings across different sites. We also performed new analyses that were not possible with the other data sets. In our analyses of the Site 2 data we found that parent education was a powerful predictor of student ELL status. Such a finding was not possible with the other data sets because parent education information was available only in the Site 2 data.

The results of our analyses of the Site 1 and Site 2 data were consistent with those presented in our earlier report (Abedi & Leon, 1999). The Site 1 and Site 2 results confirm our earlier findings:

1. In all subject areas, English language learners, particularly those with limited English proficiency, perform substantially lower than native English speakers. That is, a gap between the performance of ELL students and native English speakers can clearly be seen.

2. The gap between ELL and non-ELL students increases as the language load of the assessment tools increases.

3. The linguistic complexity of test items may act as a source of measurement error in the assessment of English language learners.

There are also findings specific to Site 1 and Site 2. Analyses of data from Site 1 suggest that the confounding of language and performance in lower grades is less serious than in higher grades. For example, in Grade 3, the native English speakers outperformed the bilingual students by a small margin. The performance gap between bilingual students and native English speakers increased as the grade level increased.

Another interesting finding from the Site 1 data was the importance of background variables on student performance. In a multiple regression with content-based test scores (math and reading), gender, and ethnicity as predictor

variables, ethnicity showed the highest predictive power in predicting student bilingual status.

The Site 2 data provided a unique opportunity for studying the relationship between student English language proficiency and test performance. The data included a large population of ELL students, large enough to enable us to perform subgroup analyses at the categories of many different background variables. The data also provided us with information that was not available in the other data sets. Variables such as parent education and information on students' family socio-economic status made Site 2 data more useable.

The results of the Site 2 multivariate analyses, which were cross-validated, indicated that student family characteristics might be more important than we originally thought. For example, parent education proved to be the single most important variable when studying the impact of language on performance. The Site 2 data also enabled us to provide a more comprehensive picture of the performance of test items across the language proficiency categories. Some test items from the standardized achievement tests were shown to be more difficult for ELL students. We identified those items and we cross-validated our findings with another group of students.

References

Abedi, J. (1997). *Dimensionality of NAEP subscale scores in mathematics* (CSE Tech. Rep. No. 428). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (1998). *NAEP math performance and test accommodations: Interactions with student language background* (Draft Deliverable to NCES, Contract No. RS90159001). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (TOEFL Res. Rep. 9). Princeton, NJ: Educational Testing Service.

Bailey, A. L. (2000/2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663, pp. 79-100). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 17-46). Hillsdale, NJ: Lawrence Erlbaum Associates.

Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*, *26*, 371-391.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Series. Ninth edition, Form T.* San Antonio, TX: Harcourt Brace.

Hoover, H. D., Hieronymus, A. N., Dunbar, S. B., & Frisbie, D. A. (1996). *Iowa Tests of Basic Skills, Form M.* Chicago, IL: Riverside Publishing.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64(1),* 55-75.

Pedhazur, E. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Holt, Rinehart, and Winston.

Appendix

Grade 2 Results

Table A1.1

Mean, Standard Deviation, and Number of Students for ITBS Subsection Scores at the Different Grade/Level Combinations (NCE Scores) for Grade 2

Test level	Grade	Bilingual status	Math concepts	Math problem solving	Math computation	Reading
8	2	Non-Bilingual				
		M	44.82	44.41	46.58	39.36
		N	25,712	25,712	25,609	25,586
		SD	21.03	21.17	22.21	20.59
		Bilingual				
		M	52.34	49.06	54.60	42.59
		N	1,798	1,801	1,799	1,786
		SD	20.53	21.26	21.64	18.43

Table A1.2

Percentage of Over-Achievement of Non-Bilingual Students Over Bilingual Students on Reading and Math Subsections for Grade 2

Test level	Grade	Math concepts	Math problem solving	Math computation	Reading
8	2	–14.4	–9.5	–14.7	–7.6

Note. Math estimation and math data interpretation subsections are not available in Grade 2.

Table A1.3

Summary Results of Principal Components and Reliability Analyses for Grade 2

Subsection	Number of components Eigenvalue > 1	Percent of variance of 1st component	Reliability (α) bilingual	Reliability (α) non-bilingual
Math problem solving	4	18.01	.84	.83
Math concepts	4	16.29	.82	.82
Math computation	6	19.25	.85	.87
Reading	5	22.93	.89	.90

Note. Math estimation and math data interpretation subsections are not available in Grade 2.

Table A1.4

Internal Consistency Coefficients Adjusted by the Number of Items for Grade 2

Subsection	Unadjusted		Adjusted	
	Reliability (α) bilingual	Reliability (α) non-bilingual	Reliability (α) bilingual	Reliability (α) non-bilingual
Math problem solving (30 items)	.84	.83	.90	.89
Math concepts (31 items)	.82	.82	.88	.88
Math computation (30 items)	.85	.87	.90	.91
Reading (43 items)	.89	.90	.90	.91

Note. Math estimation and math data interpretation subsections are not available in Grade 2.

Table A1.5

Item-Level Response Differences Between Bilingual and Non-Bilingual Students (DBN) for Grade 2

Subsection	No. of items	Minimum	Maximum	Average DBN
Math problem solving	30	–.10	.03	–.04
Math concepts	31	–.13	–.01	–.07
Math computation	30	–.14	–.04	–.07
Reading	43	–.13	.09	–.05

Note. Math estimation and math data interpretation subsections are not available in Grade 2.

Table A1.6

Results of Multiple Regression Analysis for Grade 2

Variable	*B*	SE *B*	ß	*t*	Sig *t*
Math total	.0005	.0001	.043	5.761	<.0005
Reading	–.0009	.0001	–.075	–9.994	<.0005
Gender	.0009	.0030	.002	.328	.7430
Ethnicity	1.0090	.0140	.413	72.161	<.0005
Constant	.0119	.0050			
$R = 0.411$				$R^2 = 0.169$	

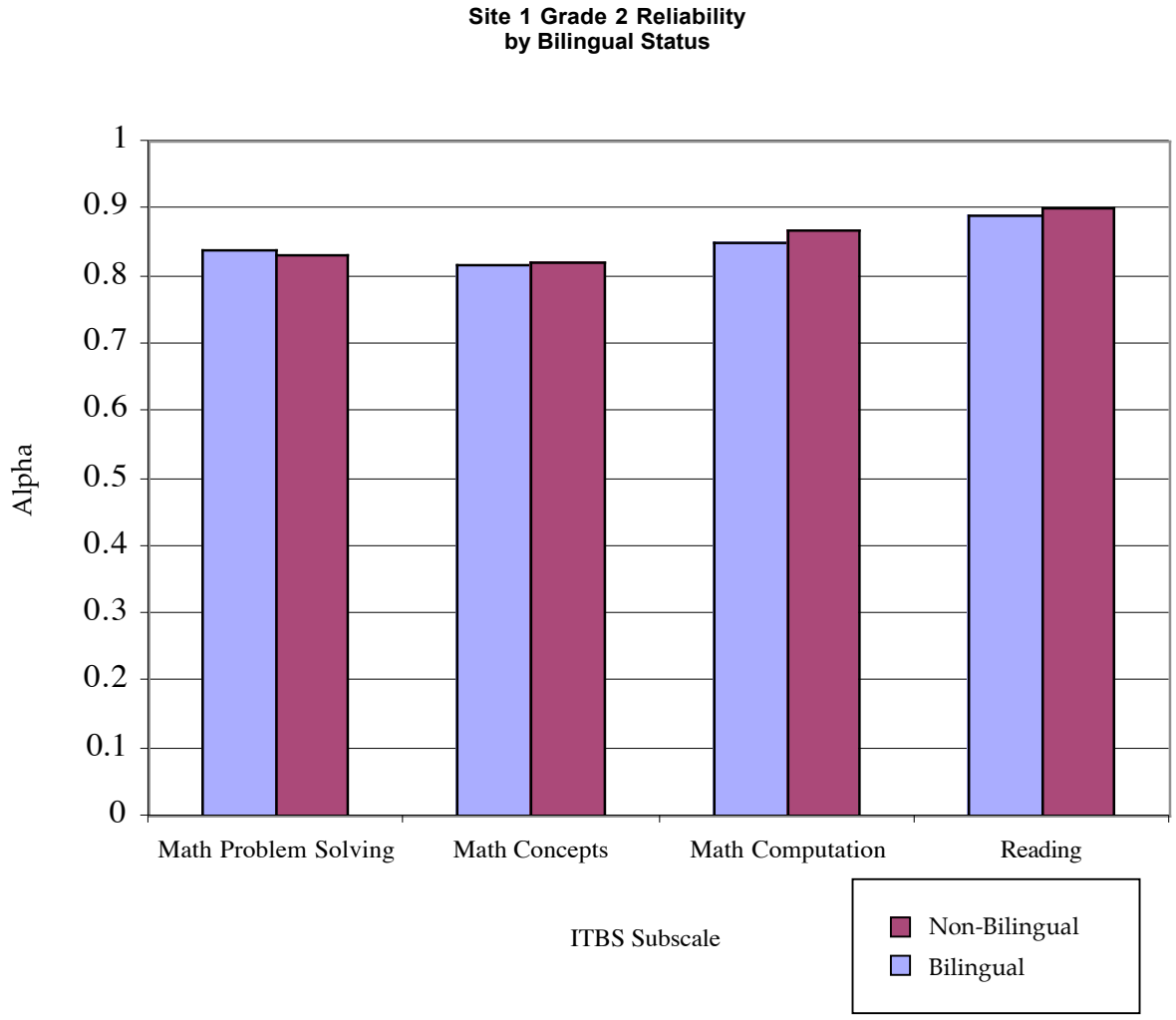


Figure A1.1. Site 1 Grade 2 reliability alpha coefficients.

CHAPTER 2

STUDENTS' CONCURRENT PERFORMANCE ON TESTS OF ENGLISH LANGUAGE PROFICIENCY AND ACADEMIC ACHIEVEMENT

Frances A. Butler and Martha Castellon-Wellington[1]

Summary

An overriding concern with large-scale content assessments is the validity of their use with English language learner (ELL) populations. One approach to addressing this issue is to compare student performance on a measure of content knowledge to concurrent performance on a language proficiency measure to determine whether students who perform at specified levels on the language assessment perform similarly on the content assessment. The purpose of this study was to investigate the relationship between same-student performance of ELL students on a standardized content assessment and a concurrent test of English language proficiency. The research sample for this study consisted of 778 3rd-grade students and 184 11th-grade students in two southern California school districts. The students were designated by their districts as English only (EO), fluent English proficient (FEP), or limited English proficient (LEP). All students took two standardized tests: the Stanford Achievement Test Series, Ninth Edition (Stanford 9; Harcourt Brace Educational Measurement, 1996) and the Reading/Writing Component of the Language Assessment Scales (LAS; Duncan & De Avila, 1990).

The results of the study show distinct differences in performance on the content subtests by the district-designated language categories. As expected, the LEP students in the sample performed less well than the non-LEP students. For both 3rd grade and 11th grade, the EO students outperformed FEP and LEP students on all the Stanford 9 subtests, with the FEP students outperforming the LEP students. At 3rd grade, the FEP students performed slightly lower than the EO students but considerably better than the LEP students. However, the gap between FEP students and EO students was considerably widened by 11th grade.

[1] The authors wish to thank Jamal Abedi, Alison Bailey, Rich Brown, Richard Durán, Joan Herman, Milagros Lanauze, Jim Mirocha, Don Powers, Lisle Staley, Robin Stevens, and David Sweet for their insightful comments on earlier versions of this chapter. In addition, a special thank you is extended to Seth Leon and Jim Mirocha for conducting the analyses reported here.

One group of LEP students in the third grade, however, those who met the criteria for redesignation, performed almost on a par with the EO students on the Stanford 9 subtests, above average in terms of norm group (NCE) scores. Content performance differences based on LAS English language proficiency categories—competent, limited, non-reader or writer—for third grade show that for the competent reader and writer categories, EO, FEP, and some LEP students (those who meet district redesignation criteria) scored at the mean (NCE 50) or above on the Stanford 9 subtests. LEP students who fall into the "competent" categories but do not meet all district redesignation criteria generally do not reach the national norm for average performance, possibly suggesting that the LAS criterion for competent performance may not be adequate for determining whether ELL students can handle the type of language found on content assessments. Another possible factor mediating LEP student performance is opportunity to learn. If students are not exposed in the classroom to the material on the content assessments, they cannot be expected to do well even if their language skills are improving.

Research Focus

The goal of the research reported in this chapter was to compare the performance of students on a standardized content assessment with concurrent performance by the same students on a measure of English language proficiency and thereby better understand the relationship between language proficiency, as measured by traditional language assessments, and student performance on tests designed to measure knowledge and skills in specific content areas. The results of these analyses augment the findings from the earlier extant data analyses (Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2000/2005) with regard to the language proficiency variable. The previous work did not include independent measures of language skills but rather looked at student performance on content measures by district- or state-designated language categories such as LEP/non-LEP and bilingual/non-bilingual.[2] This work provides an independent language proficiency measure against which performance on a content assessment can be examined. Though the primary research question being addressed in this study asks what the

[2] School districts may base their language designations on results from a commercially available test of English proficiency or on results from their own assessment method. Students given a LEP designation on school in-take may remain in that designation category for a number of years; thus students' levels of English proficiency at the point they take a standardized content assessment (which could be as much as 2 to 3 years later) may not be accurately reflected by their designation categories in the extant data sets.

relationship is between same-student performance of English language learners (ELLs) on a standardized content assessment and a concurrent test of English language proficiency, language proficiency data are also available for the EO students in the study. These data add a dimension not always considered in research with ELL students. Student performance on the content assessment, then, is examined based on proficiency categories established by the language proficiency test.

Participants

The data were collected in two southern California school districts—an elementary school district and a high school district—during spring 1999. Both ELL students and students who were native speakers of English in the 3rd grade and in the 11th grade participated. In total, 778 3rd-grade students from nine elementary schools were tested. Of these students, 296 (38%) were categorized by the school district as mainstream English only (EO), 77 (10%) were categorized as FEP, and 409 (52%) were categorized as LEP. These designations were determined upon each student's arrival in the district, whether at kindergarten, 1st, 2nd, or 3rd grade. Consequently for some students the designation is older than others.

At the high school level, 184 11th-grade students from three high schools were tested. Of these students, 115 (63%) were categorized as EO, 30 (16%) were categorized as FEP, and 39 (21%) were categorized as LEP.[3] At 11th grade, students designated LEP are either newly arrived in the district or have been in the district for some time and have weak language skills. All of the designations above were based on test scores independent of the test scores used in this study.

All study participants took the standardized achievement test and the language proficiency test. The number of students tested at the 11th grade was considerably less than at the 3rd grade due to the smaller number of ELL students in the high school district.

Instruments

The two primary test instruments used were the state mandated Stanford Achievement Test Series, Ninth Edition (Stanford 9; Harcourt Brace, 1996), and the Language Assessment Scales (LAS; Duncan & De Avila, 1990). The LAS Reading and Writing Components were administered approximately 1 month after the

[3] Some of the tables may not reflect the 3rd-grade and 11th-grade numbers exactly as reported here due to missing data.

regular district administration of the Stanford 9. In addition to these tests, the third-grade students took the Early Oral Reading Assessment (Jimerson & Klein, 1999) and the Third Grade District Writing Assessment as part of the regular district testing program. Analyses in this report focus on the first two assessments. Additional analyses, which include student performance on the latter two tests as well as the impact of opportunity to learn (OTL) on third graders in the content area of math, are provided in Staley (2005). The Stanford 9 and the LAS are described in turn below.

