

A Study Of Content Area Assessment For English Language Learners

(Contract No. T292B010001)

Final Report

Prepared for:

**Office of English Language Acquisition and Academic Achievement for Limited
English Proficient Students,
U.S. Department of Education**

Prepared by:

Anne Katz, Patricia Low, Jim Stack, and Sau-Lim Tsang

**ARC Associates, Inc.
1212 Broadway, Suite 400
Oakland, CA 94612**

September 2004

Table of Contents

I. INTRODUCTION.....	3
A. BACKGROUND OF THE STUDY.....	3
B. REVIEW OF LITERATURE.....	6
i. <i>Early Measures of Language Proficiency</i>	7
ii. <i>Oral Language Proficiency and Academic Achievement</i>	8
iii. <i>Inadequacy of the Testing Instruments</i>	10
iv. <i>Models of Language Proficiency, Language Use and Language Ability</i>	13
v. <i>Academic Achievement and Language Proficiency</i>	22
vi. <i>The Assessment of English Development and Academic Achievement of English Language Learners</i>	27
vii. <i>Integrating Factors in the Assessment of English Learners</i>	28
viii. <i>Redefining English Language Proficiency</i>	29
II. RESEARCH QUESTION	31
A. ASSUMPTIONS.....	32
B. ADDITIONAL CONSIDERATIONS	33
i. <i>Explaining “meaningful” and “valid”</i>	33
ii. <i>Defining “language proficiency”</i>	33
iii. <i>Examining the heterogeneity of ELLs as a group</i>	33
III. RESEARCH SETTING	34
A. THE STATE FRAMEWORK FOR ACCOUNTABILITY.....	34
B. THE DISTRICT CONTEXT	35
IV. STUDY DESIGN.....	36
A. OVERVIEW OF EL DATA.....	36
B. STUDENT LANGUAGE PROFICIENCY STATUS.....	37
C. SAN FRANCISCO UNIFIED SCHOOL DISTRICT TESTING FRAMEWORK	39
D. PROJECT MEETINGS AND ADVISORY TEAMS	39
i. <i>Research Team</i>	39
ii. <i>The Study Team</i>	40
iii. <i>The Advisory Committee</i>	40
V. DATA ANALYSIS AND RESULTS	41
A. COMPONENT 1: ACHIEVEMENT PATTERNS OF ELS AND EOS OVER TIME	41
i. <i>Test data</i>	41
ii. <i>Limitations of data</i>	41
iii. <i>Analyses of Data and Results: Descriptive statistics</i>	43
iv. <i>Analyses of Data and Results: Correlations</i>	45
v. <i>Findings</i>	51
B. COMPONENT 2: CLASSROOM OBSERVATIONS	51
i. <i>Selection of schools, classrooms and students</i>	52
ii. <i>Data Sources</i>	54
iii. <i>Findings from observation and interview data</i>	56
iv. <i>A closer look at the language production of RFEP students compared to EO students</i>	59
C. COMPONENT 3: ANALYSIS OF CALIFORNIA ENGLISH LANGUAGE DEVELOPMENT TEST (CELDT).....	61
i. <i>Background of the CELDT</i>	61
ii. <i>Step 1: Examining the CELDT scores themselves</i>	63
iv. <i>Step 3: The academic achievement of the students meeting the probable reclassification criteria</i>	65
VI. DISCUSSION OF FINDINGS.....	68
VII. POLICY IMPLICATIONS.....	70
VIII. RESEARCH IMPLICATIONS AND NEXT STEPS.....	71
IX. REFERENCES.....	73

I. INTRODUCTION

Under the auspices of the U.S. Department of Education, ARC Associates, in conjunction with San Francisco Unified School District (SFUSD), conducted a study to determine when assessments of academic achievement are appropriate for English language learners (ELLs).¹ This study was funded through a grant from the U.S. Department of Education, Office of English Language Acquisition and Academic Achievement for Limited English Proficient Students (OELA), #T292B010001.

The study was carried out from September 2001 until June 2004. This report will present findings from three major components of the study. In the first component, we will describe the results of our analysis of student assessment data compiled by SFUSD to explore the relationship between content area testing scores and students' English language proficiency. In the second component, we will report on the results of classroom observations of English learners to describe their academic classroom performances. In the third component, we will present the results of an auxiliary study of California's English language proficiency test that grew out of project findings. To conclude, we will discuss findings from across the three components and explore both policy implications as well as directions for future research.

A. Background of the Study

Schools in the 21st century face renewed challenges to change the way children are educated. Federal, state and local initiatives as well as business interests (Johnson & Packer, 1987; U.S. Department of Labor, 1992) have called for improvements in the delivery of educational services, greater student achievement, and better preparation of graduates for the workplace. Key to these initiatives has been the standards-based reform effort as articulated in the Goals 2000 legislation and used as the basis for No Child Left Behind (NCLB), the reauthorized version of the Elementary and Secondary Education Act signed into law in January, 2002. Standards reform, which targets

¹ Throughout this document, English language learners are sometimes referred to as ELLs. In the case of data and information from the San Francisco Unified School District and occasional other sources, the term "English learner" or EL is used to mean the same population of students.

learners across the educational spectrum, focuses on having all children achieve to high standards. An integral part of this movement involves describing explicitly what those high standards should be within content areas. Thus, during the 1990's, discipline-specific standards were developed across content areas including mathematics, science, English language arts, history, English as a second language, civics, foreign language, the arts, geography, economics, and social studies.

Content-area standards have served as a basis not only for setting learning targets, but for connecting curriculum, instruction, and assessment efforts. By articulating what all students should know and be able to do within specific disciplines, standards have provided a blueprint for creating grade-level curricula, determining the content of textbooks, and designing assessments. Note that while assessment has always been an important component of educational programs, in standards-based reform efforts, assessment has taken on increased importance since it provides the basis for accountability systems designed to monitor whether all students reach targeted standards (LaCelle-Peterson & Rivera, 1994). However, content standards were developed, for the most part, with mainstream, English speaking, middle class students in mind. As reform efforts have been undertaken in diverse contexts, it has become apparent that schools need to grapple with how to incorporate the needs of English language learners into this agenda (Miramontes, Nadeau, & Commins, 1997). While the rhetoric of content-area standards refers to their use with all students, the standards do not address such instructional issues as how to teach content material while students are still acquiring a second language, nor do they address assessment issues such as how English language learners can demonstrate knowledge of content material when tested in English.

This lack of attention to how English language proficiency may impact content area assessment systems is a recurring issue within the standards movement and poses problems for schools facing sweeping demographic changes within their classrooms (Rivera, 1999). Across the United States, increasing numbers of immigrant children have been entering K-12 classrooms. According to the 1990 census, the number of school-age children speaking languages other than English rose 39% during the 1980's; those who reported not speaking English very well increased by 83% (Numbers

and Needs, 1993; Waggoner, 1988). Another analysis of student enrollment data from the 1985-86 school year through 1991-92 revealed a 68.6% growth in the number of English language learners while total school student enrollment only increased by 6.1% (Olsen, 1994). As the data show, the total proportion of language minority students has increased greatly in U.S. schools.

As schools struggle to serve all their students and to assist them in achieving to high standards, they need to have more confidence in the tools they use to ensure equitable opportunities. Assessment is an important component within educational systems and is a major provision of NCLB, which requires schools to track the achievement of students over time and show demonstrable yearly progress in raising student achievement in reading and math. Thus, in addition to providing feedback to teachers, students and parents about individual student growth and progress, student achievement data are being used to determine how well or poorly schools are serving those students. It is because of this accountability function that assessment serves a key role in the standards-reform effort. Moreover, many states, including California, now expect all students to pass standardized tests tied to new standards in content areas (i.e, Language Arts, Math) for promotion and graduation (Marzano and Kendall, 1996). However, English language learners present enormous challenges to the assumptions underlying most test construction.

One such assumption is that test takers will share a high degree of comparability. That is, they will have had similar life experiences, cultural and/or linguistic similarities, and equitable learning experiences (President's Advisory Commission on Educational Excellence for Hispanic Americans, 2000). Clearly, the accountability function of assessment rests on such comparability. Yet if students who take tests are at varying degrees of English language proficiency, that comparability disappears. The research questions addressed in this study were designed to determine at what point along the second language acquisition continuum educators can regain confidence in the results of standardized tests conducted in English with English language learners. The answers to those questions will have an impact on educational programs both locally and at the national level.

B. Review of Literature

As American schools strive to provide equitable educational opportunities for all, educators face the challenge of how to best serve students learning English as a second language at the same time that they are learning content knowledge in a range of subject areas (Butler and Stevens, 1997). American educators and researchers have pointed out the increasingly “high stakes” nature of mandated standardized tests, with test scores being used to decide, for example, whether students enter gifted and talented program, move on to the next grade and graduate from high school (Coltrane, 2002; Butler and Stevens, 2001; Hakuta, Butler and Witt, 2000; and Kopriva, 2000). At the school level, scores on standardized tests are used in accountability systems to determine school effectiveness. As large scale, standardized tests acquire more importance, their role in the education of English language learners (ELLs) has come under scrutiny and criticism. The Council of Chief State School Officers, for example, asserted that current large-scale academic assessments for ELLs, “should be supplemented to avoid test bias” (Kopriva, 2000). Furthermore, Butler, Orr, Gutierrez and Hakuta (2000) described the SAT-9 test used in California as “not designed to measure English development and academic achievement for LEP students” (p. 152).

