High-Stakes Testing and English Language Learners: Questions of Validity

Elizabeth A. Mahon
Durham Public Schools, North Carolina

Abstract

The purpose of this study was to understand relationships between English proficiency and academic performance for a group of English language learners (ELLs) from 4 elementary schools. Descriptive and inferential statistics were used to examine scores from the Language Assessment Scales, the Woodcock-Muñoz Language Survey, and the Colorado Student Assessment Program. Findings showed that English proficiency was significantly related to English academic achievement, even for ELL students who had been in U.S. schools for 3 years or longer. Furthermore, the 5th-grade ELL cohort had greater increases in reading and writing scores compared to all Colorado 5th graders. This led to a slight closing of the achievement gap. Lastly, Spanish achievement, especially when combined with English proficiency, predicted English achievement.

Standards- and assessment-based reform has been widely implemented in U.S. schools since the passage of Goals 2000 (1994) and the No Child Left Behind Act (2002). These laws require that all students meet challenging academic standards and that schools be held accountable for the progress of all their students. The ideas of high standards, challenging tasks and accountability for all students sound advantageous for English language learners (ELLs).[1] With standards-based reform, there is hope that attention and resources will be directed at these ELL students and other minority-group students.

However, several voices warn against optimism with the standards- and assessment-based reforms for ELL students. The equity benefits of the standards-based movement are "more an aspiration than a certainty" (McLaughlin & Shepard, 1995, p. 11), and such benefits are "not a foregone conclusion" (August, Hakuta & Pompa, 1994, p. 5). LaCelle-Peterson and

Rivera (1994) echoed this uncertainty with respect to the reform of assessments: "Efforts to reform assessment as part of systematic reform do not clearly bode well or ill for ELLs; while there are evident grounds for hope, there are equal grounds for caution" (p.13).

The grounds for caution with using large-scale, standardized assessments with ELLs are far reaching. Not only is there historical evidence of misuse of such assessments with minority populations (Baca & Cervantes, 1998), but there are numerous questions about how to accurately assess the content knowledge of ELLs in schools today. There is not a consensus on the optimal stage of second-language development at which to begin testing ELL students (August & Hakuta, 1997; Figueroa & Hernández, 2000; García & Pearson, 1994). There are questions about cultural biases of standardized tests (Valdés & Figueroa, 1994), and about what level of cultural competence is needed for ELL students to successfully negotiate state standardized tests.

Interpretation of ELL test scores is another area replete with questions. If accommodations are used with ELL students, it is not known how comparable the scores are to the general population (Koenig, 2002). English proficiency continues to be a confounding factor with interpreting test scores. Sandoval and Durán (1998) ask:

> What inferences can be drawn from the use of tests with individuals limited in their command of English? What inferences can be drawn when the tests have been administered so that the instructions or the substance and content of the task have not been completely understood by the examinee? (p. 181)

In all of the legislation, little direction was given as to how to operationalize the large-scale application of standardized testing to a culturally and linguistically diverse population. With this study, I examined the interpretation of achievement test scores for ELLs in three areas: (a) the relationship between English proficiency and English academic achievement, (b) the progress in closing the achievement gap for ELLs, and (c) the relationship between Spanish and English achievement scores. I briefly describe each area below.

English Proficiency and Academic Achievement

Validity is the key concept used to frame the research questions about the relationship between English proficiency and academic achievement. Validity has always been concerned with the meaning and interpretation of test scores. Historically, validity was thought of as a triptych: construct, content and criterion-related validity. However, current conceptualizations of validity use the metaphor of building an argument for the use and interpretation of scores. Validity is defined as an evaluative judgment based upon a collection of evidence. The evaluation is applied to the interpretation and use of scores,

not to the test itself (Linn, 1999; Messick, 1989; Moss, 1992; Shepard, 1993). As stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999),

> The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed test score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. (p. 9)

Validity is measured in degrees, and refers to the amount of evidence available to support the use and interpretation of the test scores in particular situations.

Researchers use the concept of construct irrelevant variance to describe "the degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA, APA & NCME, 1999, p. 10). Other authors use the words "ancillary" or "irrelevant" to refer to the intervening factor (Linn, 1999; Sandoval & Durán, 1998). With English achievement tests, English proficiency may be considered a source of construct irrelevant variance. For example, the third-grade reading test, the Colorado Student Assessment Program (CSAP), measures progress towards the state reading standard which requires students to read and understand a variety of materials. Theoretically, students could demonstrate reading comprehension in any language—Spanish, Arabic, Portuguese, or Swahili. However, when ELL students are required to take a test in English, all the tasks require a certain level of English fluency, though fluency in English is not the targeted construct. Low levels of English language proficiency could be considered an ancillary skill, or a source of construct irrelevant variance.