Stanford Achievement Test Series, Ninth edition. The Stanford 9 is a standardized, multiple-choice achievement test that measures content knowledge and skills across a range of content areas at kindergarten and above.[4] Table 2.1 provides the content areas for which Stanford 9 scores were available at each grade in our sample.

Language Assessment Scales. The LAS (Duncan & De Avila, 1990) is a test designed to measure the English language proficiency of ELL students in grades K-12. It is frequently used by schools for determining whether ELL students are fluent or limited in their English language proficiency. The LAS consists of reading, writing, and oral components. Only the Reading and Writing Component scores were available for this study. The LAS Reading Component for 3rd grade is a 45-item reading subtest; at 11th grade, the reading subtest consists of 55 items. At 3rd grade, the LAS Writing Component consists of ten items; at 11th grade, the Writing Component consists of five items and an essay. The sections contained in each subtest and the number of items per section are listed in Table 2.2.

All of the Reading Component subtests consist of multiple-choice items. The items generally focus on discrete elements of vocabulary and usage with the exception of items in the Reading for Information section, which focus on the retrieval of details from the text. The Writing subtests consist of writing single sentences at the 3rd-grade level and writing single sentences along with a 1-page essay at the high school level. For a content analysis of the LAS Reading Component (Forms 1A, 2A, and 3A) see Stevens, Butler, and Castellon-Wellington (2000).

[4] Although scores from this test were used in our analyses, researchers did not have access to the actual test content.

Table 2.1

Stanford 9 Subtests

Subtest	No. of items	Description of subtest
		3rd Grade
Reading	30	Reading Vocabulary (synonyms, multiple meanings, contexts)
	54	Reading Comprehension (reading passages consist of recreational, textual, and functional texts)
Mathematics	46	Problem Solving (subtopics include: concepts of whole number computation, number and sense numeration, geometry and spatial sense, measurement, statistics and probability, fraction and decimal concepts, patterns and relationships, estimation, problem solving strategies)
	30	Procedures (number facts, computation using symbolic notation, computation and context, rounding)
Language	18	Mechanics (capitalization, punctuation, usage)
	20	Expression (sentence structure, content & organization)
	10	Study Skills (dictionary skills, general reference sources, organizing information)
		11th Grade
Reading	30	Reading Vocabulary (synonyms, multiple meanings, contexts)
	54	Reading Comprehension (reading passages consist of recreational, textual, and functional texts)
Mathematics	48	Problem Solving (subtopics include: problem solving strategies, algebra, statistics, probability, functions, geometry from a synthetic perspective, geometry from an algebraic perspective, trigonometry, discrete mathematics, conceptual underpinnings of calculus)
Language	24	Mechanics (capitalization, punctuation, usage)
	24	Reading Comprehension (reading passages consist of recreational, textual, and functional texts)
Science	40	Content areas include: earth and space science, physical science, and life science
Social science	40	Content areas include: history, geography, civics and government, economics, culture

Table 2.2

LAS Subtests

3rd grade: Form 1A	No. of items	11th grade: Form 3A	No. of items
		Reading	
Vocabulary	10	Synonyms	10
Fluency	10	Fluency	10
Reading for information	10	Antonyms	10
Mechanics and usage	15	Mechanics and usage	15
		Reading for information	10
		Writing	
Finishing sentences	5	Finishing sentences	5
What's happening?	5	Let's write (essay)	

Data Analysis

The data analyses that follow are presented first for the 3rd grade, then for the 11th grade. The analyses include descriptive statistics, correlations, and performance trends across proficiency levels and content subtests.

Third Grade

The third-grade students are categorized as EO, FEP, or LEP, as they were designated in the district database. Although there is a redesignated fluent English proficient (RFEP) category in the database, no third-grade students in the sample had that designation since third graders are typically redesignated at the end of the school year.

Descriptive statistics. Tables 2.3 and 2.4 provide descriptive statistics for the third-grade students.

Table 2.3 presents the standard score means, standard deviations, medians,[5] and ranges for the EO, FEP, and LEP groups on the LAS Reading and LAS Writing tests. For LAS Reading, the mean for EO students was 92.6 (SD = 10.6), for FEP 90.2 (SD = 13.6), and for LEP 80.5 (SD = 15.3). For LAS Writing, the mean for EO students was 79.0 (SD = 11.1), for FEP 76.4 (SD = 9.0), and for LEP 69.4 (SD = 12.7). As expected, the EO students performed best in both the reading and writing skill

[5] The medians are included in Tables 2.3 through 2.6 to present a picture of how the distributions deviate from a symmetric distribution.

Table 2.3

Standard Score Descriptive Statistics for LAS Reading and Writing by Language Proficiency Category (Grade 3)

	n	*M*	*SD*	Median	Min.	Max.
EO						
LAS-R	292	92.6	10.6	96.0	31.0	100.0
LAS-W	280	79.0	11.1	80.0	3.0	100.0
FEP						
LAS-R	77	90.2	13.6	93.0	24.0	100.0
LAS-W	75	76.4	9.0	73.0	57.0	97.0
LEP						
LAS-R	409	80.5	15.3	84.0	36.0	100.0
LAS-W	383	69.4	12.7	70.0	13.0	97.0

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient. LAS-R = LAS Reading test; LAS-W = LAS Writing test.

areas, with FEP students next, followed by LEP students. All three groups had higher scores on the reading test than on the writing test (the two are scaled comparably), which suggests that regardless of the language proficiency category, all of the third graders in the study were stronger in reading than in writing, at least as the two skills are measured by the LAS. The minimum and maximum scores and the standard deviations for each group on both tests show a considerable range of performance within as well as across the proficiency groups. The maximum scores for LEP students for both reading and writing indicate that some of the students in that group were performing within limited and competent ranges. The mean on LAS Reading for the LEP students (80.5) suggests that those students as a group were competent readers according to LAS guidelines for score use.[6] By contrast, the minimum scores for EO students were surprisingly low on a test designed to assess second language proficiency, indicating that some native speakers were extremely weak in their reading and writing skills. These findings will be discussed further in conjunction with student performance on the Stanford 9 subtests.

[6] The *LAS Examiner's Manual for Reading/Writing* (Forms 1A and 1B) provides the following competency levels: For reading—a standardized score of 0-59 = Competency Level 1, non-reader; a standardized score of 60-79 = Competency Level 2, limited reader; a standardized score of 80-100 = Competency Level 3, competent reader. The same standardized scores and competency levels apply to writing.

Table 2.4 provides the Normal Curve Equivalent (NCE)[7] means, standard deviations, medians, and ranges for the EO, FEP, and LEP groups on the Stanford 9 Reading, Math, and Language subtests. For Stanford 9 Reading, the mean for EO students was 55.9 (*SD* = 18.9), for FEP 48.4 (*SD* = 15.4), and for LEP 31.5 (*SD* = 14.2). Again, the EO students were the highest performing group, followed by FEP and LEP students in that order. For Stanford 9 Math, the mean for EO students was 58.9 (*SD* = 21.8), for FEP 51.4 (*SD* = 21.3), and for LEP 42.4 (*SD* = 17.6). The same pattern holds with EO students performing best, followed by FEP and then LEP students. For Stanford 9 Language, the mean for EO students was 54.3 (*SD* = 21.0), for FEP 48.7 (*SD* = 17.5), and for LEP 35.0 (*SD* = 15.1). EO students, as a group, again outperformed the FEP and LEP students.

To test the significance of the differences between mean test scores for EO, FEP, and LEP students, a single-factor multivariate analysis of variance (MANOVA) model was used. In this model, language proficiency—EO, FEP, and LEP status—was used as the between-subjects variable with Stanford 9 Reading, Math, and Language scores used as the outcome variables. The overall model was significant (Wilks's Lambda .648, $F = 59.55$, $p < .001$). The univariate analysis indicated that the

Table 2.4

Normal Curve Equivalent Descriptive Statistics for Stanford 9 Reading, Math, and Language by Language Proficiency Category (Grade 3)

	n	*M*	*SD*	Med.	Min.	Max.
Reading						
EO	294	55.9	18.9	57.3	6.7	99.0
FEP	68	48.4	15.4	47.4	15.4	84.6
LEP	392	31.5	14.2	32.3	1.0	67.7
Math						
EO	296	58.9	21.8	58.7	1.0	99.0
FEP	73	51.4	21.3	51.1	1.0	99.0
LEP	408	42.4	17.6	41.1	1.0	93.3
Language						
EO	294	54.3	21.0	53.2	6.7	99.0
FEP	70	48.7	17.5	48.5	10.4	99.0
LEP	399	35.0	15.1	33.7	1.0	82.7

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient.

[7] NCEs are used to provide comparability across data sets from other school districts and states.

mean reading scores were different across the three language proficiency categories (F = 189.90, DF = 2 and 738, $p < .001$). Similarly the math scores were significantly different across the language proficiency categories (F = 55.61, DF = 2 and 738, $p < .001$). The language test means also showed significant differences across the three categories of language proficiency (F = 99.31, DF = 2 and 738, $p < .001$). As expected, these results indicate a real, quantifiable difference in performance on a standardized content assessment for third-grade students with differing language ability in the language of the assessment.

In addition to looking at the group means for these students, it is important to consider the ranges as well. The minimum and maximum scores suggest, just as with the LAS scores, considerable variability within each group. Of particular note is the performance of the LEP students on all three subtests. In every instance there are LEP students with maximum scores that are above average; that is, above the NCE mean of 50. In Math and Language especially, the LEP maximums are high. This information, when coupled with the LAS data, suggests that there may be students currently classified as LEP who, in terms of language ability as measured by the LAS, actually belong in the RFEP category. To explore this possibility, redesignation criteria used by the district were applied to the LEP group to determine whether, in fact, any students in the sample currently designated LEP would more appropriately be classified as RFEP.[8] Forty students in the third grade LEP category met the criteria. For the purposes of the remaining third-grade analyses reported here, those students 40 are included as a separate group designated RFEP. Table 2.5 provides the revised descriptive statistics for the LAS Reading and Writing student sample based on the hypothetical redesignation of the 40 LEP students to RFEP status. Statistics for the EO and FEP groups are unchanged from Table 2.3.

For LAS Reading, the new mean for LEP students was 78.8 (SD = 15.2), down slightly from 80.5. The LAS Reading mean for the newly created category RFEP was 96.2 (SD = 3.4). The relatively high mean for the RFEP students reflects the application of the redesignation criteria, which require a score of 80 or better on both

[8] The districts' redesignation criteria required students to receive a Level 3 (*competent*) rating on the LAS Reading and Writing subtests, a Level 5 on LAS Oral, and performance at the 36th percentile or better on the Stanford 9 Reading subtest. Three of the four measures—LAS Reading, LAS Writing, and Stanford 9 Reading—were available for the students in the sample, so those three scores were used as criteria for moving students from LEP to RFEP status for purposes of analyses. Had the LAS Oral score been available, there may have been some differences in the students moved to RFEP in the study.

Table 2.5

Standard Score Descriptive Statistics for LAS Reading and Writing by Language Proficiency Category With RFEP Added (Grade 3)

	n	*M*	*SD*	Med.	Min.	Max.
EO						
LAS-R	292	92.6	10.6	96.0	31.0	100.0
LAS-W	280	79.0	11.1	80.0	3.0	100.0
FEP						
LAS-R	77	90.2	13.6	93.0	24.0	100.0
LAS-W	75	76.4	9.0	73.0	57.0	97.0
RFEP						
LAS-R	40	96.2	3.4	97.0	84.0	100.0
LAS-W	40	85.3	5.0	83.0	80.0	97.0
LEP						
LAS-R	369	78.8	15.2	82.0	36.0	100.0
LAS-W	343	67.5	12.0	70.0	13.0	97.0

Note. EO = English only; FEP = fluent English proficient; RFEP = redesignated FEP; LEP = limited English proficient. LAS-R = LAS Reading test; LAS-W = LAS Writing test.

LAS Reading and Writing, as well as a 36th percentile ranking or above on the Stanford 9 Reading test. Interestingly, the group mean for the students in the RFEP category appears higher than that of the EO students. Due to the unequal sample size, the difference in the two means could not be tested for significance.

For LAS Writing, the new mean for LEP students was 67.5 (*SD* = 12.0), down from 69.4. The LAS Writing mean for RFEP students was 85.3 (*SD* = 5.0). The descriptive statistics for performance on LAS Reading and Writing show that the 40 RFEP students were a highly proficient group in terms of English language ability, performing as well as many EO and FEP students on the language tasks being assessed.

Figures 2.1 and 2.2 provide visual representations of the distributions for LAS Reading and Writing by percentage of cases. FEP and RFEP students are combined in the figures due to the small number of students in each category.

Figure 2.1 demonstrates the negatively skewed nature of the LAS Reading distribution for all language proficiency groups. Eighty percent of EO students and a slightly higher percentage of FEP/RFEP students had a reading score between 90-100. Another 8% of EO and 9% of FEP/RFEP students had scores between 80-89. A

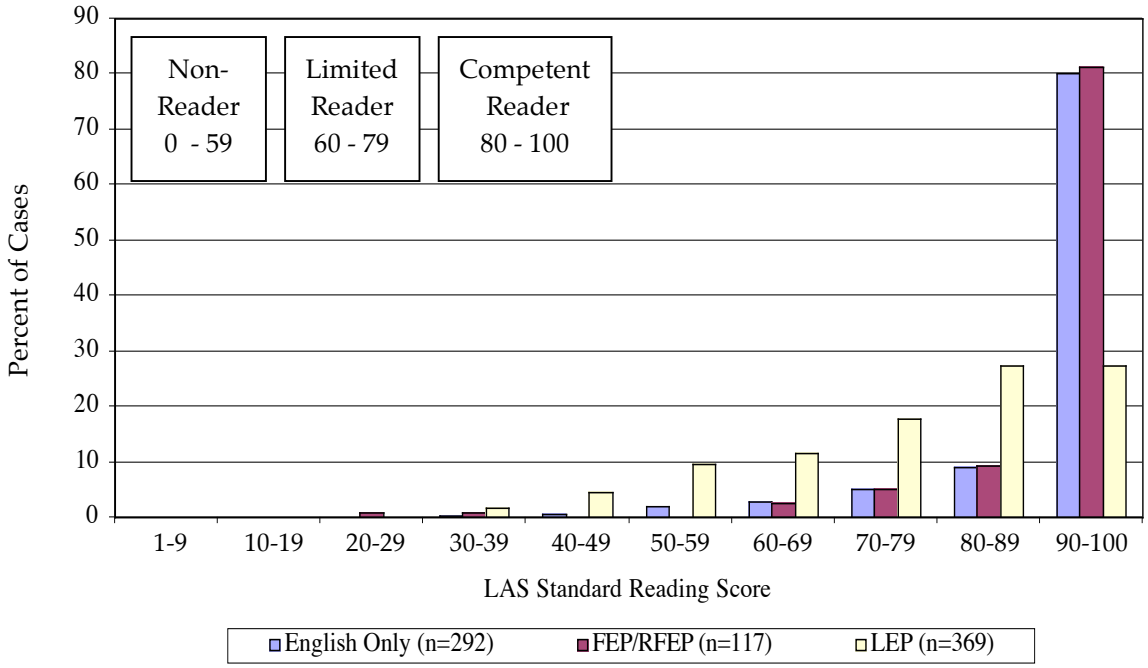


Figure 2.1. Distribution of standard scores for LAS Reading by percentage of cases for language proficiency categories (Grade 3).