Educators and researchers have come to agree upon the following four topics of research as being central to understanding how to create appropriate and valid assessments for ELLs, particularly with regard to achievement in the content areas.

1. When should students be tested in their first language?
2. What kinds of test accommodations are appropriate for English language learners?
3. What are the best methods to measure the development of an English language learner’s English until that student reaches a level at which s/he can take an achievement test in English?
4. At what point does testing an English language learner yield meaningful results? (Butler and Stephens, 2001; National Clearinghouse for Bilingual Education, 1997).

The central question of this study: “When along the language proficiency continuum does testing a student in the second language in the content areas yield meaningful and valid results?” relates to question four above. Underlying this question are assumptions based upon current understanding of the relationship between English language proficiency and academic achievement. In the sub-sections that follow, we review the literature and research relevant to the assessment of English language learners, including models and theories of language proficiency, research focusing upon the link between language proficiency and academic achievement and current research on the assessment of English language learners, with an emphasis on assessment in the content areas. Thus, this review seeks to frame the central question of this study in a thorough understanding of the complexity of defining English language proficiency and recognizes that this complexity has a direct impact on the creation of valid, equitable content area assessments for English language learners.

i.. Early Measures of Language Proficiency

Measuring language proficiency is, in itself, a complex task. Much of the research has centered on oral skills as the main construct for measuring students’ growth in English proficiency. In an examination of oral English language proficiency tests, Graham and Acosta (1979) found a lack of consistency in the results from different measures of oral English proficiency. They reported correlations ranging from .39 to .82 between the Bilingual Inventory of Natural Language (BINL), the Language Assessment Battery (LAB), the Language Assessment Scales (LAS), and the Bilingual Syntax Measure (BSM). When they examined whether there was consistency in the way pairs of tests classified students, they found averages of 56% for the BINL and BSM, 65% for the BINL and LAS, and 77% for the BSM and LAS. Gilmore and Dickerson (1979) compared the classification of students among five tests used in Texas - BINL, BSM, LAS, the primary Acquisition of Language (PAL), and the Schutt Primary Language Indicator Test (SPLIT) and also found lack of agreement in the classification of students.

Cervantes and Nakano (1979), comparing classifications of oral language proficiency of second language student in the first and third grades who were tested on

both the BSM and the LAS, found classification agreements only 14% of the time. Ulibarri, Spencer and Rivas (1981) in a study for the California State Department of Education compared students' English level classifications on the BSM, LAS, and BINL. The percentages of agreement between pairs of tests ranged from 27% for the BSM and LAS to 65% for the BINL and LAS. While they found lack of agreement in the classifications of second language students among the BSM, LAS, and BINL, Ulibarri, Spencer, and Rivas (1981) noted an underlying agreement among the tests. When they used raw scores from the various tests instead of the nominal classifications such as NEP, LEP and FEP, they found moderate to high correlations between the tests, offering evidence that the tests were measuring similar oral language abilities. Thus, although the cutoff scores recommended for classifications and program decisions were not comparable from one test to another, the tests may have been measuring the same language abilities.

ii. Oral Language Proficiency and Academic Achievement

In the same way that the studies concerning the classification of language minority students by oral English language proficiency tests do not agree in their findings, studies dealing with oral English language proficiency tests as predictors of academic achievement have often been contradictory. The student results on oral English proficiency tests do not always correlate with student results on standardized tests of academic achievement in English. Oral English language ability seems only mildly related to academic achievement. For minority language students in grades 1, 3, and 5, DeAvila, Cervantes, and Duncan (1978) found a correlation of .41 between oral language proficiency scores on the BSM, LAS, and BINL and reading scores on the Comprehensive Tests of Basic Skills (CTBS). For students in grades 7-12 they reported a correlation of .49.

The Ulibarri, Spencer, and Rivas study (1981) reported correlations between three oral language tests (BSM, LAS, and BINL) and academic achievement as measured by the CTBS, the California Achievement Test (CAT), and the Stanford Achievement Test (SAT). For Hispanic students in grades 1, 3, and 5, the correlations

ranged from .08 to .47. The oral language tests were poor predictors of academic achievement.

In a study of second through sixth grade students, Saville Troike (1984) examined oral language as a predictor of academic achievement. From the scores on three English language tests (the Northwest Syntax Screening Test, the Functional Language Test, and the Bilingual Syntax Measure), she found low correlations with reading scores on the CTBS of .291, .138, and .258 respectively. From videotaped student interviews, quantitative indices of English language proficiency, such as verbosity, measured length of T-units, and grammatical accuracy, were gathered. Saville-Troike found that "accuracy in English morphology and syntax in spoken language appears to make little difference in academic achievement"(1984: 216). In fact her study pointed out that the students least likely to succeed academically as would be predicted by their oral language scores were some of the strongest academically, and conversely, "the lowest academic achievers in our sample were among the most successful at interpersonal communication" (1984:216). None of the these results provides support for a strong relationship between oral English language proficiency and academic achievement.

On the other hand, Ulibarri, Spencer, and Rivas in plotting the results mentioned above state that the "figures show an increase in average achievement scores as one goes from NES to LES to FES on the LAS, BSM, and BINL" (1981:71). Also two-way analyses of variance for the LAS and the BSM indicated that the categories of NES, LES, and FES do discriminate between increasing levels of academic achievement. In other words, there is correspondence between growth in oral English language proficiency and growth in academic achievement in English.

Studies in reading achievement in English and oral English language proficiency offer support for a positive relationship between oral language proficiency and academic achievement. Tregar and Wong (1981) examined the relationship between native language (L1) reading comprehension and second language (L2) reading comprehension. They used the Oral Dominance Test of the Boston Public Schools as a measure of oral English language proficiency, and they used cloze reading tests of the same school district to measure reading comprehension. For Hispanic and Chinese

students in grades 3 through 5, they noted a higher correlation between L1 reading comprehension and L2 reading comprehension than between L2 oral ability and L2 reading comprehension. However, for students in the middle grades (6, 7, and 8), there was a higher correlation between L2 oral ability and L2 reading comprehension than between L1 reading comprehension and L2 reading comprehension.

In another analysis of English reading achievement and English language proficiency among language minority sixth grade students, Fetter (1983) found significant differences in reading achievement in English between students classified as limited English proficient (LEP) and those classified as fluent English proficient (FEP) and as English only (EO). The LEP students scored, much lower in reading achievement than the other two groups. In an analysis of the reading subskills, EO scores were always the highest with LEP scores significantly lower. However, the FEP scores tended to be close to the EO scores. In this study, as in the previous study, academic achievement of older students increased as oral English language proficiency increased. These studies would appear to offer evidence for a positive relationship between oral language proficiency in English and reading achievement in English for older students.

However, there is no strong consensus in these studies as to the nature of the relationship between oral English language proficiency and academic achievement. From one point of view, the lack of agreement in the various studies examining the relationship between oral English language proficiency and academic achievement lies with the oral language proficiency tests (Dieterich, Freeman, and Crandall, 1979, Rivera and Simich, 1981). From another point of view the inconsistency in the results lies in the lack of an adequate theoretical framework with which to relate language proficiency and academic achievement (Cummins, 1980).

iii. Inadequacy of the Testing Instruments

The confusion in the findings concerning the effectiveness of oral English language proficiency as a predictor of academic achievement may be attributed to the language proficiency tests themselves. The development of these assessment instruments has been influenced by various trends in linguistic theory. According to Rivera and Simich (1981), early assessment measures were influenced by linguistic

perspectives and psychometric methodology which focused on testing discrete aspects of language. Because of the influence of these trends the instruments placed a great deal of emphasis on formal grammar and syntax. These oral language proficiency tests tended to measure only surface aspects of language and emphasize linguistic features (Dieterich, Freeman, and Crandall, 1979). Thus, these tests have been criticized for not assessing language as it is used in real communication (Carroll, 1981).

There has been a trend toward communicative language testing based on the sociolinguistic concept of "communicative competence". This concept was offered by Hymes (1972) as a criticism of and corrective to Chomsky's conception of competence. For Chomsky, linguistics is concerned with a theory of competence which attempts to set forth the linguistic rules that can generate and describe the grammatical sentences of a language. According to Hymes (1972) this conception of competence is too narrow: it leaves no place for the idea of the appropriateness of what is said to the situation and context in which it is said. Competence should also account for "when to speak, when not,....what to talk about, with whom, when, where, and in what manner" (Hymes, 1972: 277-278). Therefore, in addition to assessing linguistic competence, oral language proficiency tests should also measure communicative competence.

Oral English language proficiency tests have been developed to measure communicative competence, and the dual construct of linguistic competence and communicative competence has been used in investigations of academic achievement in English. With Hispanic high school students, Politizer and Ramirez (1981) used the Bahia Oral Language Test (BOLT) to assess linguistic competence and they used a test developed by themselves and others at the Center for Educational Research at Stanford (CERAS) to measure communicative competence. They noted that both linguistic competence and communicative competence were related to academic achievement as measured by the number of completed high school graduation proficiency requirements. The correlations were .54 for linguistic competence and .51 for communicative competence.

Using the same oral English language proficiency instruments and the same assessment of academic achievement, McGroarty (1982) found through principal components analysis that linguistic competence and communicative competence were

both related to academic achievement. However, she found that the degree of English linguistic competence was able to explain more of the variance in academic achievement than the degree of English communicative competence.