This study examines to what degree English proficiency influences performance on measures of academic achievement with two research questions: Is there a difference in English proficiency level among CSAP proficiency categories (unsatisfactory, partially proficient, proficient/advanced) in reading, writing, and mathematics? Further, to what extent does English language proficiency predict English academic achievement, as measured by the CSAP in reading, writing, and mathematics?

The Achievement Gap

In the past, the achievement gap literature has focused on the differences between African Americans, Hispanics and Anglos, but there has not been particular attention paid to ELLs (e.g., Jencks & Phillips, 1998; Lee, 2002; Miller, 2003). The Hispanic category includes both ELLs and English-only Hispanics, so it is impossible to distinguish how the ELLs are progressing among the subsuming category of Hispanic. Several studies from California

have set the model for how to interpret the progress of ELL students with standards-based assessments (Parrish et. al, 2002; Thompson, DiCerbo, Mahoney & MacSwan, 2002). The preferred methods include analyzing changes in the achievement gap over time, comparing mean scaled scores, calculating effects sizes to show differences, using quasi-cohorts or true cohorts, and including fluent English proficient (FEP) students.

Using a mix of these methods, Thompson et al. (2002) found that with few exceptions, the achievement gap between ELLs and all California students did not appear to be narrowing. However, Parrish et al. (2002), in a study which included all California schools, found a slight narrowing of the achievement gap, from .05 to .20 of a standard deviation. I used methods from the California studies to examine the progress of ELL students in Colorado with one research question: How does the academic achievement of the fifth grade ELL cohort, as measured by the CSAP reading and writing, compare to the English academic achievement of all Colorado students?

Spanish and English Achievement Scores

Colorado uses the CSAP reading and writing assessments in Grades 3–4 for students who are attending bilingual schools, so it was possible to compare the English CSAP scores with the Spanish scores. The Spanish versions of the tests were designed to measure the same standards as the English versions, but with materials and questions derived from Spanish writing (CTB/ McGraw-Hill, 2003; T. Quackenboss, personal communication, November 24, 2003). The theoretical support for the relationship of Spanish and English achievement comes from Cummins' theory of common underlying proficiency (Cummins, 1981), where the development of language and literacy in one language builds a base of knowledge that can transfer to a second language. In this study, I apply language transfer theory to criterion-referenced achievement tests and ask: To what extent does Spanish academic achievement, as measured by the Spanish CSAP, predict English academic achievement, as measured by the English CSAP?

Method

Participants

The data set came from a larger study, *An Analysis of Limited English Proficient Student Achievement on Colorado State Reading, Writing and Math Performance Standards*, funded by the Office of English Language Acquisition (Escamilla, Baca, Hoover, & Almanza de Schonewise, 2005). English language proficiency scores and CSAP achievement scores were collected for 200 ELLs in the fourth and fifth grades. These schools were selected

because they were typical examples of particular program models in Colorado. As schools and students were not selected randomly, there are limitations as to how much the findings generalize from this study to a larger population. Characteristics of the student groups and schools are presented so that readers may make comparisons to other student groups.

Data were gathered from four different elementary schools in two districts. In Saddle Valley district, Clay and Iris Elementary Schools used a transitional bilingual education model with 30% to 40% of students labeled ELLs. At Clay and Iris Elementary, 26% and 42% of the student population qualified for free and reduced lunch respectively. In Butler District, Linda Elementary used a pull-out ESL program model, and had approximately 17% ELLs and 22% free and reduced lunch. Pine Mountain was the dual language site, with 54% ELLs and 44% free and reduced lunch. In summary, 44% of the students were in transitional bilingual programs (n = 87), 37% were in dual language programs (n = 72), and 11% were served in ESL programs (n = 22).

Fifth graders comprised 54% of the sample (n = 107) and fourth graders comprised 47% of the sample (n = 93). The sample consisted of 53% males (n = 105) and 48% females (n = 95). Approximately 96% of the students were classified as Latino (n = 191) with Spanish as their first language. Butler County supplied data that clearly showed which students were categorized limited English proficient (LEP) and FEP. The Saddle Valley data came from the district ELL database, which tracks only students who have not yet exited LEP services, so all Saddle Valley students were labeled LEP for the study. Approximately 90% of the study group was labeled LEP (n = 178), and 10% was labeled FEP (n = 19). Program data were not given for 8% of the study group (n = 16).