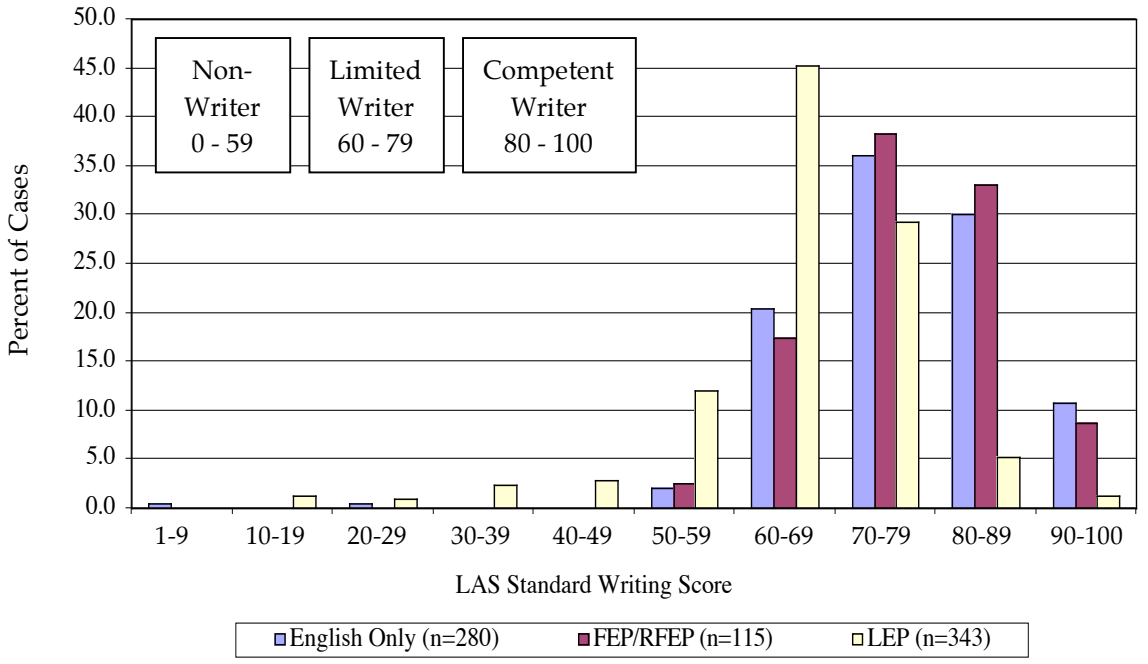


Figure 2.2. Distribution of standard scores for LAS Writing by percentage of cases for language proficiency categories (Grade 3).

little more than 50% of LEP students scored between 80-100 (27% each in the 80-89 and 90-100 ranges) and were thus competent readers according to LAS scoring criteria. Clearly the LAS Reading test was easy, as expected, for most EO and FEP/RFEP third-grade students in the study, providing no discrimination in the competent reader range (80 and above). However, for those students who were limited or non-readers, the LAS Reading Component did provide a higher degree of discrimination. That is, the LAS Reading Component captures differences in reading ability below the competent level.

Figure 2.2 shows a distribution for LAS Writing that is shaped differently from the LAS Reading distribution. Though scores still tend towards the upper end of the distribution, LAS Writing captures more variability in student performance than LAS Reading. A very small percentage of students from all language proficiency categories scored in the 90-100 range (EO, 11%; FEP/RFEP, 9%; LEP, 2%). The highest percentage of both EO and FEP/RFEP students fell into the 70-79 range (EO, 36%; FEP/RFEP, 38.5%). The highest number of LEP students, just over 45%, fell into the 60-69 range. A comparison of Figures 2.1 and 2.2 shows, as the means indicate, that for the third graders in this study, regardless of proficiency level, LAS Reading was easier than LAS Writing.

Table 2.6 provides the NCE means, standard deviations, medians, and ranges for the EO, FEP, RFEP, and LEP groups on the Stanford 9 Reading, Math, and Language subtests. The EO and FEP numbers are unchanged from those in Table 2.4. For reading, the mean for LEP students dropped from 31.5 (see Table 2.4) to 29.1 (*SD* = 12.8); for RFEP the mean is 52.0 (*SD* = 7.3). For math, the mean for LEP students dropped from 42.4 to 40.3 (*SD* = 16.5). For RFEP students, the mean is 62.1 (*SD* = 14.5), which is higher than the means for both the FEP group (51.4) and the EO group (58.9). For language, the mean for LEP students dropped from 35.0 to 22.7 (*SD* = 13.7). For RFEP students, the mean is 55.0 (*SD* = 10.6), which is slightly higher than the means for both the EO group (54.3) and the FEP group (48.7). This table shows substantial differences in performance between the RFEP and LEP groups, as well as between the RFEP group and the EO and FEP groups. The RFEP students outperformed the remaining LEP students by a considerable margin on all of the subtests and outperformed the FEP students as well, though by a lesser margin. In addition, they slightly outperformed the EO students on the language subtest and outperformed them to a greater degree on the math subtest. In reading only did the RFEP students fall slightly behind the EO students.

Table 2.6

Normal Curve Equivalent Descriptive Statistics for Stanford 9 Reading, Math, and Language by Language Proficiency Category With RFEP Added (Grade 3)

	n	*M*	*SD*	Med.	Min.	Max.
Reading						
EO	294	55.9	18.9	57.3	6.7	99.0
FEP	68	48.4	15.4	47.4	15.4	84.6
RFEP	40	52.0	7.3	50.6	42.5	67.7
LEP	352	29.1	12.8	29.9	1.0	65.6
Math						
EO	296	58.9	21.8	58.7	1.0	99.0
FEP	73	51.4	21.3	51.1	1.0	99.0
RFEP	40	62.1	14.5	61.7	32.3	93.3
LEP	368	40.3	16.5	38.3	1.0	89.6
Language						
EO	294	54.3	21.0	53.2	6.7	99.0
FEP	70	48.7	17.5	48.5	10.4	99.0
RFEP	39	55.0	10.6	54.8	41.3	82.7
LEP	360	22.7	13.7	32.3	1.0	72.8

Note. EO = English only; FEP = fluent English proficient; RFEP = redesignated FEP; LEP = limited English proficient.

Figures 2.3 through 2.5 provide visual representations of the distributions for Stanford 9 Reading, Math, and Language subtest scores by percentage of cases.

Figures 2.3 through 2.5 show a range of performance for the language proficiency groups—EO, FEP, and LEP—on the three Stanford 9 content subtests—Reading, Math, and Language. For Reading, the three groups do not overlap at the far right of the distribution. Though the EO and FEP groups have more symmetric distributions with the FEP students peaking sharply in the middle, the LEP distribution is slightly positively skewed, demonstrating a group weakness for LEP students on Stanford 9 Reading. For Math, the distributions overlap except at the extreme high end. Indeed, Table 2.6 shows that the maximum score for LEP students on Math is 89.6, falling just short of the 90-100 range. Still, Math is the strongest of the three Stanford 9 content areas reported for these third-grade LEP students. For Language as with Reading, the more closely related content area, the distributions overlap except above 80. The LEP students peak at a lower point in the distribution, but there is a clear range of performance for all groups.

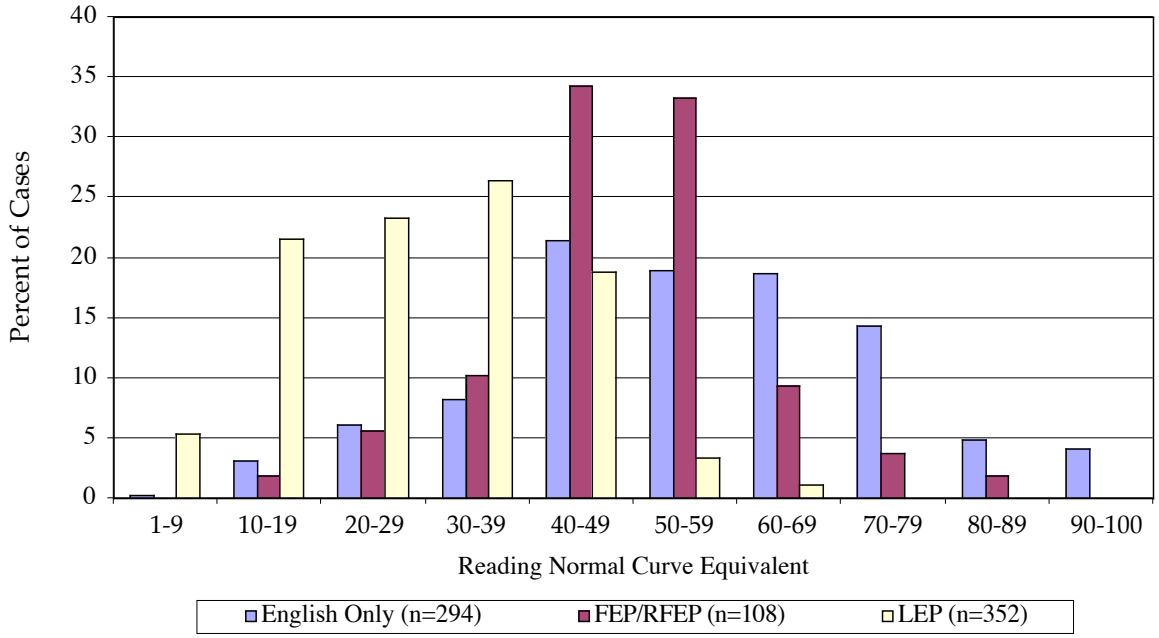


Figure 2.3. Distribution of normal curve equivalent scores for Stanford 9 Reading by percentage of cases for language proficiency categories (Grade 3).

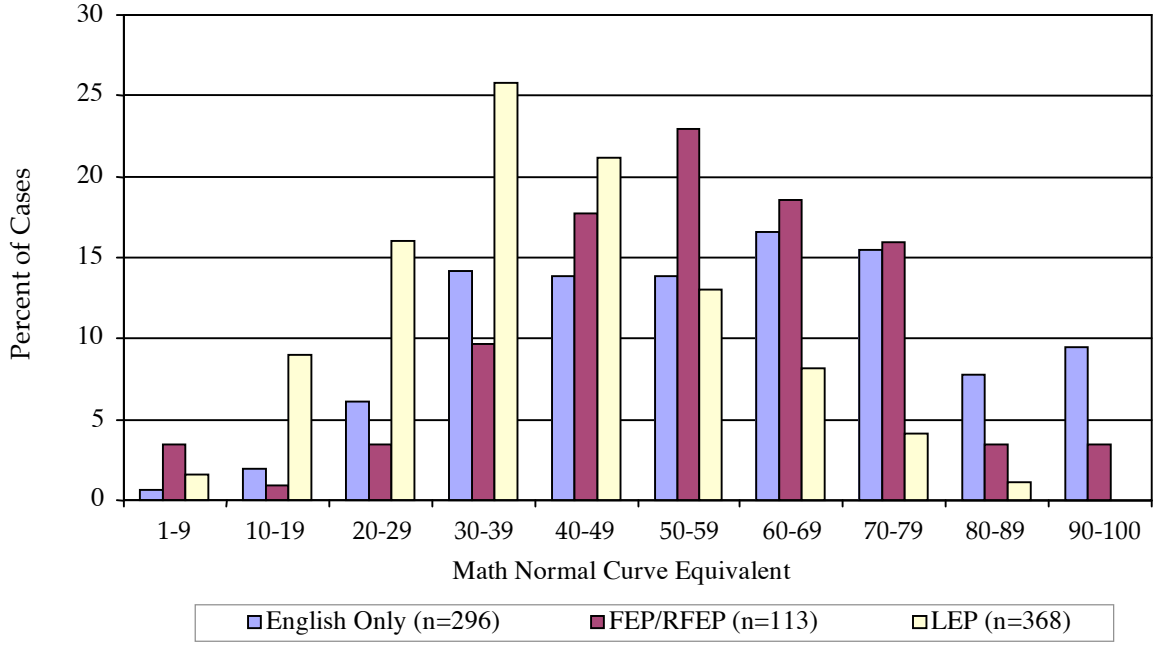


Figure 2.4. Distribution of normal curve equivalent scores for Stanford 9 Math by percentage of cases for language proficiency categories (Grade 3).

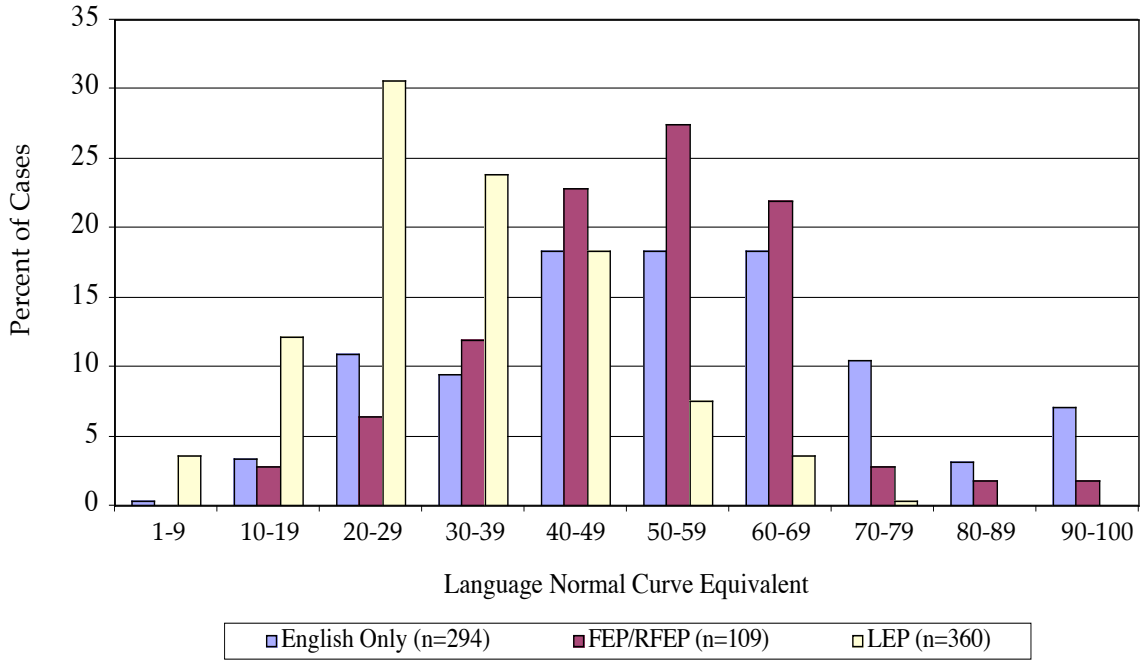


Figure 2.5. Distribution of normal curve equivalent scores for Stanford 9 Language by percentage of cases for language proficiency categories (Grade 3).