Yet, there are contradictory findings even with these communicative approaches to testing. Hayes (1982) examined the relationship between oral English language proficiency and academic achievement on the CTBS by third grade Hispanic students. To assess English oral language proficiency Hayes used the CERAS Bilingual Balance Test to measure linguistic competence and the CERAS communicative competence test to measure communicative competence. Concerning the relationship between oral English language proficiency and academic achievement as measured by reading in English, Hayes found that the communicative competence measure contributed more significantly to the prediction of levels of academic achievement than did the linguistic competence measure. This result contradicts the earlier findings by McGroarty (1982). However, Hayes noted that the combination of linguistic competence and communicative competence measures was the best predictor of academic achievement.

Thus, even the studies using more recent tests of communicative aspects of oral English language proficiency have produced the same contradictory findings as to the relationship between oral language proficiency and academic achievement which the studies using more traditional tests produced. In one case, linguistic competence was related to academic achievement; in the other, communicative competence was the best predictor of academic achievement. This may be an indication that the problem lies deeper than the language proficiency tests themselves whether they be grammar based or communicative tests. It may lie in the lack of a clear conception of what language proficiency is. According to Cummins (1980) the confusion in language proficiency testing is due to the lack of a theoretical framework that adequately relates language proficiency and academic achievement. Cummins argues that

there has been relatively little inquiry into what forms of language proficiency are related to the development of literacy skills in school contexts and how the development of literate proficiency in L1 relates to the development of literate proficiency in L2. (1980: 27).

This suggestion that the confusion lies with theoretical models of language proficiency introduces a new dimension to the exploration.

Because the assessment of the English language proficiency of minority language students and their placement in bilingual instructional programs has been based on the results of oral English language proficiency tests, studies concerning the consistency of these tests in measuring language proficiency and in classifying minority language students were examined. Among oral English language proficiency tests little agreement was found in the language classifications to which students were assigned. Studies of oral English language proficiency tests as predictors of academic achievement produced similar inconsistent results. Even for older students, when there did seem to be a relationship in that academic achievement in English increased with growth in oral English language proficiency, the results were not conclusive.

This lack of consistency in results was attributed to the inadequacy of the tests, which were grammar-based and focused on language form. Yet, even when communicative tests of oral English language proficiency were examined, the same contradictory results appeared. The lack of agreement in findings may be linked with the lack of clear conceptual frameworks of language proficiency. Bachman (1990) points out that advances in language testing are stimulated by advances in the understanding of language acquisition. If indeed the problem lies with the models of language proficiency that underlie the tests, then an examination of these models may highlight which theories would provide a grounding for the relationship between English language proficiency and academic achievement. The next section of this review will examine some of these current conceptualizations of language proficiency.

iv. Models of Language Proficiency, Language Use and Language Ability

Looking historically at the attempt to define language proficiency, North (2000) points out that the evolution of definitions of language proficiency were much affected by developments in linguistics and sociolinguistics, including Chomsky's previously mentioned distinction between competence and performance (1965) and Hymes' development of the concept of "communicative competence" (1972). North makes a useful distinction between proficiency definitions which incorporate competence and

language proficiency as measured and defined by performance tests. For example, Taylor (1988) uses the term “communicative proficiency” and defines it as the ability to make use of competence (p. 166).

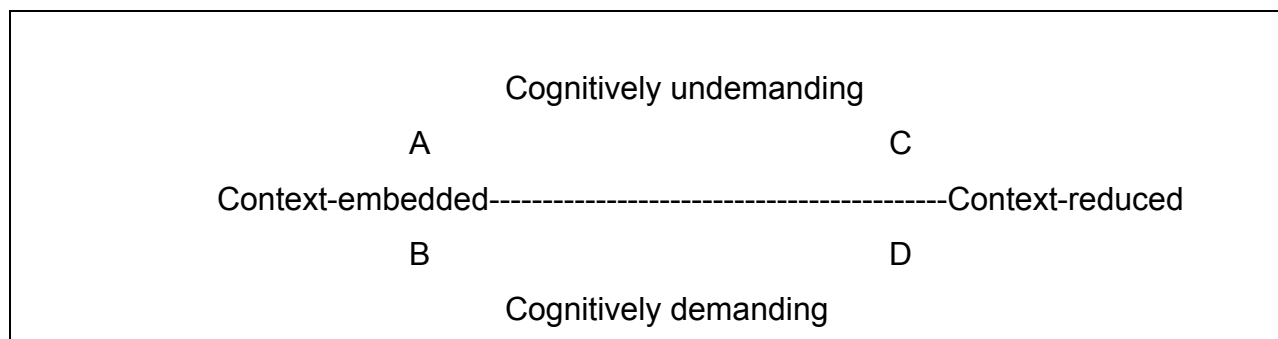
To further complicate matters, North points out that in America, the term language proficiency is historically associated with a “proficiency movement,” stemming from the American Council on the Teaching of Foreign Languages’ (ACTFL) publication of a set of proficiency guidelines (ACTFL, 1986). As a result, North hypothesizes that researchers such as Bachman (1990) begin to avoid the term proficiency all together, using instead terms such as “communicative language ability” or “communicative language use.” A model for such language use, particularly in relation to the creation of a standardized assessment, remains elusive, however, in part because a language users’ competence, by definition, may not be completely captured by a test and thus cannot be easily operationalized into proficiency descriptors or levels.

The models of language use reviewed in this section are associated with second language acquisition and language assessment but do not presume to encompass fully the complex number of factors and variables that affect English language learners when they take a test. Numerous studies reviewed in this section seem unaware of North’s international and historical perspective on the once narrowly defined term, “language proficiency,” but we have made every effort to use the terms of the researchers, which range from “proficiency” to “development” to “acquisition” to “use” to “ability.”

Cummins (1981b), for example, offers a model of second language acquisition and use that provides a contextual framework for language use offering explanatory power within an educational setting. He examines language use from two perspectives: first, in terms of the degree of contextual support available for expressing or receiving meaning, and then in terms of the degree of cognitive involvement. He describes the degree of contextual support as a continuum ranging from "context-embedded" on the one hand to "context-reduced" on the other and the degree of cognitive involvement as a continuum ranging from "cognitively undemanding" to "cognitively demanding". The cognitively undemanding communicative tasks are those in which " the linguistic tools have become largely automatized (mastered) and thus require little active cognitive involvement for appropriate performance" (1981b:13). On the other hand, cognitively

demanding tasks require active cognitive and linguistic involvement. Cummins (1981b:10) attempts to clarify these aspects of language proficiency with the following diagram of the dual continua:

Figure 1
Range of Contextual Support and Degree of Cognitive Involvement in
Communicative Activities



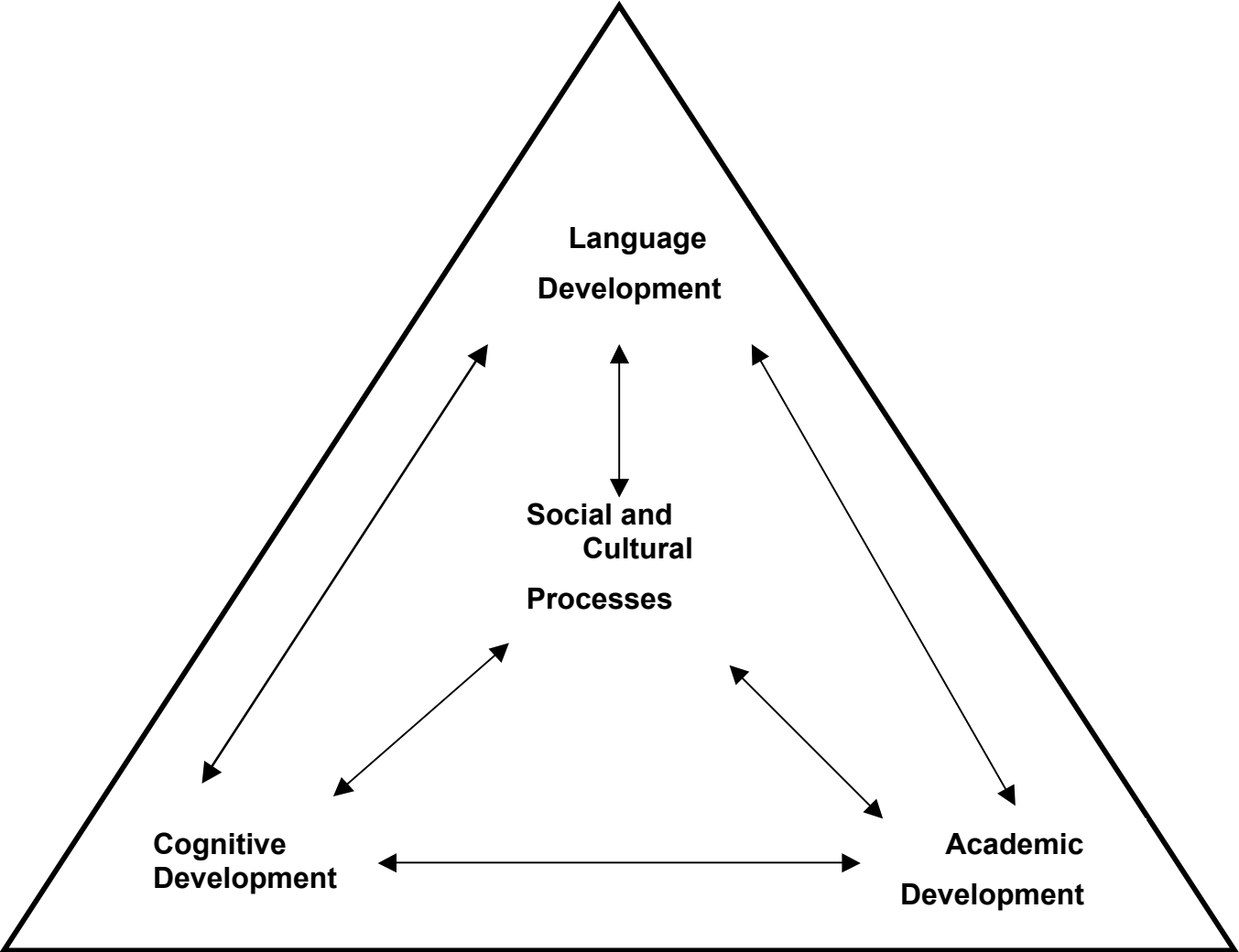
Using this model of language proficiency, any communicative task can be described according to its cognitive demand and its contextual support. Thus some communication is cognitively undemanding and context-embedded, that is, accompanied by gestures and intonation. This is the language used for face-to-face encounters and basic informal communication, and it can be located in quadrant A of the diagram. Other communication is cognitively demanding and context-reduced, that is accompanied only by linguistic cues. This is the language of schooling and literacy. It can be located in quadrant D of the diagram. For Cummins, the major aim of literacy instruction is to develop the students' ability to deal with context-reduced, cognitively demanding texts. In this view, language proficiency involves command of both basic communication skills and cognitive academic skills.