The majority of students attended schools in the two districts for at least 3 years, with 76% of the participants (n = 151) entering the school districts in spring 2000 or earlier. Approximately 15 % of the students (n = 29) attended schools in the district for 1 to 2 years, while 10% (n = 19) attended schools in the district for less than one year, entering in the 2002–2003 school year. Students may have entered Butler County and Saddle Valley from other school systems in the United States, so the number of years in the district does not represent how long students have attended U.S. schools. Data on how long students attended U.S. schools were not available through the district databases. The total number of years spent in U.S. schools would affect English proficiency, so this lack of available data is a limitation of this study.

Instruments

This study used achievement scores from the CSAP (CTB/McGraw-Hill, 2001, 2002, 2003). The CSAP is a criterion-referenced test that measures students' progress towards Colorado State Content Standards in reading,

writing, mathematics, and science. CSAP scores are reported to the public as percentages of students scoring in four different proficiency categories: unsatisfactory, partially proficient, proficient and advanced. In this study, there was only one student who scored in the advanced category, so the proficient and advanced categories were collapsed into one level (proficient/advanced) to analyze if there is a difference in English proficiency level among CSAP proficiency categories. For the remaining research questions, I used the scaled scores of the CSAP, which are an equal interval scale allowing for comparison across time for specific subject tests, as well as across grade levels.

This study also used English language proficiency scores from the Language Assessment Scales–Oral Short Form (LAS–O) (De Avila & Duncan, 1990) and from the Woodcock-Muñoz Language Survey Normative Update (WMLS) (Woodcock & Muñoz-Sandoval, 2001). The LAS–O consists of three subtests: vocabulary, listening comprehension, and story retelling. Students receive a LAS–O scaled score which runs from 0 to100. From these scaled scores, students are placed in five different proficiency levels. Level 1 with scaled scores of 0–54 are labeled non-speakers; Levels 2–3 with scaled scores of 55–74 are labeled limited speakers; and Levels 4–5 with scaled scores of 75–100 are designated fluent speakers.

The WMLS consists of four subtests: picture vocabulary, verbal analogies, letter-word identification, and dictation. Unlike the LAS–O, two of the WMLS subtests include tests of oral English abilities, as well as reading and writing skills. Students receive a WMLS scaled score that is referred to as a "broad English ability." The broad English ability scaled scores are an interval level scale, ranging from 340 to 595. From these scaled scores, students are placed in five proficiency levels: negligible, very limited, limited, fluent and advanced. The proficiency levels of 1–5 are an ordinal level scale, and are not used in this study. I used the WMLS scaled scores, which are an interval scale.

Each of these instruments has their limitations. The development of the CSAP lacked the psychometric rigor that is used to develop norm-referenced tests. There is a lack of detailed information about the psychometric properties of the test in the manuals, and a lack of research on the psychometric properties of the test. Furthermore, though the CSAP manual reports the results of the differential item analyses, little other information is given on whether there is evidence for validity with different cultural groups.

The LAS–O and WMLS are state-approved English proficiency assessments. Both tests take approximately 20 minutes to administer. Though both tests report sufficient validity and reliability information, there is a lack of information on how the scores from these tests relate to the type of English proficiency needed in classrooms. Furthermore, scores for one of the subtests of the LAS–O (listening comprehension) were not adequate to assume

reliability or validity with this subtest. For this study, LAS–O scores were not directly analyzed, so the problems with the test are not seen as affecting this study. The shortcomings of WMLS and the CSAP are a limitation of this study.

Data Analysis and Results

I used quantitative data on language proficiency and academic achievement collected from four elementary schools in two districts. SPSS was the computer program used to analyze the data.

English Proficiency

The first two research questions are presented together as both address the relationship between scores of English proficiency and scores of English academic achievement: Is there a difference in English proficiency level, as measured by the WMLS, across three CSAP proficiency categories (unsatisfactory, partially proficient, proficient/advanced) on the fourth- and fifth-grade 2003 CSAP in reading, writing, and mathematics? To what extent does English language proficiency, as measured by the WMLS, predict English academic achievement, as measured by the CSAP, on the fourth- and fifth-grade 2003 CSAP?

Both questions used the same data set. Data from 67 fourth- and fifth-grade students were used in the analyses for reading and writing. These students had both English language proficiency and English CSAP reading and writing scores from the 2003 spring testing sessions. Approximately 90% of the students ($n = 60$) in the reading/writing group attended district schools for at least 3 years. For the mathematics analysis, scores of 41 fifth-grade students from the spring 2003 testing session were used. The math CSAP was not administered until the fifth grade, so there were no fourth grade scores. For the math group, 85% ($n = 34$) of the students attended district schools for at least 3 years.

One-way analysis of variance (ANOVA) was used to determine if there were differences in means on the WMLS across CSAP subject-area proficiency categories: (a) unsatisfactory, (b) partially proficient, and (c) proficient/advanced. There was only one student who scored in the CSAP advanced category, so the proficient and advanced categories were collapsed into one level (proficient/advanced) for data analysis. I conducted an analysis of variance three times. For each analysis, I respectively used the scores from CSAP content tests of reading, writing and math (each divided into three categories) as the factor and WMLS scores as the dependent variable.