To test the significance of the differences between mean test scores for EO, FEP, and LEP students, a MANOVA model was used. The difference between this analysis and the earlier MANOVA is that the RFEP students were removed from the LEP group. Unfortunately, the RFEP group could not be included in the analysis due to restricted range for the group and the small sample size. In this model, as in the previous model, language proficiency—EO, FEP, and LEP status—was used as the between-subjects variable with Stanford 9 Reading, Math, and Language scores as the outcome measures. The overall model was significant (Wilks's Lambda .595, $F = 68.79$, $p < .001$). The univariate analysis indicated that the mean reading scores were different across the three language proficiency categories ($F = 229.23$, $DF = 2$ and 699, $p < .001$). Similarly the math scores were significantly different across the language proficiency categories ($F = 71.00$, $DF = 2$ and 699, $p < .001$). The language test means also showed significant differences across the three categories of language proficiency ($F = 126.48$, $DF = 2$ and 699, $p < .001$). These results, as expected, are similar to the results of the earlier MANOVA which also showed a significant difference in performance on a standardized content assessment for

third-grade students who have different levels of English language proficiency as measured by the LAS.

Test reliability. The reliability coefficients (internal consistency coefficients) for LAS Reading and Writing are provided in Table 2.7. Again, the RFEP group is not included because of the restricted range for the group and the small sample size.

Though the reading coefficients are somewhat higher than the writing coefficients for the three remaining groups, all of the coefficients are sufficiently high on both measures to demonstrate the internal consistency of the instruments, which shows that the tests are basically measuring the same construct across the proficiency groups—EO, FEP, and LEP. The state's item-level data for the Stanford 9 subtests for the 1999 administration were not available, and thus the reliability coefficients could not be calculated.

Test correlations. Table 2.8 provides the Pearson Product Moment correlations for the LAS standard scores for Reading and Writing with the Stanford 9 NCEs for Reading, Math, and Language. All of the correlations in the table are significant at the .001 level.

Correlations are presented for the EO, FEP, and LEP groups. The RFEP category is not included because three subtests used in the correlations—LAS Reading, LAS Writing, and Stanford 9 Reading—were used for the redesignation of LEP students to the RFEP category.

The LAS Reading correlations with the Stanford 9 subtests are higher than those of writing—the two exceptions being the EO group with the Stanford 9 Math and the FEP group with Stanford 9 Reading. Overall there was no major difference

Table 2.7

Reliability Coefficients (α) for LAS Reading and Writing Tests by Language Proficiency Category (Grade 3)

	Reading (45 items)		Writing (10 items)	
	n	α	*n*	α
EO	292	.889	280	.821
FEP	77	.916	75	.753
LEP	369	.876	343	.828

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient.

Table 2.8

Pearson Product Moment Correlations for LAS Standard Scores for Reading and Writing With Stanford 9 Normal Curve Equivalents for Reading, Math, and Language by Language Proficiency Category (Grade 3)

	Reading		Math		Language	
	n	*r*	*n*	*r*	*n*	*r*
EO						
LAS-R	280	.67	282	.53	280	.59
LAS-W	270	.42	272	.55	270	.53
FEP						
LAS-R	67	.46	72	.50	69	.55
LAS-W	65	.51	70	.48	67	.44
LEP						
LAS-R	338	.72	354	.53	346	.56
LAS-W	313	.50	328	.41	321	.37

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient. LAS-R = LAS Reading test; LAS-W = LAS Writing test.

in the magnitude of the correlations, which suggests that the relationships between performance on the language measures (LAS Reading and Writing) and the content assessment subtests (Reading, Math, and Language) are similar regardless of language proficiency.

Performance trends. Table 2.9 provides the Stanford 9 NCE means and standard deviations for Reading, Math, and Language by LAS Reading level. This table shows the differences in the language proficiency group performances across content areas for students who fall into the competent, limited, or non-reader categories based solely on their LAS Reading score. The competent readers (80-100 on LAS Reading) are the largest group; that is, more students from each proficiency category—EO, FEP, RFEP, and LEP—fall into this group than into either the limited or the non-reader group. RFEP students appear in the competent category only by virtue of the redesignation criterion for LAS Reading.

On Stanford 9 Reading, there appears to be little difference in group performance among competent EO (58.5), FEP (49.8), and RFEP (52.0) students, with all three performing about average or slightly above. However, the mean for the

Table 2.9

Stanford 9 Normal Curve Equivalent Means and Standard Deviations for Reading, Math, and Language by LAS Reading Level (Grade 3)

	Stanford 9								
	Reading			Math			Language		
LAS Reading level[a]	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Competent reader									
EO	256	58.5	16.4	257	61.4	20.6	255	56.8	19.1
FEP	62	49.8	15.1	65	53.8	20.0	62	50.9	17.2
RFEP[b]	40	52.0	7.3	40	62.1	14.5	39	56.5	10.6
LEP	211	35.0	10.4	217	46.5	14.9	213	38.1	12.9
Limited reader									
EO	18	26.1	8.3	18	33.1	11.0	18	26.0	8.8
FEP	5	34.6	11.2	7	27.4	20.3	7	30.0	5.7
LEP	92	21.8	8.3	98	33.1	13.8	95	25.0	9.4
Non-reader									
EO	6	16.9	7.4	7	27.4	13.1	7	20.3	8.9
LEP	35	12.5	7.0	39	24.0	11.8	38	20.8	8.7

[a]The LAS Reading levels by standardized scores are: Competency Level 1, non-reader, 0-59; Competency Level 2, limited reader, 60-79; Competency Level 3, competent reader, 80-100. [b]RFEP students are in the competent category by virtue only of the redesignation criterion for LAS Reading.

competent LEP students (35.0) is considerably below average.[12] For Stanford 9 Math and Language, the same trend continues with EO, FEP, and RFEP students having considerably higher means than the LEP students. For Math and Language, however, the EO and RFEP groups have closer means (Math, 61.4 and 62.1, and Language, 56.8 and 56.5, respectively) than either has to the FEP group means (Math, 53.8, and Language, 50.9). These results seem to suggest that the third-grade RFEP students in these analyses were likely candidates for redesignation at the end of the school year. Their performances across content areas were much stronger than those of the remaining LEP students and appear to be stronger than the performances of current FEP students and comparable to those of EO students.

In the limited-reader category, the EO and LEP students have means similar to each other across the three content tests. For the two language-related subtests, Reading and Language, the FEP mean is higher than the other two. For Math, the EO and LEP means are identical.

Six EO students fell into the non-reader category for Reading and seven for Math and Language. It is unclear why these students performed so poorly on the content assessments. They do not appear to be representative of their group. Only one of the low-performing EO students was receiving special services.

The LEP students in the non-reader category performed very poorly compared to the LEP students in the competent and limited-reader categories across the three content subtests. These differences in performances among LEP students highlight the range of achievement demonstrated on content assessments when students are grouped by language ability as measured by LAS Reading.

Table 2.10 provides the Stanford 9 NCE means and standard deviations for Reading, Math, and Language by LAS Writing level. This table shows the differences in the language proficiency group performances across content areas for students who fall into the competent, limited, or non-reader categories based solely on their LAS Writing score. Interestingly, with the writing score as the criterion, EO students are divided between the competent (n = 168) and limited (n = 123) categories as are the FEP students n = 30 and 37, respectively). As with Reading, RFEP students appear in the competent category only by virtue of the redesignation criterion for LAS Writing. LEP students fall largely into the limited category (n = 253) with 55 in the competent category

[12] MANOVA could not be run on the data reported in Tables 2.9 and 2.10 due to the unequal variances coupled with unequal n sizes. The assumption of homogeneity of variance was violated.

Table 2.10

Stanford 9 Normal Curve Equivalent Means and Standard Deviations for Reading, Math, and Language by LAS Writing Level (Grade 3)

	Stanford 9								
	Reading			Math			Language		
LAS Writing level[a]	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Competent writer									
EO	146	62.2	17.8	146	67.8	19.3	145	63.2	19.0
FEP	28	57.3	13.7	29	61.0	16.9	28	56.9	15.4
RFEP[b]	40	52.0	7.3	40	62.1	14.5	39	56.5	10.6
LEP	32	34.9	9.1	34	47.9	12.6	34	38.5	12.4
Limited writer									
EO	122	47.9	17.9	124	47.5	18.8	123	43.8	18.5
FEP	37	42.0	13.5	41	43.0	20.9	39	42.7	17.2
LEP	244	30.3	11.6	254	40.2	15.1	247	32.8	13.3
Non-writer									
EO	2	58.5	16.5	2	37.4	3.2	2	48.1	7.2
LEP	37	14.1	8.1	40	26.2	13.2	40	23.6	9.8

[a]The LAS Writing levels by standardized scores are: Competency Level 1, non-writer, 0-59; Competency Level 2, limited writer, 60-79; Competency Level 3, competent writer, 80-100. [b]RFEP students are in the competent category by virtue only of the redesignation criterion for LAS Writing.

and 40 in the non-writer category. Thus, for all language proficiency groups, the LAS Writing subtest proved more challenging than the LAS Reading subtest, with fewer students being categorized as competent based on their writing performance. This table again shows the disparity in performance between the LEP group and all others, as well as the differential performance of the limited and non-writer groups compared to the competent non-LEP groups. A comparison of the differences between the means for EO and FEP students in both the competent and limited categories shows that the two groups perform more similarly on the content tests when grouped according to their LAS Writing scores than when grouped by their LAS Reading scores. Also, when the students are grouped by writing scores only, there tends to be a higher level of performance on the content tests for all groups.[13]

Summary of third-grade findings. As expected, the EO students as a group in the third grade, with few exceptions, outperformed the ELL students on both the language test and the content assessment. All of the third-grade students, as a whole, performed better on the LAS Reading than on the LAS Writing, with EO and FEP students generally outperforming LEP students on both the LAS Reading and Writing. However, EO students generally outperformed those classified as FEP and LEP on all sections of the Stanford 9. Greater differences are found between EO and FEP students on each Stanford 9 subtest than on either section of the LAS. The differences in Stanford 9 mean performance between EO, FEP, and LEP students are statistically significant.

Some LEP students performed better than average on the Stanford 9. These students also performed better than average on the LAS Reading and LAS Writing. When these high-performing LEP students were redesignated (RFEP), they outperformed EO students on the LAS Reading and Writing. Further, with respect to the Stanford 9, RFEP students performed similarly to the EO students. Differences in the performances of EO, FEP, and LEP students are significant for every content area. Each Stanford 9 subtest is significantly correlated with performance on the LAS Reading and Writing for students in the EO, FEP and LEP categories.

When viewing Stanford 9 scores according to LAS Reading classifications (i.e., competent reader, limited reader, and non-reader), there is a clear distinction between competent EO, FEP, and RFEP students on the one hand, and LEP students on the

[13] The data show that although the third-grade EO students as a group performed well on both the LAS Reading and Writing, a small number of EO students nevertheless were categorized as limited readers ($n = 18$) and non-readers ($n = 6$) (see Table 2.9). For writing, 122 EO students were categorized as limited writers and two were categorized as non-writers (see Table 2.10).

other. LEP students scored considerably below students in the other categories on all sections of the Stanford 9. The distinctions between EO/FEP/RFEP and LEP are less for the limited and non-reader classifications. When viewing Stanford 9 scores according to LAS Writing classifications (i.e., competent writer, limited writer, and non-writer), we see a similar pattern emerge among competent writers; there is a noticeable difference in the performance of EO, FEP, and RFEP students and students in the LEP category. The same pattern holds in the limited writer and non-writer classifications.

Eleventh Grade

The 11th-grade students in this study were categorized as EO, FEP, and LEP. The numbers of FEP and LEP students at the 11th grade were low, which is reflective of the numbers of ELL students in the district. Because of the low numbers, the 11th-grade results must be interpreted with caution.

Descriptive statistics. Tables 2.11 and 2.12 provide descriptive statistics for the 11th-grade data. Table 2.11 presents the standard score means, standard deviations, medians,[14] and ranges for EO, FEP, and LEP groups on the LAS Reading and Writing tests.

For LAS Reading, the mean for EO students was 96.9 (*SD* = 4.7), for FEP 94.8 (*SD* = 6.0), and for LEP 85.6 (*SD* = 10.5). For LAS Writing, the mean for EO students was 81.8

Table 2.11

Standard Score Descriptive Statistics for LAS Reading and Writing by Language Proficiency Category (Grade 11)

	n	*M*	*SD*	Median	Min.	Max.
EO						
LAS-R	104	96.9	4.7	98.0	71.0	100.0
LAS-W	109	81.8	11.0	80.0	60.0	100.0
FEP						
LAS-R	28	94.8	6.0	98.0	80.0	100.0
LAS-W	29	72.9	9.1	76.0	60.0	87.0
LEP						
LAS-R	36	85.6	10.5	89.0	55.0	100.0
LAS-W	36	66.8	8.0	64.0	44.0	82.0

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient. LAS-R = LAS Reading test; LAS-W = LAS Writing test.

[14] The medians are included in Tables 2.11 and 2.12 to present a picture of how the distributions deviate from a symmetric distribution.

(*SD* = 11.0), for FEP 72.9 (*SD* = 9.1), and for LEP 66.8 (*SD* = 8.0). Just as with the 3rd graders, the 11th-grade EO students performed best in both reading and writing, with the FEP students next, followed by LEP students. Again, just as with the 3rd-grade students, all three 11th-grade groups had higher scores on the reading test than on the writing test, which suggests that the 11th-grade students in the sample are stronger readers than writers, at least in terms of the skills that are measured by the LAS. The minimum and maximum scores for each group on both tests show some range of performance within and across proficiency groups. The maximum scores for LEP students for both reading and writing indicate that some of the students in that group are performing within the competent range (80-100) established by the LAS.[15]

Table 2.12 provides the NCE means, standard deviations, medians, and ranges for the EO, FEP, and LEP groups on the Stanford 9 Reading, Math, Language, Science, and Social Science subtests. For all of the subtests, the mean for EO students was the highest, followed by FEP and then LEP students. For Reading, the mean for EO students was 57.0 (*SD* = 18.6), for FEP 37.7 (*SD* = 17.3), and for LEP 22.4 (*SD* = 10.7). For Math, the mean for EO students was 69.2 (*SD* = 22.4), for FEP 43.6 (*SD* = 19.9), and for LEP 34.9 (*SD* = 13.7). For Language, the mean for EO students was 65.0 (*SD* = 20.2), for FEP 44.3 (*SD* = 16.9), and for LEP 31.5 (*SD* = 10.1). For Science, the EO mean was 65.5 (*SD* = 20.3), for FEP 38.0 (*SD* = 17.8), and for LEP 30.6 (*SD* = 10.9). Finally, the EO student mean for Social Science is 72.9 (*SD* = 21.7), for FEP 46.0 (*SD* = 21.3), and for LEP 38.9 (*SD* = 14.8).