The question arises as to the application of this model to the case of second language students. According to Cummins, almost all people attain proficiency in the basic communication skills of everyday life in their native language. This aspect of language proficiency is acquired through interpersonal interaction and involves the negotiation of meaning between speakers/interlocutors/writers. Developmentally, Cummins sees these interpersonal basic communication skills as the foundation for the more intrapersonal cognitive academic skills; however, they may not be predictive of the cognitive academic skills. Also, it is in the area of basic communication skills that second language learners first gain mastery of a new tongue.

The literacy related and school related aspects of language proficiency are more complex. This cognitive academic dimension of language proficiency correlates highly with general I.Q. and academic achievement. This is the aspect of language proficiency where the differential development among minority language students is most noticeable. Cummins sees this cognitive academic aspect of language proficiency not only as connected with intellectual development but also as interdependent across languages, forming a common proficiency underlying both languages in the case of a bilingual person. Because of this common underlying proficiency at the level of cognitive academic performance, development in the primary language of a student can promote development in the second language. Therefore, the degree of literacy in the first language should predict the degree to which second language literacy skills will be developed in the second language.

Cummins' model of language proficiency remains largely defined in terms of linguistic skills and factors, and the lack of a strong sociocultural component is one of its disadvantages. In contrast, Collier (1995) incorporates a sociocultural aspect into her conceptual model of language acquisition for school, making it the heart of a “prism” composed of four components: sociocultural, linguistic, academic and cognitive (p. 2). She argues that students’ acquisition of a second language in school is mitigated constantly by social and cultural processes occurring in their past and present everyday lives such as societal patterns towards an immigrant group, for example cultural stereotyping, intergroup hostility and the subordinate status of a minority group, or post traumatic stress symptoms exhibited by war refugees such as depression, anxiety and aggression (p. 5). See Figure 2 on the following page.

Figure 2 Acquiring a Second Language for School



(Copyright, Virginia P. Collier, 1994)

As mentioned previously, Bachman (1990, 2002) shifts his terminology away from “language proficiency” when creating models of “language use “ and “language ability.” In relating his models to language testing, Bachman (1990) recognizes that test methods and the background characteristics of language learners influence scores as much as the students’ language skills. Thus, his models of language use and ability come closer to recognizing the influence of a sociocultural context for the language learner than Cummins. And similar to Colliers’ model of language acquisition for school, it distinguishes between linguistic and academic development.

Bachman and Palmer (1996) envision language as “the creation or interpretation of intended meanings in discourse by an individual, or as the dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation” (p. 61). They extrapolate upon this definition in their model of language use, emphasizing two forms of interaction: internal and external. Individual language users’ language knowledge, topical knowledge, affective schemata, personal characteristics and metacognitive strategies interact internally to create meanings and these same individual attributes of the language user interact externally with either the characteristics of the target language use domain or the language assessment domain. See figure 3 on the following page.

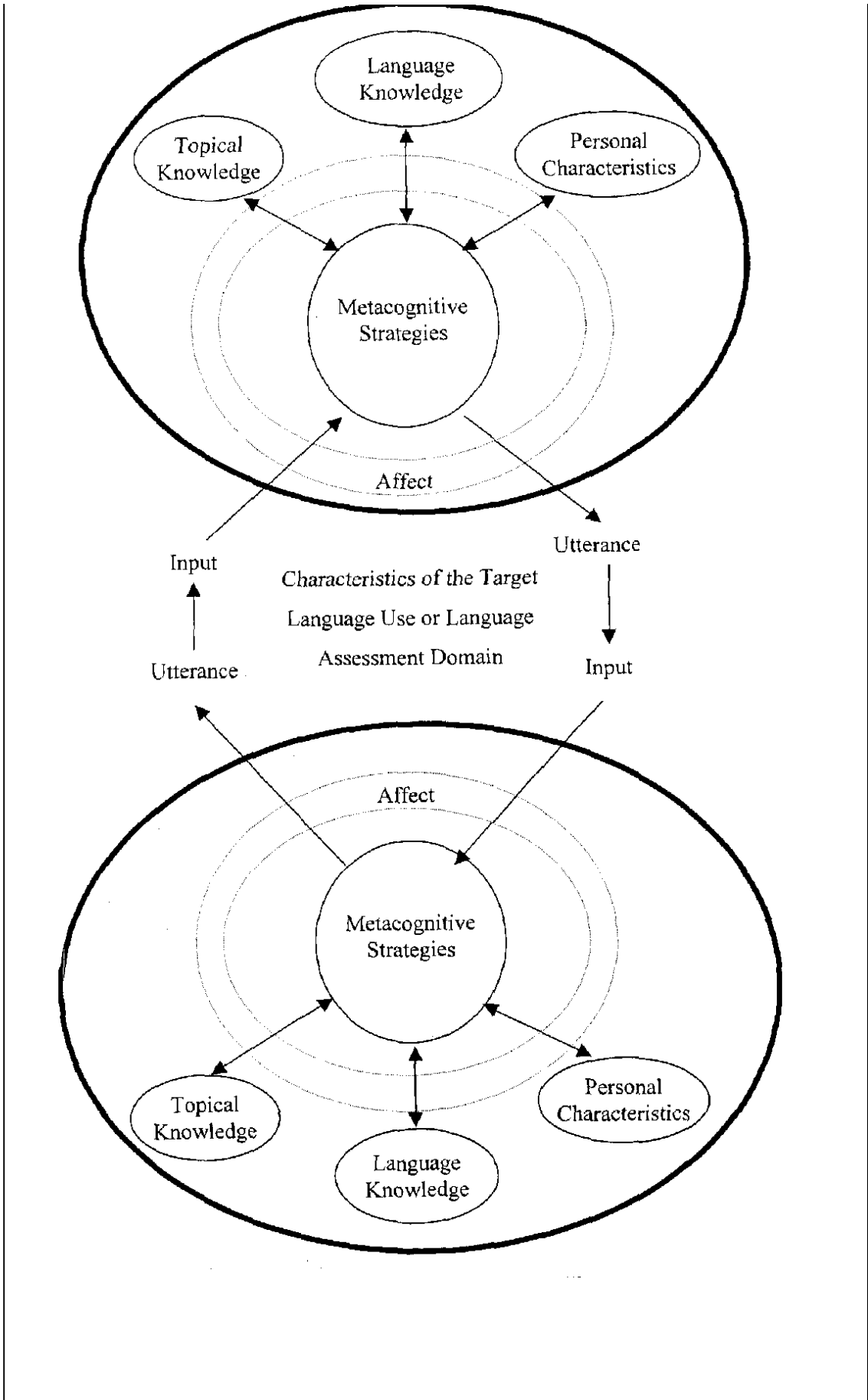


Figure 3: Model of Language Use (Copyright, Bachman and Palmer, 1996)

Building upon this model of language use, Bachman divides language ability into two main categories: language knowledge or the domain of information related specifically to language ability, and strategic competence or the metacognitive strategies engaged during language processing. Within language knowledge, there are two areas: organizational knowledge and pragmatic knowledge. These areas are further broken down: Organizational knowledge consists of grammatical and textual knowledge. Pragmatic knowledge includes propositional, functional and sociolinguistic knowledge. Under metacognitive strategies, he describes three types: assessment, goal setting, and planning.

As noted earlier by North (2000), there remains a divide between theoretical models of language use, proficiency and ability and the assessments designed for English language Learners. Perhaps this explains, in part, why when relating his model of language ability to testing, Bachman (2002) argues for the pragmatic efficacy of performance assessments that call for what Douglas (2000) describes as “the integration of specific purpose background knowledge and language ability.” While recognizing that language ability and topical knowledge are separate (if overlapping) in his model and acknowledging that it might sometimes be appropriate to assess the two attributes separately, Bachman (2002) appears to prioritize creating assessments that resemble real world tasks to ensure what he labels, “the authenticity of the assessment tasks” (p.15). Hakuta, Butler and Witt (2000) also recognize the continuing divide between theory and the practice of assessment, making an admittedly “rough distinction” between oral English proficiency and academic English proficiency (p. 4).