The Levene's test of homogeneity of variance, with alpha set at .05, showed that the assumption of homogeneity of variance was satisfied for all three CSAP content areas. For reading $p = .436$; for writing, $p = .556$, and

for math, $p = .086$. For each subject area, the null hypothesis that there is no difference in the population variances remains tenable. The assumption of normality was checked by examining the results of the normal P-P plot of the residuals, and by studying the boxplots of the distributions. Independence of groups was assumed because each set of scores came from a different student, only one set of scores was used for each student, and there was no overlap of students between groups. These tests showed that all of the assumptions were satisfied for the ANOVA analysis. To control Type I error across the three ANOVA, I set alpha for each test at .017 (.05 / 3 = .017). I used Bonferroni's multiple comparison procedure to detect which sets of means differed significantly. Bonferroni's method was selected because it reduces the likelihood of Type I error by adjusting the observed significance level for fact that multiple comparisons are made.

A post-hoc power analysis was used to determine the probability of correctly rejecting a false null hypothesis with the ANOVA analysis. The power analysis calculated the probability that it was correct to say that there was a difference between means. For this analysis, SPSS used the statistics in the sample data as if it were population data. The power analysis indicated that for all CSAP content areas, there was at least a 97% chance of correctly rejecting a null hypothesis. This was above the standard requirement of 80% for a power analysis.

The hypothesis was that as students moved from a lower CSAP proficiency category to a higher CSAP proficiency category, their scores of English language proficiency should also increase. Table 1 provides the means and standard deviations for WMLS scores by CSAP proficiency category for

Table 1

Means and Standard Deviations of Woodcock-Muñoz Language Survey (WMLS) Scores by 2003 Colorado Student Assessment Program (CSAP) Proficiency Categories

CSAP proficiency category	**WMLS**								
	Reading			**Writing**			**Mathematics**		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Unsatisfactory	24	470	18	17	467	15	11	477	21
Partially proficient	28	486	13	43	488	12	21	485	13
Proficient/advanced	15	500	11	7	506	7	9	506	8

reading, writing and math. As hypothesized, WMLS means increased among the three categories from unsatisfactory, to partially proficient to proficient/advanced for each subject area.

The results of the three ANOVA showed that for each of the subject areas, the mean differences were statistically significant: For reading, $F(2, 64) = 20.4$, $p < .001$; for writing, $F(2, 64) = 29.8$, $p < .001$; and for math, $F(2, 38) = 9.9$, $p < .001$. Thus, for each subject area, there was a difference between the WMLS means across the CSAP proficiency categories.

Bonferroni's multiple comparisons procedure showed which sets of means had significant differences. The test was performed for each content area, with the alpha level set at .05. For reading and writing, the mean differences were significant with each pair of comparisons. For mathematics, the mean differences were significant between the partially proficient and proficient/advanced categories ($p = .004$), but the mean differences were not significant between the unsatisfactory and partially proficient category ($p = .405$).

From the ANOVA we see that levels of English proficiency are related to scores of English academic achievement.

Three bivariate linear regression analyses were conducted to determine how much influence English language proficiency had on CSAP achievement. The English language proficiency scaled scores were the independent variable, whereas CSAP reading, writing and math scores were respectively used as the dependent variable for each regression. For all three linear regression analyses, normality of residuals, linearity, and homoscedasticity were confirmed by examining scatterplots, graphs of the residuals and standardized predicted values.

Table 2 presents the descriptive statistics and the coefficient of determination (R^2) for the WMLS for each CSAP subject area. The correlations between the WMLS scores and CSAP reading ($r = .79$), CSAP writing ($r = .82$), CSAP math ($r = .61$) were strong, and positive. The regression analyses showed that the WMLS scores contributed information to the prediction of each CSAP subject-area scores at a statistically significant level: For reading, $F(1, 65) = 108.12$, $p < .001$; for writing, $F(1, 65) = 132.38$, $p < .001$; and for math, $F(1, 39) = 23.201$, $p < .001$. Sixty-seven percent of the variance in reading scores could be accounted for by variance in WMLS scores ($R^2 = .67$), while 63% of the variance in writing scores could be attributed to the variance in WMLS scores ($R^2 = .63$). This relationship was not as strong with math, nevertheless, 37% of the variance in math scores could be accounted for by variance in WMLS scores ($R^2 = .37$).