For all subtests, the EO student mean was above average based on the NCE norm of 50. FEP and LEP student means were all below average.[16] The gap between EO and LEP student means was large and nearly identical across all content areas: Reading (34.6 point gap), Math (34.3 point gap), Language (33.5 point gap), Science (34.9 point gap) and Social Science (34.0 point gap). These findings are not consistent with findings from the extant data analyses (Abedi & Leon, 1999; Abedi et al., 2000/2005), which showed narrower gaps in performance between non-LEP and LEP students on the content areas of Math and Science than on Social Science and Reading; however, the number of 11th-grade LEP students in this study was small and may not be reflective of a larger or different sample.

---

[15] The *LAS Examiner's Manual for Reading/Writing* (Forms 3A and 3B) provides the following competency levels: For reading—a standardized score of 0-59 = Competency Level 1, non-reader; a standardized score of 60-79 = Competency Level 2, limited reader; a standardized score of 80-100 = Competency Level 3, competent reader. The same standardized scores and competency levels apply for writing.

[16] MANOVA could not be run on the data reported in Table 2.12 due to the unequal variances coupled with unequal sample sizes. The assumption of homogeneity of variance was violated.

Table 2.12

Normal Curve Equivalent Descriptive Statistics for Stanford 9 Reading, Math, Language, Science, and Social Science by Language Proficiency Category (Grade 11)

	n	*M*	*SD*	Med.	Min.	Max.
Reading						
EO	100	57.0	18.6	57.0	6.7	99.0
FEP	29	37.7	17.3	43.0	1.0	66.3
LEP	34	22.4	10.7	22.4	1.0	42.5
Math						
EO	102	69.2	22.4	70.1	15.4	99.0
FEP	28	43.6	19.9	43.1	10.4	86.9
LEP	33	34.9	13.7	31.5	13.1	60.4
Language						
EO	101	65.0	20.2	68.5	1.0	99.0
FEP	27	44.3	16.9	48.4	1.0	74.7
LEP	33	31.5	10.1	33.0	10.4	45.7
Science						
EO	102	65.5	20.3	71.5	17.3	99.0
FEP	28	38.0	17.8	33.0	13.1	75.8
LEP	33	30.6	10.9	29.9	1.0	56.4
Social science						
EO	101	72.9	21.7	79.6	15.4	99.0
FEP	29	46.0	21.3	44.7	6.7	86.9
LEP	34	38.9	14.8	36.5	10.4	70.9

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient.

Test reliability. The reliability coefficients (internal consistency coefficients) on LAS Reading for the proficiency categories are provided in Table 2.13.

Because of the small sample sizes for the FEP and LEP groups, the two groups were combined to compute a reliability coefficient. The reliability coefficients for LAS Reading show evidence of internal consistency among the items on the test for both the

Table 2.13

Reliability Coefficients (α) for LAS Reading (55 items) by Language Proficiency Category (Grade 11)

	n	α
EO	104	.781
FEP/LEP	64	.848

EO students and the combined FEP/LEP group. Reliability coefficients are not provided for LAS Writing because one of the items on the writing test is an essay and is weighted differently from the other five items. The state's item-level data for the 1999 Stanford 9 subtests were not available; thus, the reliability coefficients could not be calculated.

Test correlations. Table 2.14 provides the Pearson Product Moment correlations for the LAS standard scores with the Stanford 9 NCEs for Reading, Math, Language, Science, and Social Science. All but one of the correlations in Table 2.14 are significant ($p < .01$). For the EO students, LAS Writing is more highly correlated with the content subtests than LAS Reading, though for Stanford 9 Reading and Language, the correlations are almost identical with LAS Reading and Writing (Stanford 9 Reading with LAS Reading, .55, and with LAS Writing, .57; Stanford 9 Language with LAS Reading, .58, and with LAS Writing, .59). For the FEP/LEP students, LAS Reading is more highly correlated with the content subtests than LAS Writing. The correlations for Stanford 9 Language with LAS Reading and Writing are almost identical, .57 and .56 respectively. The magnitude of the correlations ranges from a low of .25 (LAS Writing with Stanford 9 Math for FEP/LEP students) to a high of .67 (LAS Reading with Stanford 9 Reading for FEP/LEP students).

Summary of 11th-grade findings. Because of the small sample sizes for the FEP and LEP 11th-grade students, the types of analyses that could be performed were restricted, and consequently the findings are limited. Consistent with the results from the 3rd grade, however, the 11th-grade students, as a whole, performed better on LAS Reading than on LAS Writing, with the EO students outperforming the FEP and LEP groups on both tests. The EO group again outperformed the FEP and LEP groups on the Stanford 9 subtests. While MANOVA could not be performed on the data to check for significant differences in the means, the point differences in the means are pronounced. In addition, differences in maximum scores across the language proficiency groups highlight the range of performance captured by the content subtests. EO group performance is consistently above average, above NCE 50, for all Stanford 9 subtests. Both FEP and LEP group performance was uniformly below average across content areas with FEP performance in the high 30s to mid 40s and LEP performance in the low 20s (Reading, 22.4) to the high 30s (Social Science, 38.9).

Discussion

The guiding research question in this study asks what the relationship is between performance of ELL students on a standardized content test and a test of English

Table 2.14

Pearson Product Moment Correlations for LAS Standard Scores for Reading and Writing With Stanford 9 Normal Curve Equivalents for Reading, Math, Language, Science, and Social Science by Language Proficiency Category (Grade 11)

	Stanford 9														
	Reading			Math			Language			Science			Social science		
	n	*r*	*p*	*n*	*r*	*p*	*n*	*r*	*p*	*n*	*r*	*p*	*n*	*r*	*p*
EO															
LAS-R	90	.55	.001	92	.42	.001	91	.58	.001	92	.38	.001	91	.46	.001
LAS-W	96	.57	.001	98	.58	.001	97	.59	.001	98	.62	.001	97	.57	.001
FEP/LEP															
LAS-R	58	.67	.001	56	.36	.007	56	.57	.001	56	.53	.001	58	.46	.001
LAS-W	61	.44	.001	59	.25	.055	58	.56	.001	59	.45	.001	61	.41	.001

Note. EO = English only; FEP = fluent English proficient; LEP = limited English proficient. LAS-R = LAS Reading test; LAS-W = LAS Writing test.

language proficiency. The study is significant because it compares concurrent performance on these two types of measures. The data reported above for both 3rd-grade and 11th-grade students clearly demonstrate content performance differences based on English language proficiency as measured by the LAS Reading and Writing Components. However, as well as looking at student performance on the content assessment vis-à-vis the language test, we examined mean differences among the district-designated language groups on the content tests. This traditional approach to looking at student performance shows that for the 3rd grade, the EO students performed significantly better than the FEP and LEP students (see Table 2.4). The EO student means are above average using the national norm of NCE 50 as the benchmark; the FEP student means are approximately average, and the LEP student means are well below average. Though the group mean differences vary to some extent across content areas, the overriding trend is, as expected, that LEP students are doing less well on content tests than non-LEP students. In every language proficiency group, however, there is a wide range of scores, with some students performing well above the mean.

The same results are true for the 11th-grade students in this study. The EO students outperformed the FEP and LEP students across all content areas, and the FEP students outperformed the LEP students, with some students in every group performing well above the mean (see Table 2.12). The 11th-grade FEP students as a group, however, were weaker on the content subtests than the 3rd-grade FEP students. Still, these results reflect general performance results on standardized achievement tests used across several states (Abedi & Leon, 1999; Abedi et al., 2000/2005). LEP students as a group were doing poorly on standardized content assessments, with some individual LEP students performing at least as well as some FEP and some EO students.

When we consider the performance of LEP students who were doing well, there is reason for optimism. One group of ELL students in the third-grade sample—LEP students who were, for the analyses in this study, redesignated RFEP on the basis of their language test scores and their Stanford 9 Reading score—outperformed the EO students on the language tests (see Table 2.5) and performed similarly to them on the Stanford 9 subtests (see Table 2.6). The performance of these RFEP students suggests that when ELL student means are in the mid 90s as measured by the LAS, RFEP student performance is similar to EO performance on content tests (see RFEP group, Tables 2.5 and 2.6), suggesting that for these students,

the content assessments are likely valid measures of their content knowledge. That is, their performance is average (50) or above in terms of norm group (NCE) scores. Although the size of the RFEP group was small ($N = 40$), their performance seems to indicate that they had acquired English sufficiently well to be able to demonstrate their content knowledge through English. While this group was excluded from some analyses due to the small number of students, the mean differences in performance and the range of scores may suggest that some students who are designated LEP upon entering school are making progress in both English and content knowledge. There is, of course, the possibility that the students redesignated RFEP were misplaced upon entering school and had a high degree of English proficiency to begin with.

The performance of some third-grade EO students in the study raises questions about the criteria ELL students are being held to in order to receive FEP or RFEP status. Of the 296 EO third graders, only 140 (47%) met the same redesignation criteria being used with ELL students. This finding gives the appearance that ELL students are being held to a standard that many EO students themselves cannot reach. The findings may indicate that a large percentage of EO students are also struggling with language, OTL, or both, or that the criteria are inappropriate.

The focus of the research reported here is on the comparison of the means for the content test by LAS proficiency categories: competent, limited, and non-reader or non-writer.[17] These results (Tables 2.9 and 2.10) show that third-grade ELL students who meet the redesignation criteria and who qualify as competent readers and writers according to LAS scoring criteria score at the mean (NCE 50) or above on the Stanford 9 subtests. LEP students in the competent reader category, however, do not reach the national norm for average performance on the content test (with the exception of Math when LAS Writing is the criterion). It is possible that the LAS Reading and Writing criterion of 80 is not the appropriate language criterion for judging whether students have sufficient mastery of English to perform similarly to non-ELL students, all other factors being equal.[18] Neither LAS subtest discriminated well among students at the higher end of the LAS proficiency spectrum (see LAS scores for EO and FEP students in Table 2.3). This finding is consistent with information provided in the *LAS Examiner's Manual* (Duncan & De Avila, 1988) and

[17] The same comparisons could not be made for the 11th grade because of the small numbers of FEP and LEP students.

[18] Note that the district criteria for redesignation include the Stanford 9 Reading score.

confirms that the LAS does not discriminate within the competent range, 80-100. The lack of discrimination within the current LAS competent range is a limitation in this study because it clouds the comparison of language and content scores. That is, without a test that discriminates well at the upper range of the language proficiency continuum for the grade level, it is difficult to tell whether students who are identified as competent by LAS are, in fact, skilled enough in the language to handle the material on content assessments.

Another possible factor in the performance of LEP students who are in the competent category is related to the students' classroom experiences. These students may not be able to perform similarly to non-ELL students, not because of their language proficiency necessarily, but rather because they have not had an opportunity to learn (OTL) the content material covered on the test. LEP students are often not exposed to the same curriculum as mainstream students because their educational focus is usually on acquiring English in special programs, so regardless of improvement in language proficiency, they may not have had access to the content covered on tests such as the Stanford 9.

Socioeconomic status (SES) is another variable that impacts ELL student performance (Abedi et al., 2000/2005). However, there was not enough variability in SES for the sample in this study to allow analysis of this variable.

The study reported here suggests that there is a strong relationship between the English language proficiency of ELL students and their performance on a content assessment. However, the specifics of that relationship are not clear because the data available to date are not sufficient to determine when ELL students have adequate English language proficiency to demonstrate their content knowledge. As mentioned above, variables such as OTL and SES mitigate student performance, as do length of time lived in the United States, ability in the first language, and home language environment (not discussed here). Each of these variables is an important part of the total picture for every student (Butler & Stevens, 1997) and should be considered whenever possible in future research. In fact, research should be designed to control for these variables.

Also, an important research goal for future studies on this topic is the procurement of item-level data on content assessments. Identifying those items on which ELL students and native English speakers perform differentially would

provide us with an opportunity to examine the degree to which the language of the test might be a threat to the validity of the content assessment.

Finally, if the language tapped by the LAS and other commonly used language assessments does not adequately mirror the language used on content assessments, it is possible for ELL students to reach competent ranges on these tests without being sufficiently skilled in the more "academic" style of language reflected in content tests. Research that contributes to a better understanding of the type of language used on content tests (Bailey, 2000/2005) and the relationship of that language to the language assessed by language proficiency measures (Stevens et al., 2000) will move us closer to assuring the validity of content assessments for ELL populations.

References

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Mirocha, J. (2000/2005). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663, pp. 1-45). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L. (2000/2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663, pp. 79-100). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Duncan, S. E., & De Avila, E. A. (1988). *Language Assessment Scales (LAS) Reading and Writing examiner's manual, Forms 1A and B, Forms 3A and B.* Monterey, CA: CTB/McGraw-Hill.

Duncan, S. E., & De Avila, E. A. (1990). *Language Assessment Scales (LAS) Reading and Writing Component, Forms 1A and 3A.* Monterey, CA: CTB/McGraw-Hill.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Series, Ninth edition, Form T.* San Antonio, TX: Harcourt Brace.

Jimerson, S., & Klein, J. (1999). *ORAL-J Grades 1, 2, & 3 reliability study: Spring 1999.* Unpublished manuscript, Counseling, Clinical, and School Psychology Program, University of California, Santa Barbara.

Staley, L. (2005). *The effects of English language proficiency on students' performance on standardized tests of mathematics achievement.* Unpublished doctoral dissertation, University of California, Los Angeles.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs* (CSE Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

CHAPTER 3

LANGUAGE ANALYSIS OF STANDARDIZED ACHIEVEMENT TESTS: CONSIDERATIONS IN THE ASSESSMENT OF ENGLISH LANGUAGE LEARNERS

Alison L. Bailey[1]

Summary

One potential threat to the validity of administering standardized tests of achievement to English Language Learners (ELLs) is the fact that the language demands of the tests may exceed the English language abilities of ELL students.[2] Performance on these assessments may therefore not be an accurate reflection of the content knowledge of ELL students if students are stymied in their efforts to answer questions by the presence of construct irrelevant language. Findings from analyses of the language demands of a standardized achievement test at 11th grade (and preliminary results at 3rd grade) are presented. These analyses were conducted to describe the nature and degree to which test items in the mathematics, science, and reading comprehension subsections of the standardized test contain potential language demands for ELL students.

Specifically, we conducted a review of items to determine potential linguistic demands (excluding content-specific material such as mathematical terminology) that might constitute construct irrelevant language. This resulted in a set of evaluative criteria to identify (a) site of difficulty in test items (stimulus passage, stem and/or response options), (b) language domain (vocabulary, syntax and/or discourse), and (c) type of linguistic demand (e.g., uncommon vocabulary, atypical parts of speech, idiomatic language). We also developed a Likert scale for language demand to rate the degree of difficulty of test items from low to high.

---

[1] I wish to thank Frances Butler, Richard Durán, Martha Castellon-Wellington, Anthony Friscia, Jim Mirocha, Robin Stevens and David Sweet for helpful comments and suggestions on this chapter, and Ani Moughamian, Seth Leon, and Rebeca Fernandez for research assistance.

[2] Language demand, for the purposes of this chapter, is being defined as construct irrelevant language that reflects an unusual or unnecessary level of linguistic sophistication. The evolution of this working definition of language demand and its operationalization will be discussed in greater detail later in the chapter.