In applying considerations of the models of language use, ability and proficiency described above to the present ongoing study, it seems essential to consider what definitions of language use, ability or proficiency underlie the assessments used in each district. For examples, students in the San Francisco Unified School District took the recently developed California English Language Development Test (CELDT), an assessment based upon the K-12 English Language Development Standards adopted by the California State Board of Education in 1999. These California ELD standards

were created by a committee of educators, who were asked to both use the national TESOL (Teachers of English to Speakers of Other Languages) standards as a base and link the California ELD standards to the already created California English Language Arts standards (Kuhlman and Nadeau, 1999). Does the CELDT reflect the definitions of language use, ability and proficiency presented in Cummins' and Bachman and Palmer's models? In fact, Hakuta, Butler and Witt (2000) identified the English Language Development (ELD) Standards and the assessments based on these standards as being a helpful future point of reference for further "policy-relevant" study (p. 15).

v. Research on Academic Achievement and Language Proficiency

The following section reviews model based basic and evaluation research. Model-based research on academic achievement in a second language is of two types: basic research and evaluation research. Hakuta and Snow (1986) distinguish between "basic research" which analyzes the linguistic, psychological, sociological, and cultural processes in human development, and "evaluation research" which usually compares educational program models or teaching methods to examine their effectiveness.

a. Model-based Basic Research

In a summary of research with Finnish immigrant children in Sweden undertaken for a report commissioned by UNESCO, Skutnabb-Kangas (1979) reported that Finnish students arriving to Sweden between 9 and 11 years old and having had several years of schooling in Finland, achieved at higher levels than Finnish students who arrived between 6 and 8 years old who had little or no schooling.

Baral (1979) examined the academic achievement of middle school age, recent immigrants from Mexico. Contrary to expectations based on previous research that recent immigrants outperform native-born Mexican-Americans, the native-born Mexican-Americans as a group consistently achieved better than the recent immigrants. In the discussion of possible explanatory factors, Baral explained that, although on the surface the achievement results argue that initial native language schooling may not have helped this group of recent immigrants learn English and succeed, this study was

consistent with the Finnish study and the ideas of language Cummins was beginning to formulate. Baral concluded that it could be variables such as the total length of schooling in the native tongue and the timing of the shift to English that explain the results and that "the full benefits of the native language approach may only be attained after prolonged instruction in the home language throughout the primary years" (1979: 12).

In a study on language use and achievement, Dolson (1985) examined the effects of Spanish language use on academic achievement of Hispanic students. Dolson compared the academic achievement of Hispanic fifth and sixth grade students whose families had maintained Spanish as the main home language with students from homes where a switch to English had occurred. The group from home context that maintained Spanish significantly outperformed the group from the homes of language shift in five of ten academic measures and had higher mean scores on the remaining five measures. What is of interest to the present study is that in the area of oral English proficiency based on a teacher judgment instrument (SOLOM) and in the area of reading achievement as measured by the CTBS, there were no significant differences between the groups.

In a study of second, third and fourth grade students of Italian-American backgrounds, Page and Ramirez (1986) examined the effects of bilingual home environments on academic achievement. Three home language environments were delineated: monolingual children in monolingual homes, monolingual children in bilingual homes, and bilingual children in bilingual homes. Additionally, the students were categorized as monolingual, passive bilingual, or active bilingual. There were no significant differences in achievement by home language environment. However, the academic skills of passive bilinguals were significantly lower than those of monolingual and active bilingual children. This would seem to support Cummins' (1979) threshold hypothesis that a child must reach a certain level of active bilingual competence before realizing the benefits from being bilingual.

A few studies have examined the length of time it takes students learning a second language in school settings to reach the level of average academic achievement of native speakers when schooled only in the second language, English. Cummins

(1981a) examined the time needed for immigrants to Canada to acquire English academic language proficiency when taught in that second language after arrival. Cummins reanalyzed the data from a study by Ramsey and Wright (1974) which involved 1,200 immigrants in the Toronto school system in grades 5, 7, and 9. Based on the age on arrival (AOA) of the immigrants Cummins worked out an average length of residence (LOR) according to the grade level of the student. He found that LOR rather than AOA has a substantial effect upon the rate at which immigrant students approach grade norms and that it takes at least five years on the average. Although in this study the older learners acquired the English academic language proficiency more rapidly than younger learners, the age on arrival did not significantly affect the eventual performance at grade norms. Cummins speculated that this finding may not be generalizable outside of the Canadian social context noting that it differed markedly from the Skutnabb-Kangas report discussed earlier.

Extending the literature on the same topic, Collier (1987) studied the average length of time required for 1,548 immigrants to the United States to reach native-speaker norms on standardized tests (50 NCE) when taught only in English after arrival. The subjects were "advantaged" second language learners, that is, they were at age appropriate grade level in their primary language when they immigrated and they were middle class or upper middle class. These students were assessed as non-English proficient (NEP) when they entered school implying that they had no previous exposure, or very little, to English. Similar to Cummins (1981a), Collier found that it took four to eight years to approach the 50 NCE in reading, language, science, and social studies. In this study age on arrival had an effect in that students arriving between the ages of 8 and 11 were the fastest achievers with seven year old arrivals slightly below their performance. LEP students arriving at ages 5 and 6 were projected to require at least 2 to 3 more years to reach comparable performance while students arriving at ages 12 to 15 scored significantly lower than all the others.

The cross sectional data on advantaged LEP students reported by Collier (1987) was further examined by Collier and Thomas (1988). One more year of data was added, and sex differences were reported. The sex differences were not practically significant, but the findings on arrival age further confirmed the earlier analysis. ESL

graduates with age on arrival of 8 - 11 years reached the 50th NCE within 3 to 6 years' length of residence, depending on the subject area and grade level when tested. ESL graduates with age on arrival of 4-7 years were significantly below the appropriate level for their length of residence, and 12 to 15 year old arrivals had the lowest scores of all.

Another study of bilingual/ESL graduates noted the relationship between time and achievement. Examining the academic achievement of 544 Chinese-American students in grades 4, 5, and 6, Lee (1985) found no significant differences between the reading achievement of Chinese-American students formerly classified LEP and having received bilingual/ ESL services, Chinese-American fluent English proficient students and non-minority background English speakers. In all areas of academic achievement, scores increased as the number of years of fluency also increased. According to Lee "this finding can be related to the work of Cummins (1981b) who argues that it commonly takes between five and seven years before language minority students are able to achieve at a level comparable to language majority students" (1985: 119).

Cummins and Nakajima (1987) conducted a study of the language proficiency and academic achievement of Japanese students in Toronto as part of the five-year Development of Bilingual Proficiency project at the Modern Language Centre of the Ontario Institute for Studies in Education (Harley et. al., 1990). Similar to Cummins earlier findings (1981a), it appeared that students required about 4 years of instruction after arrival to Canada to attain grade level norms in English reading. However, there was a tendency for students who arrive at the age 6-7 to make somewhat more rapid progress toward grade level norms than those who arrive at older ages. They also found that, when length of residence is controlled, a significant relationship between reading achievement across languages is evident. While writing performance was found to be less closely related across languages than was reading, Cummins and Nakajima speculated that it may have been a function of the different types of measures (standardized reading tests and non-standardized writing tests). Generally the data were consistent with other studies in supporting the interdependence of cognitive academic skills across languages and the time needed for attaining grade norms in English academic tasks.

Another study of language interaction focused on low-achieving Hispanic students and contrasted perceived and actual linguistic competence. Cummins and Miramontes (1989) reported on a qualitative analysis of the linguistic performance of four Hispanic bilingual students whose language dominance was not clearly defined and whose academic achievement was perceived by teachers to be limited by their language abilities in both English and Spanish. Through participant observation over a two-month period, the researchers found that although none of the students were fully proficient in English, the teachers had underestimated their linguistic and academic abilities in both English and Spanish based on their classroom performance. The strengths and capabilities of the students were not being used in the school setting.

b. Model-based Evaluation Research

An effort at systematic application of a theory of bilingual education and implementation of programs based on the implications of the theory was undertaken by the California State Department of Education and selected local school districts in 1982. This "Case Studies Project" was based on Cummins' conceptualization of language proficiency and its cross lingual dimensions (Gold and Tempes, 1987). Literacy and content instruction was initially introduced in the primary language in grades one and two along with an ESL component. English literacy was introduced in the third grade and content instruction in both languages continued through the upper elementary grades. The operationalized achievement goal of the project was that all students initially identified as LEP would score at or above the 50th percentile in the areas of reading and mathematics on the Comprehensive Tests of Basic Skills after seven years in the program. After two years 44% of the second grade students initially identified as LEP were scoring at or above the 50th national percentile in reading. Thirty-nine percent of the third graders were above the 50th national percentile. The authors reported steady increases in the percentage scoring above the national average from 16 % the first year, 27% the second year, 36% the third year, and 39% the fourth year. In the less language dependent area of Mathematics, more than three-fifths of the participants had scored above the 50th national percentile.

The evaluation of a preschool program in Carpinteria, California (Keating, 1984) highlighted the oral dimension of academic language proficiency. Spanish speaking preschool children historically scored much lower on school readiness tests compared with their English-speaking peers. The tests emphasized the comprehension and production of the instructional language used in school, i.e. following directions, responding to questions, and so on. The project students were exposed to a variety of language enriching experiences in their mother tongue. At the time of elementary school entry these children outperformed Spanish speaking controls on both English tests and Spanish tests. Additionally, they compared favorably with English-speaking controls on readiness skills for kindergarten.

Krashen and Biber (1988) reviewed evaluation results from five school districts in California implementing programs based on the Cummins' framework. On standardized tests three of the districts showed scores at or above the national average by grades 5 and 6. While the data presentation in the review was inconsistent and the nature of the comparisons groups not always clear, which makes interpretation of results difficult, the review does report achievement data which is important for understanding the growth of second language learners.