In sum, both the ANOVA and regression analyses showed that WMLS scores were related to CSAP scores in all three content areas. The bivariate linear regression analyses demonstrated that WMLS scores were predictors of reading, writing and math CSAP scores. The relationship between English proficiency and CSAP reading and writing was stronger than between English proficiency and mathematics.

Table 2

Descriptive Statistics and Correlations for Colorado Student Assessment Program (CSAP) and Woodcock-Muñoz Language Survey (WMLS)

Variable	*n*	*M*	*SD*	*r*	R^2
Reading					
CSAP	67	541	57[a]	.79	.67
WMLS	67	484	18		
Writing					
CSAP	67	435	44	.82	.63
WMLS	67	485	17		
Math					
CSAP	41	450	49	.61	.37
WMLS	41	488	18		

Note. All correlations are statistically significant at $p < .001$.

[a]The standard deviation for the 2003 CSAP for all Colorado fourth- and fifth-grade reading is between 65–69, for the 2003 CSAP fourth- and fifth-grade writing is 55–57, and for 2003 fifth-grade math is 72. The standard deviations for the ELL group reported here are smaller than the standard deviations reported for CSAP tests in general, indicating less of a score spread for the ELL group.

The Achievement Gap

How does the English academic achievement of the fifth-grade ELL cohort, as measured by the CSAP reading and writing, compare to the English academic achievement of all Colorado students from spring 2001 to spring 2003? Line graphs and effect sizes were used to show the difference between the scores of the ELL cohort and all Colorado students. Effect sizes were calculated to show the change across the years for each group and the change in the achievement gap between the two groups. In all cases, the average of the standard deviations for each year was used in calculations. The standard deviations for the CSAP were between 60 and 90 for reading, and approximately 50 for writing.

The fifth-grade ELL cohort is the group of students who attended fifth grade in 2003, but started their CSAP testing in third grade in 2001. Of the 107 fifth graders in the database, only 22 students had English reading CSAP scores for all 3 years of testing. This low number was primarily due to the large number of students who took CSAP reading assessments in Spanish ($n = 62$ in 2001, $n = 39$ in 2002), as well as to students moving out of the district. Results for the 3-year reading longitudinal analysis, thus, were based upon 22 scores. This group is referred to as the 3-year ELL cohort. This small sample size is a

Table 3

Colorado Student Assessment Program (CSAP) Reading 2001–2003 for 3-Year ELL Cohort and All Colorado Students

Group	CSAP scores						Effect size
	2001		2002		2003		2001–2003
All Colorado students	*M*	*SD*	*M*	*SD*	*M*	*SD*	
	566	76	584	66	608	69	.60
ELL cohort	495	90	533	66	560	65	.88
Score difference between groups	71		51		48		.31

limitation of the achievement gap analysis. However, it was also possible to analyze growth in English CSAP reading scores from only 2 years of testing, from 2002 to 2003. Forty-three fifth-grade students had English CSAP reading and writing scores for both 2002 and 2003. This group is referred to as the 2-year ELL cohort in the following discussion. The scores of the 2-year cohort were also used for the analysis of the CSAP writing, as the writing assessments were first given in 2002. These groups of ELL students were compared to all Colorado students, which included all Colorado students (approximately 55,500) who took the CSAP test in English at a particular grade level.

Table 3 reports descriptive statistics for the reading CSAP for the 3-year cohort and all Colorado students. The scores for reading for both groups increased from 2001–2003. Colorado students increased their scores by .60 of a standard deviation while the ELL cohort improved their scores by .88 of a standard deviation. This leads to a slight closing of the achievement gap (.31 of a standard deviation), judged to be a small to medium effect size. For all 3 years, there is a gap in achievement between Colorado students and the ELL cohort (.71–.86 of a standard deviation). The means for the 3-year ELL cohort fell in the partially proficient CSAP category, while the means for all Colorado students fell in the proficient category.

Table 4 shows the same trend with the achievement gap for reading and writing CSAP scores of the 2-year cohort. Both the 2-year ELL cohort and all Colorado students improved their scores from 2002 to 2003. However, the mean score of the ELL group improved slightly more than the mean score for all Colorado students, which leads to a slight closing of the achievement gap for both reading (.12 effect size) and writing (.17 effect size). For both years, the CSAP mean for reading and writing for the 2-year ELL cohort fell in the partially proficient category, while the CSAP mean for all Colorado students fell in the proficient category.

Table 4

Colorado Student Assessment Program (CSAP) Reading and Writing, 2003–2003 for 2-Year ELL Cohort and All Colorado Students

Group	Mean scaled score		Effect size
	2002	2003	2002–2003
Reading			
All Colorado students	584	608	.36
ELL cohort	527	559	.47
Score difference between groups	57	49	.12
Writing			
All Colorado students	485	502	.31
ELL cohort	435	461	.50
Score difference between groups	50	41	.17

Note: Standard deviations for writing and reading were similar for all Colorado students and the 2-year ELL cohort. For writing, standard deviation was 51–57, and for reading, standrd deviation was 66–69.