We found that test items on the 11th-grade mathematics and science subsections of the standardized test included general vocabulary that was evaluated as uncommon or used in an atypical manner, 60% and 75% for mathematics and science items, respectively. A slightly lesser percentage of items on both subsections contained syntactic structures that were evaluated as complex or atypical constructions. Just a quarter of the items contained discourse demands. The reading comprehension subsection contained high percentages of items with vocabulary and syntax demands, but more than half of the items in this subsection also had discourse-level demands. In addition, the results of the linguistic demand ratings found that the reading comprehension items contained a higher degree of difficulty in vocabulary and syntax compared to the items in the mathematics and science subsections. This is consistent with our findings from the extant data analyses (Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2000/2005; Butler & Castellon-Wellington, 2000/2005), which show that there is a larger gap in test performance between ELL students and English proficient students from various school districts on the reading comprehension subsections of standardized content assessments than on the mathematics and science subsections. The findings are generally replicated in a preliminary evaluation of the language demands of test items on a 3rd-grade assessment**.** Differences between ELL and non-ELL student performance on 3rd-grade reading and math items were correlated with language demand ratings of the items. These correlations are only suggestive but provide a framework for a potentially fruitful avenue of research.

Introduction

In this chapter we first present rationale for the language analysis of standardized achievement tests. Specifically, we discuss how academic language at lexical, syntactic, and discourse levels may impact the test performance of ELL students. Next, we describe the development of language demand rating scales. The rating scales were designed to target language that was not content-specific (e.g., ignoring specialized mathematics vocabulary) but was still likely acquired in an academic context rather than in less formal environments. Evaluation criteria were devised by which we could rate each test item on the reading comprehension, mathematics, and science subsections of a standardized achievement test for the 11th grade, with replication at the 3rd grade for reading comprehension and mathematics only.

Rationale for Language Analysis of Standardized Achievement Tests

As part of the larger initiative of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) to investigate the validity of assessing ELL students, we decided to make a closer examination of the language of standardized achievement tests. It is actually the concern that such tests may place too great an English language demand on ELL students, a demand that gives rise to a threat to the validity of administering standardized achievement tests to ELL students. Consequently, performance on these tests may reflect the English language abilities of ELL students rather than their knowledge of the content material the tests are designed to measure (e.g., mathematics skills, scientific knowledge, etc.).

Moreover, larger differences in performance between ELL students and English proficient students have been found in reading comprehension than in mathematics and science subsections of the Stanford 9 (Harcourt Brace Educational Measurement, 1996) and the Iowa Tests of Basic Skills (ITBS; Hoover, Hieronymus, Dunbar, & Frisbie, 1996) in extant data from various school districts nationwide (see Abedi & Leon, 1999; Abedi et al., 2000/2005; Butler & Castellon-Wellington, 2000/2005). This, in itself, suggests that the greater English language load of the reading comprehension subsections presents a barrier to the performance of ELL students. The analysis and findings reported here will provide a more refined picture of how the language of the different subsections differs and likely impacts student performance across the content areas. Indeed, the results of the linguistic analysis of the content tests have subsequently played a role as one of five types of evidence used in operationalizing academic language (see Bailey & Butler, 2002/2003, 2004, forthcoming).

Our quantitative analysis of items may share a superficial resemblance to prior research in other domains of psychometric research. One domain of research has examined test item difficulty by examining the number of students able to answer particular items correctly. However, we rate the difficulty of a test item's degree of language demand—the language of the test item itself. Thus, our rating of difficulty is not based on student performance on the test item. Our analysis also shares much in common with the adaptations made to existing assessments normed for one age group for eventual use with another age group. The linguistic adaptations that are made in order to make an assessment age-appropriate in such circumstances are among the language demands with which we will be concerned (e.g., familiarity of vocabulary). However, we also pay careful attention to culturally appropriate

language such as the culturally embedded uses of language (e.g., idioms and metaphor) that have been argued to be less familiar to ELL students (e.g., Montero, 1993).

Defining Academic Language

Our first undertaking in attempting to quantify the language demands of standardized achievement tests required that we define academic language. This type of language, be it at the lexical (vocabulary), syntactic (grammar), or discourse level, was the target of our analysis and stands in contrast to both the specialized content-specific language, such as the conceptual terminology of mathematics (e.g., parallelogram, velocity, equation), and the everyday informal speech that ELL students may acquire outside the classroom environment. Rather, academic language is a mode of communication (spoken/written) that is not specific to any one content area, but is nevertheless a register or a precise way of using language that is often specific to educational settings. For example, formal vocabulary, such as *examine* and *cause,* that children encounter at school contrasts with everyday vocabulary, such as *look at* and *make,* that they encounter in less formal settings (Cunningham & Moore, 1993). The distinction between informal and formal oral language is one made by Cummins (1980) and can be described as the difference between Basic Interpersonal Communication Skills (BICS), acquired and used in everyday interactions, and Cognitive Academic Language Proficiency (CALP), acquired and used in the context of the classroom. Shefelbine (2000) has made academic language one of four necessary components in a model of the reading acquisition process, along with decoding skills, reading fluency and comprehension strategies. The Cognitive Academic Language Learning Approach (Chamot & O'Malley, 1994, 1996) is a program that operationalizes CALP with ELL students and offers the following definition of academic language, which is the one we adopt in this analysis of test item language. Academic language, according to Chamot and O'Malley (1994) is

> the language that is used by teachers and students for the purpose of acquiring new knowledge and skills . . . imparting new information, describing abstract ideas, and developing students conceptual understanding. (p. 40)

To this definition we add two features. First, academic language implies the ability to express knowledge by using recognizable verbal and written academic formats. For example, students must learn acceptable, shared ways of presenting

information to the teacher so that the teacher can successfully monitor learning. These formats or conventions may or may not be explicitly taught as part of a curriculum, but their use is expected of all students. Second, academic language is most commonly used in decontextualized settings. These are settings where students do not get aid from the immediate environment in order to construct meaning. There is little or no feedback on whether they are making sense to the listener or reader, so students must monitor their own performance (spoken or written) based on abstract representations of others' knowledge, perspectives, and informational needs (e.g., Snow, 1991).

Thus, students learn to recognize and make sense of the varied conventional ways of presenting academic material in decontextualized settings. For example, test items often present students with sentence fragments either in the question stem or in the answer options that require students to be familiar with sentence completion as a test item format. We argue that the test-taking situation is the epitome of the academic decontextualized setting requiring academic language proficiency. The test-taking routine is a conventional script with specific structures that need to be learned, and during the test students obviously receive no feedback from the test writer, the grader, or their teacher.

Developing Language Demand Rating Scales

Operationalizing Language Demand

We turn now to the development of a rating scale for assessing the language demands of standardized content assessments in the domains of vocabulary, syntax, and discourse. The process of operationalizing and reliably identifying different degrees of language demand in each of the three domains has been, and continues to be, an extremely complex issue for this area of the validity study. Our definition of language demand as construct irrelevant language that reflects an unusual or unnecessary level of linguistic sophistication is a working definition that has evolved as we have read the available literature and solicited and received input from various colleagues in the field. While we acknowledge that it is difficult to objectify the language demands of test items because it requires us to quantify linguistic features in terms of levels of processing demand, we had at least two guiding criteria.

First, there are linguistic complexities in some items that are not present in others. For example, complex clausal structures can be rated as more linguistically

demanding than simple clausal structures. Thus, items that include complex clauses will be rated higher on language demand in the syntax domain than those items that do not include complex clauses. This approach to operationalizing language demand is independent of the level of language proficiency of the students taking the test; a complex clause is a complex clause regardless of the individual reading the clause. However, it will be a barrier to comprehension if the individual's language proficiency is not at a level to process complex clauses. This is more likely to be the case for an ELL student than a native English-speaking student. For example, an ELL student who is a reasonably proficient speaker of everyday (BICS) English but who may not have had as extensive an exposure to complex syntax, idioms, and depth of vocabulary (e.g., antonyms, synonyms, etc.) as a native speaker of English of the same age may find some test items more challenging because his or her language proficiency level may not match the demands of the language on these items. An ELL student may easily read an interrogative sentence such as "Who do you think will win the game?" because it is the sort of language that he or she may encounter in widely read materials, such as newspapers and magazines. In other words, this form of written language more closely resembles the language a student hears in everyday speech. However, the same request for information may be conveyed in very different language in a standardized content assessment. For example, the following question is fictitious, but indicative of the linguistic complexity of test items: "What is your best estimate of which of the teams will win?" This version of the question includes not only the unfamiliar use of "best" (to mean "most accurate"), but also an embedded wh-question in the second clause: ". . . which of the teams will win?"

Second, we also had in mind the extraneous use of language in test items that may only add to the linguistic processing load and not to a student's understanding of a test item. We therefore include in our operational definition of language demand the sort of language that may in fact not be useful to any reader, ELL or non-ELL, but that may be more problematic to ELL students in a test situation because they may read at a slower pace than native speakers of English. In the example given above, the use of "What is your best estimate . . ." may be unnecessarily complex; the question may be more straightforwardly expressed as "Which team is likely to win?"

The process of refining the definition of language demand will continue with further examination of content assessments and further examination of the literature

in this area. Most recently we have conducted observations of language use in science classrooms and extensive linguistic analyses of textbooks in order to establish language-based profiles for the different content areas specifically at fifth grade (Bailey, Butler, LaFramenta, & Ong, 2001/2004; Bailey, Butler, Stevens, & Lord, forthcoming; Butler, Bailey, Stevens, Huang, & Lord, 2004). We hope that our efforts in this area will afford the opportunity for us to make criteria available to help others in the area of academic English test development, curricula development, and future research studies that require evaluation of the language of test items.

Procedures

The three subsections of the standardized content assessment we analyzed for language demand comprised approximately 40 to 60 items per subsection (see Table 3.1). The assessment included a reading comprehension section, with different stimulus passages using authentic published texts, both narrative and expository in nature; a mathematics section, with questions using mathematical formulas with some attendant language,[3] and questions with language-rich problems set in the context of everyday activities; and a science section, with items that varied in their use of formulas, lists, visual stimuli, and language-rich problems set in the context of everyday activities.

To make an evaluation of the language demands on these three content areas, we proceeded through the following three steps:

Table 3.1

Subsections of the 11th-Grade Standardized Assessment Examined

Reading	Reading comprehension (reading passages consisting of autobiography, expository and literary texts)
Mathematics	Mathematical concepts and problem solving
	Mathematical computation[a]
Science	Science (consisting of life, earth, and physical science topics)

[a] The mathematical computation subsection was examined and all items were rated as having no language/no language demands at all. Therefore we excluded this mathematics subsection from further analysis and discussion.

[3] These items require some language processing, unlike items in the separate mathematical computation subsection, which is comprised almost exclusively of mathematical formulas and no language.

Step 1: We conducted an initial reading of all test items to determine the range of potential linguistic demands placed on students.

Step 2: We developed a qualitative coding scheme to identify the following:

1. site of difficulty in item passage: stimulus passage, stem and/or response options;

2. affected language domain: vocabulary, syntax and/or discourse; and

3. specific type of language difficulty: for example uncommon vocabulary, atypical parts of speech, non-literal use of language (see Appendix 3.A for entire list of types of demand). The types of demand on comprehension of test items were derived from the text readability literature (e.g., Noonan, 1989), the literature on the language of mathematics (e.g., Mestre, 1988; Nesher & Katriel, 1986; Saxe, 1988; Spanos, Rhodes, Dale, & Crandall, 1988), and with the help of the project advisory board that contained educationalists, applied linguists, Chicano studies researchers, and experienced bilingual education teachers (see also Abedi, Lord, & Hofstetter, 1998).

Step 3: We developed a language demand rating scale for test items with difficulties identified in Step 2. The domains of vocabulary and syntax were rated 0 = no/low demand, 1 = some demand, 2 = moderate demand, and 3 = high demand for all three of the subsections. Connected discourse was rated 0 for absent or 1 for present in the test items on the reading comprehension and mathematics subsections, but rated 0, 1, or 2 for test items on the science subsection (see Appendix 3.B). The latter reflects a distinction in the science items between the absence of discourse-level demands in a test item at one extreme, and connected or extended discourse at the other extreme (e.g., use of anaphoric reference, temporal and causal connectors that are the hallmarks of extended discourse), with information presented in multiple sentences (e.g., unrelated lists) using no intersentential connectors as the intermediate level of discourse style. This three-way distinction was prevalent only in the science subsection because of the nature of science items.

Reliability was calculated as a percentage of exact agreements (the number of rating agreements divided by the number of agreements and disagreements) between two independent coders. The percentage of agreements across content areas (calculated on just three of the six reading passages due to the use of the remaining three passages for training) and language domains ranged from 60% to 100%. There were differences in the degree of agreement between coders across the

different language domains. Discourse was the most reliably rated language domain, with reading, math and science subsections having exact agreement of 72%, 100%, and 86%, respectively. The syntactic domain in the reading, math, and science subsections had exact agreements of 66%, 75%, and 86%, respectively. The vocabulary domain had exact agreements for the reading, math, and science subsections of 72%, 60%, and 80%, respectively. Syntax and vocabulary ratings were likely less reliable than discourse ratings because they had more gradation within the rating schema.[4]

We also see from these percentages that, overall, the science subsection of the test was the most reliably rated across all three language domains, reaching the desirable 80% threshold in all domains, whereas the reading and math subsections were both less consistently rated across the language domains. The discourse domain in the math subsection, however, was scored in total agreement, most likely because discourse is less commonly used in math items so its presence, when it is used, is very salient to coders. It is possible that reliability for the reading subsection was somewhat low because of the large amount of language to be rated and because the content-specific vs. construct irrelevant dichotomy is less obvious in the context of the general interest reading passages that were employed on the reading subsection. All disagreements were resolved by consensus between the two coders before further analyses were performed. However, a future goal is to achieve greater specificity in the rating guidelines. This will allow for improved reliability between raters in further evaluations of test items.

Potential Language Difficulties for ELL Students: Example of a Fictitious Test Item

The following example of a test item with potential language demands is fictitious and is provided for illustrative purposes only. Though not an actual item from the test, this "dummy" item is representative of the types of items we analyzed and provides a comparable level of language demand to that found on actual items.

> Mice were randomly assigned to two diet regimens by a biologist working in his lab. Altogether he tended 14 animals. However, he raised five mice with low protein and nine with normal levels of protein. Then, as he fed them, he

[4] The proportions of exact agreements between two codes for rating the third-grade math section were 70%, 79%, and 96% for vocabulary, syntax, and discourse, respectively. Coding of the reading comprehension section was done by consensus because this section was used in development and training.

monitored their health. After just three days, five of the mice began to grow sick. The biologist concluded that lack of protein had reduced the immune systems of these mice to a level subject to disease."

1. **Vocabulary and syntax demand:** *lack of protein had reduced the immune systems of these mice to a level subject to disease.* The meaning of the word "subject" in this context is uncommon, used to mean "left open to," rather than its more common meaning—the content of a class (e.g., "the subject today was science and technology"). The word "subject" is also a syntactic demand in that it is used as a verb in this sentence structure rather than as a noun, which is its more typical part of speech.
2. **Complex syntactic demand:** *Then, as he fed them, he monitored their health.* This is just one example of a syntactic complexity in this passage. The "left-branching" of the sentence construction may prove to be a demand on students expecting English sentences to follow the less complex subject-verb-object word order, rather than the initial adverbial clause found here before the main clause.
3. **Discourse demand:** In this stimulus passage, the reader must make connections across several utterances to create meaning. The use of cohesive ties, such as the pronoun "he," to refer back to previously introduced nouns, namely "the biologist," and the use of logical and temporal connectors such as "then" and "however" each require the reader to make meaningful connections between the information presented in a new sentence and information already presented in prior sentences. Thus, such features of connected discourse increase the language processing demands.