Hakuta, Butler and Witt (2000) examined test data from four school districts, two in California and two in Canada. Their study builds upon the findings of Cummins and Collier, concluding that limited English proficient (LEP) students take three to five years to achieve oral English proficiency and four to seven years to develop academic English proficiency. The authors also noted in their analysis of the two California school districts, the significant effect of socioeconomic status on English language acquisition (p. 13).

vi. Research on the Assessment of English Development and Academic Achievement of English Language Learners

The studies thus far conducted on the assessment of English language learners have emphasized the importance of understanding the effect of academic “language load,” particularly in standardized test items, on the performance of ELL test takers. In

their analysis of test data comparing ELL and native speakers of English, for example, Abedi, Leon and Mirochi (2000) found that the gap between ELLs and native speakers increased as the “language load,” of the assessment tools increased. However, while defining language load as “linguistic complexity” of test items, the authors admit, they did not analyze the test items used (p. 4). The authors do make note of a concurrent study by a colleague that has begun analyzing the linguistic demand of test items, creating a language demand rating scale which considers the difficulty of, for example, the type of test question and complexity of vocabulary and syntax (Bailey, 2000).

Other studies have identified a mismatch between English language proficiency as measured by standardized tests commonly used in school districts and the academic language load of content tests. (Butler and Castellon-Wellington, 2000; Stephens, Butler and Castellon-Wellington, 2000) Additional studies have begun to identify and describe academic language in the content areas (Bailey, Butler, LaFramenta and Ong, 2001), which may lead to a greater understanding of what the threshold of language ability for ELLs is that allows them to adequately demonstrate their content knowledge on assessments.

In addition, Abedi, Leon and Mirochi (2000) confirm the importance of studying family background characteristics of ELLs when interpreting their assessment scores, noting in their analysis that parent education is the most important variable in their analysis of how language background influences the standardized test scores of ELLs (p. 44).

vii. Integrating Factors in the Assessment of English Learners

The studies of language proficiency and academic achievement that have been reviewed reflect the confusion surrounding models and definitions of language proficiency, use, and ability, particularly when an attempt is made to apply these definitions and models to the assessment of English language learners. If we accept the assumption that operationalizing a definition of language proficiency in an assessment of English language learners requires assessment designers to focus upon pragmatic, performance based skills, we are still left grappling with, in the interpretation and analysis of these assessment scores, how to take into account sociocultural factors

such as the effect of the authenticity of the assessment task and characteristics of the test taker, particularly having to do with socio-economic background.

In an overview of language testing research, Bachman (1991) points out that a language test score cannot be interpreted as an indicator of language ability only. Instead, interpretation must take into account, “the characteristics and content of the test tasks, the characteristics of the test taker, and the strategies the test taker employs in attempting to complete the test task (p. 677). The process of interpretation is further complicated if we take into account the interaction of the factors described.

In addition, Bachman (1991) presents a definition and model for judging the “authenticity” of language testing tasks. His definition identifies two types of authenticity, situational and interactional. Situational authenticity has to do with how closely the test method resembles real life target language usage. Interactional authenticity takes into consideration how involved the test taker’s language ability, including language knowledge and strategies, are when completing the test task.

Even if a definition of language proficiency as defined by tests given to English language learners in a school district or state seems grounded in current theory, it remains to be seen whether this definition of English language proficiency or development is reflected in content area assessments.

viii. Redefining English Language Proficiency

a.. Academic language proficiency in the classroom

The foundation of the present study is in providing a clear description of the language proficiency required for students to be successful within the academic contexts of the classroom and, as a proxy of classroom performance, on assessments. While discussions of language proficiency are often centered on a generalized notion of basic language ability, a series of recent related studies in measurement issues involving the academic achievement of ELs has provided more situated notions of language proficiency. In these studies, language proficiency is described in terms of the functions of language required in specific contexts (Hawkins, 2004).

Academic Literacy is a term that has been defined differently in many academic contexts (Scarcella, 1999). Bilingual educators are well acquainted with the BICS/CALP

(Cummins, 1981) distinction of language acquisition. Thus it raises the issue of what we really mean by the term and how it applies to what our students need to be academically successful. At the secondary level, development of higher order literacy skills is required if students are expected to master local, state and national standards for language arts and other content areas. Level of English literacy has also been used as the best criterion for determining readiness to meaningfully participate in English language assessments (National Research Council, 1998).

b.. Academic language on tests

Research has also described the language features found on language proficiency tests and compared them to descriptions of the language of classrooms and content assessment. In a study designed to describe and compare the language and performance of 7th grade ELs on tests of language proficiency and achievement (Stevens, Butler & Castellon-Wellington 2000), text analyses revealed limited correspondence between the two tests. Butler & Castellon-Wellington (2000) suggest that “competent performance” on a commonly-used language proficiency test such as the LAS may not provide sufficient evidence for determining whether or not ELs can handle the academic language load of content assessments. Bailey and Butler (2003) assert that “Academic language proficiency (ALP)” needs to be clearly defined using not a single proficiency or standardized test but including national and state content standards, English as a second language standards and information about teacher expectations and school language (p. 6). By creating such a framework, Bailey and Butler reason that we will be able to identify a “validity/language proficiency threshold,” which up until now has been elusive because individual school districts and states have used such varying requirements for identifying English proficiency (p. 37).

Fillmore and Snow (1999) examined prototype test items for a high school qualifying examination for one of the 23 states that has adopted this requirement. Their analysis reveals that students must have competence in academic English to do well on the test. The language used in the tests is not different from that ordinarily used in school textbooks and academic discussions about science, mathematics, literature or social studies. Selected examples of what students must know in order to deal with these tests successfully follow.

Students must analyze texts, assessing the writer's use of language for rhetorical and aesthetic purposes, and to express perspective, mood, etc.; extract meaning & information from texts, and to relate it to other ideas and information; evaluate evidence and arguments presented in texts, and to critique the logic of arguments made in them; recognize and analyze textual conventions used in various genres for special effect, to trigger background knowledge, or for perlocutionary effect; recognize ungrammatical and infelicitous usage in written language, and make necessary corrections to texts in grammar, punctuation and capitalization; use grammatical devices for combining sentences into concise and more effective new ones, and use various devices to combine sentences into coherent and cohesive texts; compose and write an extended, reasoned text which is well developed and supported with evidence and details; interpret word problems--recognizing that in such texts, ordinary words may have quite specialized meaning; extract precise information from a written text, and devise an appropriate strategy for solving the problem based on information provided in the text.

Holistic assessments have the potential to solve the problem of tracking individual achievement (Hewitt, 1995), particularly of English language learners. However, the validity and reliability of the rubrics and scoring guides intended to convert such data into quantifiable data are often unknown. Thus, standardized tests are used for assessing curricula, programs, or for providing peer comparison information. In the context of high-stakes, standards-based reform efforts such as the one in California, standardized tests fulfill the accountability function required by standards-based reform and so are also used to determine whether schools are doing their jobs effectively. More information is needed, however, to find out when such assessment are appropriate and can provide meaningful information about what English language learners know and can do.

II. RESEARCH QUESTION

This study sought to answer the following research question as stated as a priority of the original grants announcement:

When along the language proficiency continuum does testing a student in the second language in the content areas yield meaningful and valid results?

The above question is a complex one. As the literature review in the previous section shows, many factors affect ELLs' language and content area learning, Therefore, we extended our inquiry into this question via the following sub-questions:

- How many years of learning English in U.S. schools does it take before content area testing of ELLs can yield meaningful and valid results?
- How does degree of English language proficiency affect the testing validity of different content areas?
- What are the profiles of the ELLs who acquire a sufficient level of English to take part in meaningful content area testing?

A. Assumptions

Our approach and findings were based on assumptions which, in and of themselves should be made explicit and examined. For example: 1) To operationalize the research questions, the study used the CELDT to measure language proficiency and the SAT-9 to give us information about content area achievement. 2) The study defined content area achievement, using student achievement on the SAT-9 exam, with students' reading scores being the independent variable and students' math scores being the dependent variable. Thus, the study assumed that the SAT-9 would be a valid indicator of content achievement and that the SAT-9 scores would account for the differential impact of language proficiency levels in the different content areas. 3) The findings of this study were based on math test scores, not other content specific test scores such as science and history.

B. Additional Considerations

i. Explaining “meaningful” and “valid”

While any test result can be “meaningful” in the sense that an interpretation can be made from the scores, the corresponding notion of “validity” is more problematic, especially when we narrow our reference point to high stakes testing and accountability. Test results may not have a direct relationship with what goes on in the classrooms, for example, especially when the tests students take are commercially prepared assessments loosely tied to a state framework and not to a specific school’s curriculum. The classroom observation results from our study illustrate the discontinuity between classroom performances and high stakes testing.

ii. Defining “language proficiency”

Beginning with the review of literature and ending in the discussion of findings, one of the recurring issues in this study is how academic researchers, test designers, state boards of education and local school districts differ in their definitions of English language proficiency and what Bailey and Butler (2003) have described as “academic language proficiency (ALP).

iii. Examining the heterogeneity of ELLs as a group

While limited English proficiency is the common denominator creating the targeted group of students in this study, recent research (Abedi, Leon, & Mirocha, 2000) provides evidence that ELLs as a group are not a homogenous group. Within group differences in performances on tests suggests that we need to examine student background variables more carefully to understand variation in student achievement. As a school district representing a wide variety of languages and ethnicities, SFUSD offered a wealth of data to explore such within group differences. Our analyses in the present study indicate some difference in the pattern of achievement between Spanish and Chinese home language students.