In summary, the cohort analysis showed that English academic achievement in both reading and writing increased for both the 2-year and 3-year ELL cohorts from 2001 to 2003. The increases were judged to be medium to large in strength (.47–.88 effect size). In both reading and writing, there was a greater increase in scores for the ELL students than for all Colorado students, which leads to a slight narrowing of the achievement gap. The change in the size of the achievement gap was judged to be small, but consistently in the direction of closing the gap.

Spanish and English CSAP

To what extent does Spanish academic achievement, as measured by the Spanish CSAP, predict English academic achievement, as measured by the English CSAP? Bivariate and multiple linear regression analyses were used to determine how much influence English language proficiency and Spanish CSAP scores had on CSAP achievement scores. Students took the Spanish CSAP one year (either 2001 or 2002), and then took the English CSAP the consecutive year (either 2002 or 2003). Scores from 92 students were used in

Table 5
Descriptive Statistics and Correlations for Spanish Colorado Student Assessment Program (CSAP) and English CSAP

Variable	*n*	*M*	*SD*	*r*	R^2
Reading					
Spanish	92	531	43	.73	.53
English	92	534	61		
Writing					
Spanish	35	544	43	.76	.50
English	35	452	42		

Note. All correlations are statistically significant at $p < .001$.

the reading analysis. Ninety percent ($n = 83$) of the students in the reading group had attended schools in the district for at least 3 years. Scores from 35 students were used in the writing analysis. There were fewer writing scores because English and Spanish writing assessments were not administered until 2002. For writing, 83% of students ($n = 29$) had attended schools in the district for at least 3 years.

I conducted two separate bivariate regression analyses, one for reading and one for writing. For both subject areas, the English CSAP scaled scores were the dependent variable, and the Spanish CSAP scaled scores were the independent variable. Table 5 presents the descriptive statistics and the linear regression analysis for the Spanish and English CSAP in reading and writing. The correlations between the Spanish and English reading ($r = .73$) and between Spanish and English writing ($r = .76$) were strong and positive. The linear regression analysis showed that the Spanish CSAP scores contributed information to the prediction of the CSAP scores at a statistically significant level: For reading, $F(1, 90) = 101.85$, $p < .001$; for writing, $F(1, 33) = 33.31$, $p < .001$. Fifty-three percent of the variance in English reading scores could be accounted for by variance in Spanish scores ($R^2 = .531$), while 50% of the variance in English writing scores could be attributed to the variance in Spanish writing scores ($R^2 = .502$).

This result led to questions about what factors could account for the other 50% of the variance in English CSAP reading and writing scores. I conducted two additional multiple regression analyses, respectively adding English proficiency scores as measured by WMLS, to the two bivariate regressions, as a second independent variable. Scores for 48 students were available for the reading analysis, while scores for 26 students were available for the writing analysis.

Table 6 shows that in both content areas, the correlations between English CSAP scores and Spanish CSAP scores (r = .602 for reading, r = .585 for writing) and between English CSAP scores and WMLS scores (r = .795 for reading, r = .776 for writing) were strong and positive, indicating a linear relationship with English CSAP scores for both of these variables. The multiple regression analyses for reading and writing are summarized in Table 7. For reading, R^2 = .726, $F(2,45)$ = 59.561, p < .001; for writing, R^2 = .728, $F(2,23)$ = 30.751, p < .001. For both CSAP reading and writing, approximately 73% of the variance in English CSAP scores could be explained by reference to WMLS scores and Spanish CSAP scores.

For reading, WMLS scores played a major role in predicting CSAP English scores: ß = .659, $t(44)$ = 7.72, p < .001. The Spanish CSAP scores also played a statistically significant role, though Spanish scores were not as heavily weighted as the WMLS scores: ß = .336, $t(44)$ = 3.938, p < .001. The writing analysis was very similar to the reading analysis: WMLS scores played a major role in predicting English CSAP writing scores (ß = .656, $t(23)$ = 5.706, p < .001). The Spanish CSAP scores contributed at a significant level as well (ß = .374, $t(23)$ = 3.258, p < .003), but were not as heavily weighted as the WMLS scores.

In summary, the multiple regression analysis showed that both English language proficiency and Spanish academic achievement contributed to the prediction of English academic achievement for reading and writing. When Spanish academic achievement alone was considered, approximately 50% of the variance in English CSAP scores could be explained by the variance in Spanish scores. When English language proficiency was added to the equation, approximately 73% of the variance in English CSAP scores could be explained by the linear combination of Spanish CSAP scores and English language proficiency scores.