Results of the Language Demand Analyses

First, we report the percentage of items with identified language demands by language domain for each of the three subsections of the assessment we examined. Second, we report the mean difficulty rating each language domain received, again separately by subsection. Third, we describe results of the same analyses conducted on the third-grade-level math and reading subsections. Finally, we report the correlations between the item-level difference scores for ELL and non-ELL students' performance at the third grade and the item difficulty rating we assigned to the corresponding third-grade items.

Prevalence of Language Demands in Test Items

Mathematics and science subsections. Figure 3.1 shows the percentage of test items on the three subsections that contained language demands in the areas of vocabulary, syntax and discourse. Approximately two thirds to three quarters of the test items on the mathematics and science subsections, respectively, had general vocabulary rated as uncommon or used in an atypical manner. Note that this is not

content vocabulary specific to the fields of mathematics and science. For the purposes of this analysis, we assume that such specialized content vocabulary has been, or should have been, taught explicitly to all students, both ELL and English proficient. However, we acknowledge that having the opportunity to learn all content material, including the necessary content-specific vocabulary, may be less assured for ELL students because they may be taught at a slower pace than English proficient students.

In Figure 3.1 we also see that one half to two thirds of test items in the mathematics and science subsections have syntactic structures evaluated as complex or atypical in their construction. Connected discourse demands are not as prevalent in test items, with only about one quarter of items presenting students with discourse-level processing demands. However, in the case of discourse demands in the science subsection, we have additional rating information because of the 0, 1, 2 rating scale that reflected language demands beyond the level of the sentence but without connected discourse (e.g., synthesis of information presented in a list format). When the 16 test items that were rated as 1 (i.e., non-connected discourse) are combined with those rated as 2 (connected discourse), the percentage of science items given a discourse demand rating even of a minimal sort increases from 24% to 56%.

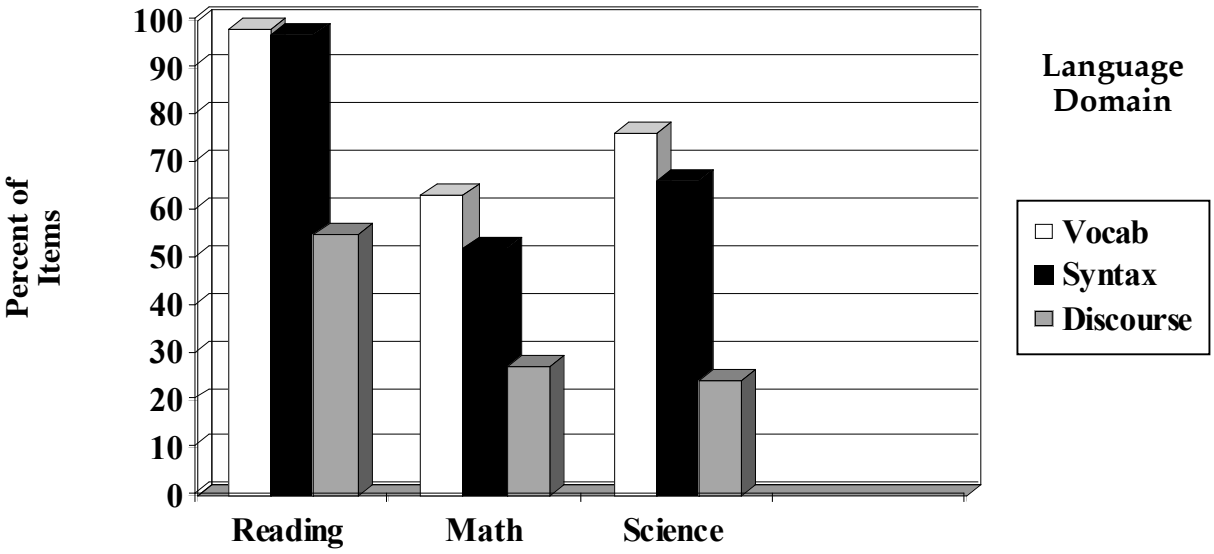


Figure 3.1. Percentage of items with language demands across content areas by language domain.

Reading comprehension subsection. Vocabulary and syntax demands were common to most test items in the reading comprehension subsection. However, in contrast to the mathematics and science subsections, more than half of the reading comprehension items also had connected discourse-level demands that require students both to process multiple clauses to extract meaning and to make sense of information presented in less familiar print genres (e.g., autobiography that may be familiar from history and social science but may not be commonly encountered outside the classroom environment).

Severity of Language Demands

Figure 3.2 shows that reading comprehension test items were rated as containing a higher degree of language difficulty compared with the mathematics and science items. That is, not only do more items contain language demands in the reading comprehension subsection, as shown in Figure 3.1, but those demands are rated as much more difficult. In the domains of both vocabulary and syntax, the mean difficulty rating is approximately 2 (0-to-3 scale) on the reading comprehension subsection, whereas the mean rating for these two domains on the mathematics and science subsections is approximately 1.[5]

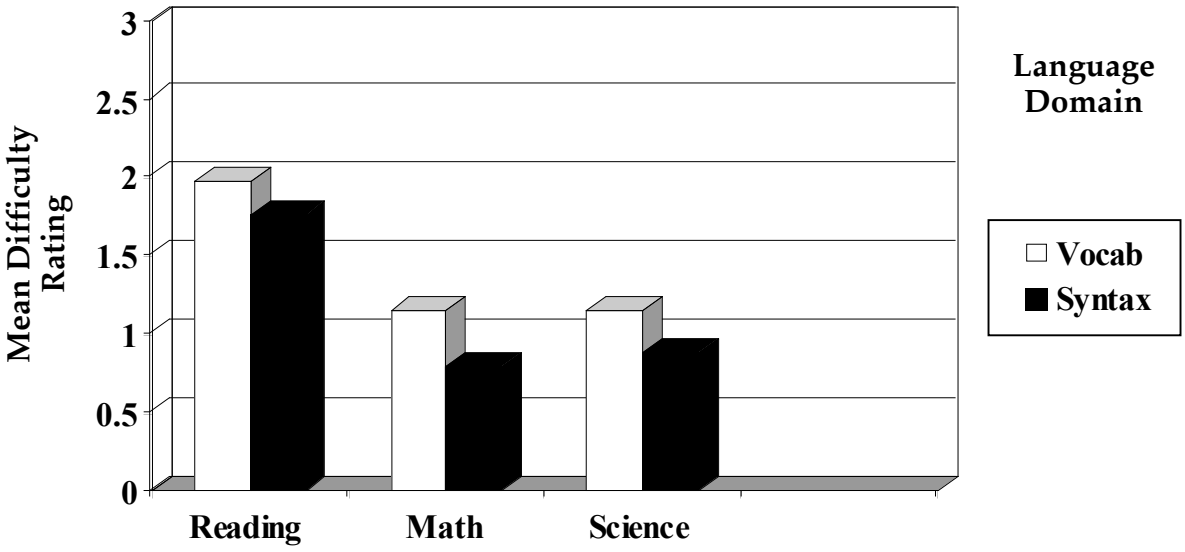


Figure 3.2. Mean difficulty rating of items across content areas by language domain.

[5] The discourse domain does not yield a mean difficulty score different from the percentage score given in Figure 3.1 due to the binary nature of the scale.

Language Demands in Math and Reading Test Items at the Third Grade

We also have preliminary results of an evaluation of the language demands of the math and reading subsections of a 3rd-grade content assessment.[6] These results generally replicate those found with the 11th-grade test items, with the reading subsection containing more items with vocabulary, syntax and discourse demands than the math subsection. The patterns within subsections were similar across the two grade levels with the notable exception of discourse demands that were found in the vast majority (92%) of reading subsection test items on the 3rd-grade test but in only approximately half of the test items on the 11th-grade test.

In terms of severity of language demands in the vocabulary and syntax domains, the math subsection had mean difficulty ratings of .64 and .46 respectively, and the reading subsection had mean ratings of 1.25 and .97 respectively. This pattern of difference between the two subsections replicates the pattern obtained with the 11th-grade test items.

Differential Item Performance by Third-Grade ELL and Non-ELL Students and Language Demand Ratings of Items

We conducted preliminary correlational analyses of the linguistic demand ratings of test items and the mean difference in item-level performance of third-grade ELL students and non-ELL students on the math and reading subsections of a standardized test of achievement.[7] We hypothesized that those items that most differentiated the performance of the two types of students would have greater language difficulty ratings and those items where there was little difference in scores between ELL and non-ELL students would have lower difficulty ratings. The preliminary results showed significant correlations in just two areas. First, there was a correlation between discourse demand and the difference in performance of ELL and non-ELL students in the math subsection ($r = .32$, $p = .02$), suggesting that when math items require language processing beyond the level of the sentence, ELL students have a more difficult time accurately answering the items in comparison to non-ELL students. Second, there was a significant negative correlation between vocabulary demands and the difference in performance of ELL and non-ELL students in the reading subsection ($r = –.40$, $p = .02$). This latter finding suggests that

[6] The percentage of items with vocabulary, syntax, and discourse demands in the math and reading subsections was 57%, 45%, 32% and 83%, 78%, 92%, respectively.

[7] These analyses were not conducted with 11th-grade test items because item-level performance data at the 11th grade were not available.

when vocabulary demands are high in a test item, the difference in performance between ELL and non-ELL students is reduced. The reduction of this gap is due to non-ELL student performance also being adversely affected by such test items.

These findings must be viewed with caution because we suspect the correlations between language demand and performance differences may be more extensive than found here due to the fact that the third-grade performance data available to us showed little difference between ELL and non-ELL students in terms of overall performance, unlike the much larger differences at other grade levels reported in chapters 1 and 2 of this report (Abedi et al., 2000/2005; Butler & Castellon-Wellington, 2000/2005). Moreover, the restricted range of the language demand rating scale may have impacted the calculation of the correlations. Therefore, in future analyses we will conduct correlations between student performance differences and more finely differentiated language demand ratings at additional (likely higher) grade levels as individual item-level performance data become available to us.

Conclusions and Implications

The findings of the language demands analysis are consistent with a reported larger difference in standardized test performance between ELL students and English proficient students on reading comprehension subsections than on either mathematics or science subsections (Abedi & Leon, 1999; Abedi et al., 2000/2005; Butler & Castellon-Wellington, 2000/2005). Most obviously, mathematics and science items often require less language processing due to greater utilization of numerical and visual stimuli. Mathematics and science items also often contain less demanding language, compared to the figurative uses of language found in literature. However, our language demands analysis reveals greater specificity about why a difference may exist between ELL students and English proficient students and between the different content areas. First, while all three content areas presented challenging syntax and vocabulary in the majority of the test items, reading comprehension did so in almost every item. Moreover, reading comprehension requires the student to process connected discourse in many more of the test items than did either the mathematics or science subsections. In addition to these differences across content areas, we found that the syntactic and vocabulary demands of reading comprehension items were actually greater in difficulty than those of the mathematics and science subsections.

Although discourse demands were not as prevalent as the other language demands examined, it is interesting to note that they do exist in all three content areas. This, coupled with the finding that syntactic, not simply vocabulary, demands play a prominent role in all content areas, has led us to begin expanding the notions behind approaches to accommodation strategies with ELL students. We suggest exploring ways to broaden the focus of accommodation strategies that traditionally address the vocabulary demands of standardized tests, by including the study of other types of language demands as we continue to gather data from a controlled study of the provision of dictionaries and extra time.

One approach to explore through future research is academic language instruction to familiarize students with the specialized vocabulary, syntactic structures, and connected discourse skills likely necessary for success on standardized content assessments. Academic proficiency instruction will provide students with the formal English language abilities not likely to be found outside the classroom environment, nor to be taught as part of content area classes. It is even conceivable that within the formal setting of the classroom, academic language may not often be modeled during instructional activities. Teachers' own oral registers may remain fairly informal (personal communication, Martin Murphy, 26th April 2000). Gee (1990) has also pointed out the limitations to academic language acquisition within classrooms because children are often not given sufficient opportunity to use scientific language themselves. Reliable exposure to academic language may therefore only be incidentally afforded through reading academic texts and other printed materials. The development of academic language tasks for use in assessment and instruction with ELL students has been proposed elsewhere (Butler, Stevens, & Castellon-Wellington, 1999). The findings of the language demands analysis suggest likely merit in explicit instruction (e.g., having students construct their own everyday and academic versions of the same concepts), as well as in assessment of academic language proficiency itself to determine whether ELL students are indeed linguistically equipped to succeed on standardized content assessments independent of their content area knowledge.

Finally, the development of criteria for identifying the nature of language demands in written texts more broadly (assessments, textbooks, media products) could also be of value to test and curricula development. For example, our work could prove useful to such organizations as the Council of Chief State School Officers, which is involved in writing practical guidelines for the development of

content assessments used for testing ELL students (Kopriva, 1999). Most imminent, we see our work being utilized in the development of an assessment of academic language proficiency. Such an assessment could be used to identify performance at various levels of language demand, which in turn could be used to match students with the appropriate achievement tests in terms of language demand level—the level having been independently established for all such standardized achievement tests.

References

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Mirocha, J. (2000/2005). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663, pp. 1-45). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., & Butler, F. A. (2002/2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of English language learners K-12. *Language Assessment Quarterly, 1*, 177-193.

Bailey, A. L., & Butler, F. A. (Forthcoming). A conceptual framework of academic English language for broad application to education. In A. L. Bailey (Ed.), *Language demands of students learning English in school: Putting academic language to the test.* New Haven, CT: Yale University Press.

Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2001/2004). *Towards the characterization of academic language in upper elementary science classroom* (CSE Tech. Rep. No. 621). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., Butler, F. A., Stevens, R., & Lord, C. (Forthcoming). Further specifying the language demands of school. In A. L. Bailey (Ed.), *Language demands of students learning English in school: Putting academic language to the test.* New Haven, CT: Yale University Press.

Butler, F. A., Bailey A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (CSE Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F., & Castellon-Wellington, M. (2000/2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663, pp. 47-77). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F., Stevens, R., & Castellon-Wellington, M. (1999). *Academic language proficiency task development process* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language approach.* Reading, MA: Addison-Wesley.

Chamot, A. U., & O'Malley, J. M. (1996). The cognitive academic language learning approach: A model for linguistically diverse classrooms. *The Elementary School Journal, 96,* 259-273.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14,* 175-187.

Cunningham, J. W., & Moore, D. W. (1993). The contribution of understanding academic vocabulary to answering comprehension questions. *Journal of Reading Behavior, 25,* 171-180.

Gee, L. (1990). *Social linguistics and literacies: Ideology in discourses*. London: The Falmer Press.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Series. Ninth edition, Form T.* San Antonio, TX: Harcourt Brace.

Hoover, H. D., Hieronymus, A .N., Dunbar, S. B., & Frisbie, D. A. (1996). *Iowa Tests of Basic Skills, Form M.* Chicago, IL: Riverside Publishing.