III. RESEARCH SETTING

A. The State Framework for Accountability

California, like other states across the U.S., has engaged in standards-based reform efforts. Standards, adopted or developed across content areas, inform the state frameworks which determine the direction of instruction and assessment throughout the state. While some states developed their own assessments, California adopted the Stanford-9 test as its required, standards-based assessment for district accountability. In 1999, California established a new assessment program, the Academic Performance Index (API), which ranks every school in the state according to a formula that combines school characteristics (e.g., socioeconomic status, ethnic/ racial makeup) and SAT-9 scores². Schools are assigned a numerical score between 200-1000, a rank between 1 (worst) and 10 (best) and a growth target for the year. Schools and teachers that reach these targets will be given financial incentives by the state. Schools that do not reach the targets face sanctions.

The API also defines subgroups—e.g., statistically significant ethnic/racial and socioeconomically disadvantaged populations— at each school. Each subgroup is also assigned a growth target which schools must meet in order to receive the awards. However, the API does not include ELLs as a separate group. The API is the fruit of a decade in which standards and assessment have taken educational center stage in California. While the focus on high standards and accountability for all students has been a positive step in the direction of national school reform, little has been provided to help schools step up to the new expectations. Nor has this effort taken into consideration the testing issues surrounding second language learners. While schools need to be held accountable for the education of English language learners, the state has not made clear at what point it is appropriate to test them in English for content area achievement.

² As of the 2002-03 school year, the SAT-9 exam was replaced by the CAT-6 exam.

B. The District Context

SFUSD is a microcosm of the national phenomenon of increasing student diversity. The district is located in northern California in the city and county of San Francisco, an urban, coastal city with a long history of attracting immigrants from Asia, Central America, South America and elsewhere. One of the most densely populated cities in the United States, its racially and linguistically diverse population of approximately 770,700 lives within a 46.4 square mile area.

The student population of SFUSD reflects the city's diverse population. According to the February 2002 report of the district's Bilingual Education Task Force, approximately one-third of the total district enrollment consisted of English Learners (ELs). These ELs spoke 64 different languages with the five largest groups being Chinese (various dialects) 43%, Spanish 37%, Filipino 4.9%, Vietnamese 3.1% and Russian 2.7%. Half of the students in the district were language minorities, many of them redesignated Fluent English Proficient (FEP) students with ongoing language and academic content area needs.

Many ELs live below the poverty line in a city with an exaggerated cost of living aggravated by an influx of wealthy Silicon Valley/e-commerce professionals. In 1999, ELs comprised 39% of students in the district's Title I program; of them, Latinos were most likely to qualify for Title I; however, both Latino and Chinese students were over represented in relation to their proportion of the total school population. Parents with limited education and maximum economic stress struggle to prepare their children adequately and to support them once they enter school. Faced with the double tasks of language and content acquisition, this group of students is possibly the most socially and academically vulnerable at every turn.

In order to fully grasp the context for English language learners in San Francisco, one must understand the history of civil rights in San Francisco's public schools. First, it should be noted that after a series of suits against the district, San Francisco Unified School District is currently under a consent decree mandate to desegregate its schools. This legal consent decree is the result of a suit by the San Francisco National Association of Colored People (NAACP) begun in 1978 and concluded in 1983 (Biegel, 2003). Second, in 1974, SFUSD was the setting for the Supreme Court Decision, Lau

versus Nichols, which stated that SFUSD had denied equal opportunity to non-English speaking students by “failing to provide them with any special language instruction” (Arizona State Department of Education, 1977). The Lau versus Nichols decision, in conjunction with the Equal Educational Opportunities Act of 1974, created a mandate to implement educational remedies for language minority students, including bilingual education programs.

In spite of the passage of Proposition 227, the “English only” initiative in 1998, which caused the dismantling of required maintenance bilingual programs in all California public schools, SFUSD continues to offer a plethora of educational alternatives for English language learners and its general population of students, including two way bilingual immersion programs in Spanish, Chinese (Cantonese), Korean and Filipino. Bilingual instruction also occurs in Japanese. SFUSD maintains an explicit commitment both to the instruction of English learners and bilingual education. The “Guiding Principles” of the Bilingual Education Task Force state that SFUSD seeks to:

Provide and promote the opportunity for all students to develop competence in two or more languages, academic competence, and a positive self-image and attitudes towards other cultures... (SFUSD, 2002)

IV. STUDY DESIGN

A. Overview of EL data

Our study utilized the vast amount of English proficiency, content area testing, and student background data of the 27,683 English learners (ELs) of SFUSD to explore systematically the relations among content area testing scores, the students’ English language proficiency, and a variety of background variables. This large number of ELs allowed us to analyze and compare subgroups without the limitation of small subgroup sizes. The study drew on test data from school years 2000-2001 and 2001-2002 and on classroom observation data collected during the spring of 2003.

In the 2000-01 school year, there were 18,624 ELs enrolled in SFUSD. The following is the distribution of the ELs according to the major language groups and grade levels (K-11).

HOMELANG	K	01	02	03	04	05	06	07	08	09	10	11
CAMBODIAN	11	9	18	13	17	17	13	15	19	22	12	9
CHINESE	844	1120	1014	1019	803	650	446	413	363	437	435	502
FILIPINO	50	58	69	78	79	71	61	92	77	71	74	68
JAPANESE	24	27	16	16	11	4	4		6	2	8	5
KOREAN	12	17	21	19	16	16	8	6	8	10	8	8
SPANISH	725	750	799	801	729	642	471	432	389	455	366	351
VIETNAMESE	59	65	60	51	49	37	40	34	32	27	32	45
OTHER	103	115	146	120	134	120	92	97	97	110	98	105

B. Student language proficiency status

When first enrolled in the school district, every student is processed by the central intake center where the student's demographic information is collected and his/her English language proficiency is assessed. According to the assessment, students are classified into three categories:

- English Only (EO) — Student is from an English speaking background.
- Initial Fluent English Proficient (IFEP) — Student is from a non-English background but is proficient in English.
- Limited English Proficient (LEP³) — Student is from a non-English speaking background and is not proficient in English.

There is a fourth category. As a LEP student progresses and acquires English proficiency in school and satisfies a set of criteria established by SFUSD, he or she may be re-classified as Re-designated Fluent English Proficient (RFEP). Thus, all students in SFUSD fall within the four language proficiency categories. For the purpose of this study, we defined English Learners (ELs) to include both LEP and RFEP students. Note that we included the scores of RFEP students so that the array of EL student performances represented the full spectrum of proficiency levels.

Table 1 shows the numbers of students in the four categories by language backgrounds in spring, 2001. According to Table 1, 30% of the total SFUSD student population, or 17,912, were limited English proficient (LEP) and 16% of the total SFUSD student population, or 9,771, were students who had been reclassified as fluent English proficient (RFEP). Thus, 46% of the total SFUSD student population, or 27, 683

³ We use the term LEP to conform with federal and state guidelines related to English learners.

students were the students who were English learners in the schools. The LEP group was receiving English language development services, and the RFEP students had been recipients of these services in the past but were considered fluent and not in need of further services at the time of the study. This large number of English Learners allowed the project to analyze and compare subgroups without the limitation of small subgroup sizes.

Table 1: Number of student in San Francisco Unified School District, April 2001.
Home Language

	English	Chinese	Spanish	Other	Total
English Only	22763	205	175	202	23345
IFEP	2861	2180	1772	1598	8411
LEP	41	7886	6840	3145	17912
RFEP	31	6165	1495	2080	9771
Total	25696	16436	10282	7025	59439

Table 1 also shows that Chinese-speaking students⁴ accounted for 51% and Spanish-speaking students accounted for 30% of the total EL students in the district. Our study focused on these two groups as both groups had sufficient numbers to ensure the power of analysis within subgroups.

⁴ Chinese-speaking students represent a range of different Chinese dialects. The largest dialect groups are Cantonese and Mandarin.

C. San Francisco Unified School District Testing Framework

The following SFUSD assessment data sets⁵ were identified as sources for the project to utilize in addressing the research question:

Assessment	Students	Language	Content	Response Type
Stanford 9	All Grades 2-11	English	All: Reading, Math, Language, Spelling; HS: Science, Soc. Sci.	Selected Response
California English Language Development Test	LEP Grades K-12	English	English Language Development: Oral (Listening/Speaking), Reading, Writing	Oral Production, Selected Response, Short Answer, Essay

In addition, SFUSD collected and compiled background information for all students at their first enrollment in the school district. The data are maintained and updated yearly in a centralized database, Student System Master Student Record (SSMS). SSMS consists of the following information: Birth Date; Birth Place; Year of Entry into US; Parents' Education Background; Home Language; Family Income Indicators; English Language Proficiency; GPA; and additional information.

D. Project meetings and advisory teams

i. Research Team

The research team met twice monthly for the first year and monthly in year two to review analyses, interpret results, and plan additional analysis. The research team consisted of staff members of ARC Associates and SFUSD:

Project Director: Sau-Lim Tsang (ARC)

Senior Research Associate: Anne Katz (ARC)

Senior Research Associate: Patricia Low (ARC)

Research Associate: Juan Sanchez (ARC)

⁵ In addition to these tests, SFUSD also collects data from the following tests: California English Language Arts Standards, High School Exit Exam, Language and Literacy Assessment Rubric, Integrated Writing Assessment, and SABE 2.