Summary of Findings

The small number of students and the shortcomings of the CSAP, WMLS, and LAS–O limit the generalizability of the current study. We do not know the total number of years that students spent in U.S. schools, though we know that the majority of students attended Bultler and Saddle Valley for 3 years or more. We cannot be sure that the findings would be replicated with assessments from other states, or with other more comprehensive measures of language proficiency. We also cannot be sure whether the findings would translate to other populations. For these reasons, this study is exploratory and cannot unequivocally establish the relationship between language proficiency scores and CSAP scores. Frustrating as these limitations may be, the discussion about the complex relationship between academic achievement and language proficiency variables warrants attention, and the findings from this study provide quantitative starting points for this discussion.

Table 6

Correlation Matrix and Descriptive Statistics for Woodcock-Muñoz Language Survey (WMLS) Scores and Spanish Colorado Student Assessment Program (CSAP) Scores with English CSAP Scores

Variable	English CSAP	Spanish CSAP	WMLS Scores
Reading (n = 48)			
English CSAP scores	1.0	-	-
Spanish CSAP	.602	1.0	-
WMLS scores	.795	.404	1.0
M	544	519	482
SD	57	90	20
Writing (n = 26)			
English CSAP scores	1.0	-	-
Spanish CSAP	.585	1.0	-
WMLS scores	.776	.322	1.0
M	452	541	488
SD	38	37	15

Table 7

Multiple Regression Analysis of Spanish Colorado Student Assessment Program (CSAP) and Woodcock-Muñoz Language Survey (WMLS) Scores on English CSAP

Content Area	Independent variable	*B*	ß	*t*	*p*	Partial correlation
Reading (n = 48)	WMLS	1.934	.659	7.72	< .001	.72
	Spanish CSAP	.214	.336	3.938	< .001	.51
Writing (n = 26)	WMLS	1.624	.656	5.706	< .001	.77
	Spanish CSAP	.381	.374	3.938	.003	.56

Note. For reading, all correlations are significant at the $p < .002$ level. For writing, all correlations are statistically significant at the $p < .001$ level except the relationship between WMLS scores and the Spanish CSAP scores, which is not statistically significant, $p = .054$.

Evidence from this study suggests that English proficiency scores were confounded with CSAP scores. English language proficiency acted as a source of construct irrelevant variance. For reading and writing, 63–67% of the variance in English CSAP scores could be accounted for by variance in WMLS scores. For mathematics, 37% of the variance in CSAP scores could be accounted for by variance in WMLS scores. It is notable that the majority of students (76%) had been in Butler and Saddle Valley for at least 3 years, indicating that it takes many students longer than 3 years to develop academic English fully.

For the fifth-grade cohort in the study, there was a slight closing of the achievement gap between ELLs and English-only students (.12–.31 of a standard deviation). All Colorado students showed a mean score gain each year, but the mean score for ELLs rose more than general population. This finding concurs with Parrish et al. (2002), where the achievement gap closed slightly(.05 to .20 of a standard deviation) in California from 1998 to 2001.

Lastly, there was a strong correlation between CSAP scores on English and Spanish reading (r = .73) and English and Spanish writing (r = .76). Additional analyses with the Spanish data showed that approximately 73% of the variance in English CSAP scores was accounted for by the combination of English proficiency scores and Spanish CSAP scores.

Implications

At the time of this study, ELLs were exempted from English CSAP testing for their first 3 years in U.S. schools. Even after 3 years in U.S. schools, however, English proficiency continued to affect ELLs' performance on the CSAP. The CSAP is not accurately measuring academic achievement of ELLs to the same degree as for English only or FEP students. Using a time limit of 3 years in U.S. schools is not adequate evidence for English fluency. Instead of using a time limit, this study suggests using a measure of English proficiency as an indicator of testing readiness in English. In the suggested scenario, a student would take the CSAP only after he or she had reached a certain cut-off score on the English proficiency measure.

Several policy reports ask at what point along the continuum of language proficiency it becomes valid to test ELL students in English (August & Hakuta, 1997; Valdés & Figueroa, 1994). From the ELL students in Butler Valley who took both the WMLS and the English CSAP in 2003 (N = 67), 15 students scored at the proficient level on the reading CSAP, 7 scored at the proficient level on the writing CSAP, and 9 scored proficient on the math CSAP. Seventeen students scored proficient on one or more CSAP test. This small group of students scored at a WMLS broad English proficiency level of 3.5 to 4, which is in the *limited* (Level 3) to *fluent* range (Level 4). The numbers of this study

are too small to make a statement about the threshold of English proficiency needed for CSAP testing, but the general idea behind the analysis could readily be applied to a larger group of students. For example, districts could analyze the language proficiency scores of ELL students who score in the proficient range on the CSAP. A minimum threshold of language proficiency could be established based upon this analysis. This would be an alternative to the 3-year time limit for deciding if to test ELL students.