Kopriva, R. T. (1999). *Ensuring accuracy in testing for LEP students: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Mestre, J. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 201-219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Montero, E. (1993). *Linguistic and cultural influences on differential item functioning for Hispanic examinees in a standard secondary level achievement test.* Unpublished doctoral dissertation, Florida State University.

Nesher, P., & Katriel, T. (1986). Learning numbers: A linguistic perspective. *Journal for Research in Mathematics Education, 17,* 100-111.

Noonan, J. (1989). Readability problems presented by mathematics text. *Early Child Development, 54*, 57-81.

Saxe, G. B. (1988). Linking language with mathematics achievement: Problems and prospects. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 47-62). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shefelbine, J. (2000, February). *The role of academic language in reading and comprehending.* Presentation at the Seeds University Elementary School, Winter Quarter Colloquium Series, University of California, Los Angeles.

Snow, C. E. (1991). Diverse conversational contexts for the acquisition of various language skills. In J. Miller (Ed.), *Research on child language disorders* (pp. 105-124). Austin, TX: Pro-Ed.

Spanos, G., Rhodes, N., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix 3.A

Evaluation Criteria

1. Source in test	2. Language domain	3. Type of demand
Stimulus Passage	Vocabulary	Uncommon usage
Question		Nonliteral usage (idioms)
Answer		Manipulation of lexical forms
	Syntax	Atypical parts of speech
		Uncommon syntactic structures
		Complex syntax
		Academic syntactic format
	Discourse	Uncommon genre
		Need for multi-clausal processing

Appendix 3.B

Language Demand Rating Scales

Vocabulary

Compose a sentence using the target word in a non-content-specific context. Next, judge the word to be uncommon or not. If deemed uncommon still in everyday speech, list the word as a potential demand. Scores should be given as follows:

Number of uncommon words	Score
1–2	1
3–4	2
5+	3

Additionally, words that have multiple meanings or lexical forms that have been manipulated should be similarly evaluated as potential demands. For example, depth of vocabulary is often needed for answering items, as in the case of synonymous or near synonymous usage and similar classes of lexical items (i.e., feelings, attitudes). These receive one point each as an uncommon word.

Syntax

Consider whether each sentence is written in the clearest possible way. Locate specific syntactic issues (e.g., left branching, multiple clauses, extraneous clauses) that may impose demands. Sentence fragments, with the exception of one-word fragments used as sentence completions for test items (i.e., function like a cloze test in sentence final position) may impose a demand. Each instance should be scored as follows:

Number of syntactic words	Score
1	1
2	2
3+	3

Discourse

Consider whether or not the student is required to synthesize information across sentences. Scores should be given as follows:

- Single sentence question = 0
- Required to make clausal connections between concepts and sentences = 1

Science discourse scoring only (see previous discussion of rationale for 3-point science discourse scale):

- Single sentence question = 0
- Presentation of sequential facts with no synthesis required = 1
- Required to make clausal connections between concepts and sentences = 2

CHAPTER 4

GENERAL DISCUSSION AND RECOMMENDATIONS

Alison L. Bailey, Frances A. Butler, and Jamal Abedi

Summary

In this final chapter of the report we summarize what we have learned from the three previous chapters regarding the validity of administering large-scale content assessments to students who are English language learners (ELLs). First, we discuss the question of validity itself and how the research conducted here may or may not provide answers due to the complex nature of both the ELL populations we studied and the data available. Second, we discuss in more detail the technical concerns that these studies raise, including the statistical limitations of the studies, definitions of ELL student populations, and availability of appropriate data. Finally, we provide recommendations for the assessment of ELL students and recommendations for future research in this area.

Issues of Validity

The goal of the efforts reported in this document was to explore the technical issues of validity around the use of large-scale content assessments with English language learners. Each of the chapters in this report has as its basic focus the role of language in standardized content assessments. The question of whether assessing ELL students with large-scale content tests is a valid practice is one that many school districts have asked with no definitive answers provided to date. Numerous ELL students have been excluded from large-scale content assessments in the past because the validity of administering such assessments to these students was called into question. Specifically, the language demands of the assessments may be so great for these students as to invalidate the assessment of content knowledge. Ultimately, we have not known whether the performance of ELL students primarily reflects their language abilities or their content knowledge.

In chapter 1 of this report, we found that ELL student performance suffers in those content area subtests that are thought to have greater language complexity than others.[1] These findings suggest that student language proficiency impacts

[1] Chapter 3 discusses types of language complexities present in content assessments.

performance on standardized content assessments according to the nature of the English language demands of the content area assessed. Though this finding is not surprising and has already been widely assumed, the studies in chapter 1 provide statistical evidence, across multiple school districts and across multiple states, of weaker ELL student performance in contrast with much higher English only (EO) student performance in general and in content areas that have greater language demands.

The study in chapter 2 allowed us to examine student performance on language proficiency assessments and concurrent performance on content assessments. It has provided baseline data for identifying a threshold of language ability needed to determine whether ELL students' content assessment performance is considered a valid measure of their content knowledge. We saw evidence of some ELL students, those designated in this report as fluent English proficient (FEP) and as redesignated fluent English proficient (RFEP), performing on a par with EO students (50 Normal Curve Equivalent [NCE] or above) on the content assessment subtests, suggesting that for those students, performance on the content test reflected their content knowledge. Other ELL students, however, those designated as limited English proficient (LEP), while scoring in the competent range on the language assessment, did not score on a par with EO students on the content assessment subtests. These results suggest that further differentiation of language performance in the upper proficiency range will help to determine whether these particular students are struggling with language, content, or both.

Chapter 3 in this report has provided greater detail about the nature of the language demands in different content areas on large-scale assessments. More specificity about what constitutes a language demand will enable us to identify test items that may not be valid with ELL students who have limited English proficiency. This line of research, it is argued here and elsewhere (e.g., Stevens, Butler, & Castellon-Wellington, 2000), can also inform test development and student instruction. For example, development of a language test that emphasizes the academic language needed for accurate assessment of content knowledge could be used as an indicator of ELL readiness to take content tests. That is, for students who perform well on a measure of academic language, performance on content assessments is likely to be valid, providing opportunity to learn is not an issue

Technical Concerns

The complexity of issues concerning the assessment of ELL students is evident from the literature and from the data presented in this report. It is clear that a combination of factors and variables interact in these students' assessment. These interactions make the assessment outcomes difficult to interpret. In more technical terms, we believe that the interpretation of assessment outcomes for ELL students is confounded by background variables including language proficiency—the focus of this report. We demonstrated that multiple variables, including level of English language proficiency, student ethnicity, parent education level, and family income level, are significant predictors of ELL student performance in content areas. We know that all of these predictors are correlated, but due to the high level of confounding in the data available to us, we do not know how much unique contribution each variable has and how important each is in the assessment of ELL students. We note, however, that several of these variables impact EO student performance as well.

In order to better understand the roles of the multiple variables that affect student performance, we need access to valid, complete, and reliable data. Though data available to us for the work reported here have been useful in answering some of our research questions, limitations to these data curtailed our ability to thoroughly explore all of the trends that emerged. Among the limitations with the existing data are these:

1. The lack of uniformity in defining ELL students. Terms such as ELL, FEP, LEP, and bilingual are used in the national dialogue about students who are acquiring English as a second language. Unfortunately, these terms are often operationalized differently across school sites within a district, across districts, and across states, causing difficulties with respect to data interpretation. For example, some districts and states have redesignation criteria that are based on different measures or different cut scores. Furthermore, students are not redesignated at the same time during the school year across districts and states. Therefore, student designations may not be accurate at the time research data are compiled or collected

2. The lack of comprehensive data sets. Often existing data files do not include important data elements such as student ethnicity, parent education level, and family income because the data were not collected for research purposes. In addition, item-level data are often not available.

3. Limitations regarding the aggregation of small numbers of ELL students across districts. Though we are interested in ELL/EO student comparisons at the national level, the variability in student background variables and the

designation criteria across districts do not allow us to combine data sets in order to make large-scale comparisons. This issue is even more critical when studying assessment issues by subgroups of ELL students.

4. The limitations of language assessments. A major weakness in the study of ELL student assessment is the lack of a standard instrument that can be used to assess English language proficiency in a way that is parallel to the way language is used on the content assessments. The content of currently available commercial language proficiency tests may not be adequate to measure the level of language proficiency necessary for standardized achievement tests.

Recommendations for Assessment

Currently there are two approaches widely taken for the use of large-scale content assessments with ELL students. The first is to exclude ELL students from testing; the second is to include them in the testing process knowing that the interpretation of their test scores may be problematic. The first approach, in our view, is unacceptable because if ELL students are not tested, information on their achievement is, in effect, absent from any decision making that impacts their school careers. The results of the research reported here suggest that there is reason for concern with the second approach; we propose two, not mutually exclusive alternatives that would serve to make the second approach more viable. The first is the identification of an English language learner validity threshold through the use of a metric for defining the language proficiency of ELL students. This alternative is discussed below. The second alternative is the use of test accommodations with ELL students. A much-needed standard procedure for implementing accommodations would be an outgrowth of an established validity threshold for academic language proficiency. The use of accommodations has already received attention (Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, Hofstetter, & Baker, 2000; Butler & Stevens, 1997; Castellon-Wellington, 1999; Olson & Goldstein, 1997). Currently, work underway by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) will examine the use of English language and bilingual dictionaries among other types of accommodations with ELL students in several states.

Identification of an ELL Validity Threshold

An important consideration underlying the research reported here was the goal of identifying and/or recommending a threshold level on a widely used language proficiency test that would indicate when ELL students' performance on a

standardized content test would be valid from a linguistic standpoint. The language test used in chapter 2 did not provide adequate specificity about student language at the upper range of proficiency, and thus is not a likely candidate for establishing a threshold. However, the notion of identifying a threshold of language proficiency is still viable with a test that provides a clear indication that the language complexity of the content assessment is not a barrier to student performance. Butler and Stevens (1997, p. 22) provide a flow chart that incorporates an academic language proficiency assessment as part of a decision-making process for providing test accommodations for ELL students.

The use of an academic language proficiency assessment would allow for another option in assessing English language learners: Include ELL students in the testing process but assess only their growth in English proficiency until they reach the language proficiency threshold. In other words, for accountability purposes, students who do not reach the threshold would take a measure of English growth at the same time other students take a content assessment. The state of Illinois is currently taking this approach with the Illinois Measure of Academic Growth in English (1999). In order to establish a validity/language proficiency threshold, we propose the development of a nationwide metric for defining the academic language ability of ELL students. It is to a discussion of that metric that we now turn.

A Nationwide Metric for Defining the Language Ability of ELL Students

As mentioned above, one stumbling block to both research and policy with ELL students is the lack of uniformity in how school districts and states operationally define these students through their designations such as LEP, FEP, RFEP, and bilingual. The lack of uniformity is due in large part to the different approaches states take to making their designations. A nationwide metric, a language test that allows for clear, objectively defined parameters for ranges of linguistic performance, would help remove this stumbling block and make articulation of ELL student performance uniform. The metric would specify academic language proficiency characteristics aligned with the type of language used on content assessments. It would be drafted based on additional study of the academic language requirements for successful performance on content assessments and would require participation of language experts as well as policymakers. OBEMLA, the Office of Bilingual Education and Minority Languages Affairs,[2] could play a critical advisory role in

[2] Now OELA, the U.S. Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students.

this effort by bringing states together in a way that will allow alignment of English language testing in grades K-12. The first step would be to convene a panel of experts to discuss test development and policy issues and to produce guidelines for moving the effort forward. Such an effort could be facilitated by CRESST. The group would need to include language testing experts, applied linguists, teachers, psychometricians, and policymakers. The intent would be to build on current research that suggests the critical need for sensitivity to issues of academic language in language test development for ELL populations. Initial CRESST efforts in the development of academic language tasks (Butler, Stevens, & Castellon-Wellington, 1999) and work from other sources (e.g., initiatives in Illinois, New York, and California) could serve as a point of departure.

Recommendations for Future Research

The research that has been reported here shows a clear relationship between language proficiency and performance on content tests for ELL students. However, these findings are strongly tempered by a number of major concerns and considerations. Though language is likely a dominant factor for ELL students, English language proficiency does not explain all the variation we found in students' content performance. In addition, we have been unable to attribute causality because of the nature of the data. Where they were available in the extant data sets, additional background factors also were found to play a role in predicting student performance, namely parent education level and family income level. Opportunity to learn content and academic language are both also potentially important predictors of student performance. These factors were not included in the extant data sets supplied by the school districts, nor are they factors that are easily measured and quantified. Therefore, we recommend future studies in which the interactions among variables that influence student performance are further explored.

An initial step in addressing opportunity to learn content is an experimental effort that was conducted with the same third-grade students reported in chapter 2 in the area of mathematics. This work, reported in Staley (2005), controls for opportunity to learn in a specific area of mathematics (statistics and probability) by providing students with direct instruction in this specific area prior to assessment. The results may provide initial evidence of causality between language proficiency and content knowledge for ELL students. Other work at UCLA focuses in part on

the effect of opportunity to learn in the social sciences content area (Aguirre-Muñoz, 2000). We propose additional intervention studies across content areas that allow for experimental control to determine cause and effect.

Further, we recommend controlled, small-scale research studies that investigate the effect of language proficiency on the demonstration of content knowledge and that take account of opportunity to learn both content material and academic language. Because the content of large-scale assessments is often cumulative, data should be collected on educational background to help identify gaps in student exposure to content. Students who have been in the United States for only a short time or have been enrolled in special programs may not have had exposure to the content being assessed.

Data collection would include

1. student and teacher surveys and interviews on curriculum and educational background;
2. a language test that reflects the language complexity of large-scale content assessments;
3. content assessments; and
4. posttest surveys, interviews, and/or focus groups.

Research along these lines will help provide a clearer understanding of what information is needed to determine language readiness, that is, the proficiency level needed for taking content assessments.

Final Remarks

What we have learned from the work reported here is that a multiplicity of factors are statistically significant indicators of student performance. We know that ELL students who are designated as LEP perform on standardized content tests at levels that are lower than those of EO students. However, low LEP student performance does not in itself make the tests or the test data invalid. We need to better understand the roles of academic language proficiency, student background, and opportunity to learn in ELL student performance on content assessments in order to determine the effectiveness and validity of the standardized assessments being used. In addition to these factors, we need more specific information from multiple sites about student performance on the content assessments. This information should include item-level data that will permit us to analyze ELL

student response patterns and thereby provide insight into how ELL students are processing test material compared to their EO counterparts. These indicators taken together will then allow us to more confidently determine when standardized content tests are valid indicators of content knowledge for ELL students.

References

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students, NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19,* 16-26.

Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners.* Unpublished doctoral dissertation, University of California, Los Angeles.

Butler, F., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F., Stevens, R., & Castellon-Wellington, M. (1999). *Academic language proficiency task development process* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Castellon-Wellington, M. (1999). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Illinois Measure of Annual Growth in English. Test for students in bilingual education. (1999). Springfield, IL: Illinois State Board of Education.

Olson, J. F., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessment: A summary of recent progress* (NCES 97-482). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Staley, L. (2005). *The effects of English language proficiency on students' performance on standardized tests of mathematics achievement.* Unpublished doctoral dissertation, University of California, Los Angeles.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs* (CSE Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).