However, the purpose of including ELL students in assessments is to emphasize accountability for their education. Waiting for students to reach a certain level of English proficiency before standardized academic testing is not pragmatic in the era of accountability. What other options are there? If students are in bilingual programs, native language academic assessments are a viable alternative. Reforms in assessment policy need to advocate native language assessment, as measures of Spanish academic achievement give us much-needed information on the academic progress of ELLs. Portfolios of academic progress or language-simplified tests in English may be another option. However, most likely, the CSAP will be continued to be administered to ELLs, regardless of their English proficiency levels. In that case, this study suggests analyzing and interpreting the data with caution. It would be best to compare an ELL cohort to their English-only peers using scaled scores over time.

Lastly, this study substantiates the concept that it takes more than one year to learn English. Programs that claim to teach ELLs English in one year contradict the evidence that shows that English acquisition is a lengthier process. School reformers should offer programs that support the language and academic achievement of ELLs for longer than one year.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academy Press.

August, D., Hakuta, K., & Pompa, D. (1994). *For all students: Limited English proficient students and Goals 2000*. Washington, DC: National Clearinghouse for Bilingual Education.

Baca, L., & Cervantes, H. (1998). *The bilingual special education interface* (3rd ed.). Upper Saddle River, NJ: Merrill/ Prentice Hall.

CTB/McGraw-Hill. (2001–2003). *Colorado Student Assessment Program*. Monterey, CA: CTB/ McGraw-Hill.

De Avila, E., & Duncan, S. (1990). *Language Assessment Scales Oral* (LAS–O). Monterey, CA: CTB / McGraw-Hill.

Cummins, J. (1981). *The role of primary language development in promoting educational success for language minority students*. Los Angeles: California State University, Evaluation, Dissemination, and Assessment Center.

Escamilla, K., Baca, L., Hoover, J., & Almanza de Schonewise, E. (2005). *An analysis of limited English proficient student achievement on Colorado state reading, writing, and math performance standards* (Field Initiated Project No. T292B010005). Washington, DC: U.S. Department of Education, Office of English Language Acquisition.

Figueroa, R., & Hernández, S. (2000). *Testing Hispanic students in the United States: Technical and policy issues*. Washington, DC: U.S. Department of Education.

García, G., & Pearson, P. (1994). Assessment and diversity. *Review of Research in Education, 20,* 337–391.

Goals 2000: Educate America Act. Pub. L. No. 103–227(1994).

Jencks, C., & Phillips, P. (Eds.). (1998). *The black-white test score gap.* Washington, DC: Brookings Institution Press.

Koenig, J. (Ed.). (2002). *Reporting test results for students with disabilities and English language learners: Summary of a workshop.* Washington, DC: National Academy Press.

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.

Lee, J. (2002). Racial and ethnic achievement trends: Reversing the progress toward equity. *Educational Researcher, 31,* 3–12.

Linn, R. (1999). Validity standards and principles on equity in educational testing and assessment. In A. Nettles & M. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 13–33). Boston: Kluwer Academic.

McLaughlin, M., & Shepard, L. (1995). *Improving education through standards-based reform*. Stanford, CA: National Academy of Education, Panel on Standards-Based Education Reform.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Miller, G. (2003). Analyzing the minority gap in achievement scores: Issues for states and federal government. *Educational Measurement: Issues and Practice, 22* (3), 30–36.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229–258.

No Child Left Behind Act. Pub. L. No. 107–110 (2002).

Parrish, T., Linquanti, R., Merickel, A., Quick, H., Laird, J., & Esra, P. (2002). *Effects of the implementation of Proposition 227 on the education of English language learners, K–12*. Washington, DC: American Institutes for Research and WestEd.

Sandoval, J., & Durán, R. (1998). Language. In J. Sandoval, C. Frisby, K. Geisinger, J. Scheuneman, & J. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 181–212). Washington, DC: American Psychological Association.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.

Thompson, M., DiCerbo, K., Mahoney, K., & MacSwan, J. (2002). ¿Exito en California? A validity critique of language program evaluations and analysis of English learner test scores. *Education Policy Analysis Archives, 10*(7).

Valdés, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex Publishing Corporation.

Woodcock, R. & Muñoz-Sandoval, A. (2001). *Woodcock-Muñoz Language Survey Normative Update, English Form*. Itasca, IL: Riverside.

Endnote

[1]The abbreviation ELL is used throughout the study to refer to English language learners, students whose first language is not English and who are in the process of learning English.