Co-Director: Jim Stack (SFUSD)

Senior Statistician: Sophia Luk (SFUSD)

Statistician: Stella Szeto (SFUSD)

The research team recruited two sets of researchers and practitioners with expertise in the areas of assessment and second language learning to review study design, proposed data sets, and on-going analyses of data. These teams also provided another point of view that allowed project staff to reflect on these data from the perspective of other school contexts and geographic regions.

ii. The Study Team

The study team met with the research team bi-monthly to review the progress, help interpret findings, and guide the analysis. The study team members were:

Mary Ellen Gallegos, Executive Director, Multilingual Program, SFUSD

Carolyn Hofstetter, Assistant Professor, University of California, Berkeley

Ritu Khanna, Executive Director, Research, Planning and Accountability, SFUSD

Robert Linqianti, Senior Research Associate, WestEd

Lydia Stack, Supervisor, Multilingual Programs, SFUSD

iii. The Advisory Committee

In addition, an advisory committee consisting of national experts in areas related to the study also assisted and guided our analysis, reviewed our findings, and suggested policy questions and recommendations. The advisory committee members were:

Frances Butler, Senior Research Associate, CRESST/UCLA

Fred Davidson, Professor, University of Illinois, Urbana

Patricia Gandara, Professor, University of California, Davis

Haggai Kupermintz, Assistant Professor, University of Colorado

James Purpura, Associate Professor, Columbia University

Charlene Rivera, Director, Center for Equity & Excellence in Education, George Washington University

The advisory committee met three times during the course of this study. The first meeting was held on November 9, 2001, the second meeting on September 30 2002, and the last was on November 15, 2003.

V. DATA ANALYSIS AND RESULTS

A. COMPONENT 1: Achievement Patterns of ELs and EOs Over Time

i. Test data

This section presents the part of our study that focused on data from the Stanford Achievement Test, Ninth Edition (SAT/9). SAT/9 is the California State mandated achievement test administered to all students for grades 2 to 11. Scale scores of SAT/9 were the main variables of our study. We used the results from the tests administered in the last two weeks of April in 2001⁶. The following SAT/9 scale scores were available:

	2 nd to 8 th graders	9 th to 11 th graders
Reading: Vocabulary, Comprehension	√	√
Language	√	√
Mathematics: Procedures, Problem Solving	√	√
Science		√
Social Sciences		√

Since item statistics were not available from the publisher, we were unable to calculate the reliability of the test and subtest for our sample and sub-samples.

ii. Limitations of data

Our data were constrained in three ways.

⁶ California State mandated SAT/9 testing for all students beginning in 1999. SFUSD filed a lawsuit against the State claiming the mandate of such testing for LEP student was unfair. In 1999 and 2000, the district did not administer SAT/9 to a large number of LEP students whom teachers did not consider to be ready to take this English-only test. A settlement was reached in 2000 to allow the school district to provide accommodations to LEP students enrolled the district for the first year. Testing for the all students started in 2001

1. SAT/9 was administered to 2nd thru 11th graders. Therefore, our analyses were limited to those grade levels.

2. Our preliminary analysis of SAT/9 reading and math data showed that a significant number of students scored at the 1st percentile (Table 2) on at least one of the tests. An examination of the distributions of all Normal Curve Equivalent (NCE) scores showed that these students contributed to abnormal peaks in those distributions. SFUSD personnel indicated that most of these students had probably not made an effort to partake in the testing. Thus, we excluded these students from our analyses⁷.

Table 2: Number of LEP students scoring at the 1st percentile, 2001.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th	11 th	Total
Chinese	12	15	14	11	15	54	20	20	59	81	301
Spanish	64	42	53	28	19	59	35	24	61	50	435
Total	76	57	67	39	34	113	55	44	120	131	736

3. State policy allows test accommodations for first-year limited English proficient students; students' teachers make the determination as to whether accommodations are appropriate. Accommodations include extra or extended administration time, the reading of test items or questions, translation of test directions, the use of bilingual dictionaries. Table 3 shows 51% of 1st year LEP students were provided accommodations during this testing.

Table 3: Number of 1st year LEP students provided accommodations, 2001.

	Chinese-Speaking	Spanish-Speaking	Other LEP	All 1 st yr. LEPs
Number tested	415	318	253	986
With Accommodations	229	191	82	502
% with Accommodations	55	60	32	51

Since accommodated test scores are not acceptable for the purpose of our study and the majority of the 1st year LEP students were provided accommodations, the

⁷ Table 2 shows the number of students who scored at the 1st percentile on either the reading or the math test. While the total number of students (735) in this table may seem large, smaller numbers of students scored at the 1st percentile on a particular subtest. For example, a total of 298 LEP students scored at the 1st percentile on reading.

remaining LEP students would have consisted of a biased sample. Therefore, we excluded all first year students in our study.

Our study sample consisted of 9925 Chinese- and 4890 Spanish-speaking 2nd to 11th graders. Table 4 shows the distribution of the students.

Table 4: Distribution of students in the study sample by grade levels.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th	11 th	Total
Chinese											
LEP	951	849	631	447	295	304	270	293	276	293	4,609
RFEP	1	155	331	574	769	818	678	680	708	602	5,316
Total	956	1004	962	1021	1064	1122	948	973	984	895	9,925
Spanish											
LEP	592	624	543	431	369	322	265	245	202	147	3,740
RFEP	0	17	58	180	130	145	174	160	159	127	1,150
Total	592	641	601	611	499	467	439	405	361	274	4,890

iii. Analyses of Data and Results: Descriptive statistics

First, we generated descriptive statistics of our data set. The following two tables show the mean NCE SAT/9 scores of our student samples.

Table 5: SAT/9 Reading NCE scores and standard deviations in parentheses, 2001.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th	11 th
Chinese										
LEP+RFEP	55.5 (15.6)	49.8 (14.7)	53.0 (16.6)	50.0 (16.5)	52.5 (16.1)	52.7 (18.0)	49.4 (17.6)	43.5 (17.6)	43.0 (19.7)	43.3 (20.1)
LEP	55.5 (15.6)	48.1 (14.3)	47.3 (15.4)	40.1 (14.0)	38.4 (13.3)	34.5 (13.5)	32.5 (12.6)	27.3 (12.7)	24.6 (12.7)	25.3 (14.5)
RFEP	51.1 (NA)	59.1 (13.2)	63.9 (13.0)	57.7 (13.9)	57.9 (13.7)	59.5 (14.4)	56.1 (14.6)	50.5 (14.6)	50.1 (17.2)	52 (16.3)
Spanish										
LEP+RFEP	37.0 (15.3)	36.5 (13.6)	37.0 (15.3)	38.3 (16.1)	36.3 (13.9)	33.6 (15.9)	36.7 (15.5)	30.0 (14.4)	30.0 (14.8)	33.2 (16.9)
LEP	37.0 (15.3)	36.0 (13.4)	35.4 (14.7)	33.8 (13.4)	32.1 (12.0)	27.7 (12.5)	29.8 (12.6)	23.9 (11.5)	23.4 (10.7)	25.6 (13.2)
RFEP	NA (NA)	54.0 (7.5)	52.4 (12.2)	49.2 (16.8)	48.2 (11.8)	46.6 (15.1)	47.1 (13.6)	39.4 (13.3)	38.4 (15.1)	42.1 (16.4)

Table 6: SAT/9 Math NCE scores and standard deviations, 2001.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th	11 th
Chinese										
LEP+RFEP	68.1 (18.2)	68.2 (17.4)	65.4 (18.7)	67.6 (18.2)	68.9 (18.2)	68.4 (19.3)	66.7 (19.2)	70.0 (19.4)	65.0 (19.7)	66.2 (20.3)
LEP	68.2 (18.2)	66.4 (17.3)	60.0 (18.5)	58.5 (17.3)	57.6 (18.3)	55.3 (18.3)	56.3 (19.2)	59.6 (19.8)	57.2 (19.0)	60.0 (20.9)
RFEP	43.0 (NA)	78.3 (14.8)	75.7 (14.0)	74.8 (15.4)	73.4 (16.1)	73.5 (17.1)	70.9 (17.5)	74.6 (17.4)	68.4 (19.1)	69.7 (19.1)
Spanish										
LEP+RFEP	43.3 (18.1)	43.3 (17.3)	41.0 (17.3)	43.1 (17.9)	41.5 (16.2)	38.0 (15.9)	38.6 (14.6)	43.4 (16.2)	39.5 (15.2)	39.0 (17.7)
LEP	43.3 (18.1)	42.7 (16.9)	39.5 (16.8)	38.7 (16.0)	37.4 (14.1)	33.0 (12.7)	33.0 (11.8)	38.8 (13.9)	34.6 (12.6)	34.0 (15.0)
RFEP	NA (NA)	66.3 (13.7)	55.2 (15.0)	53.9 (17.9)	53.3 (15.9)	49.6 (16.7)	47.7 (14.2)	50.6 (17.1)	46.6 (16.0)	45.6 (18.9)

The data show consistently larger standard deviations for the Chinese students' math scores at all grade level indicating a flatter distribution of scores. This finding is consistent with other national data sets that show that while Chinese students represent a high proportion of high math achievers, they also have a high proportion of low achievers (Tsang, 1993).

Table 7: SAT/9 Science and Social Science NCE scores and standard deviations of high school students, 2001.

	Science			Social Science		
	9 th	10 th	11 th	9 th	10 th	11 th
Chinese						
LEP+RFEP	52.4 (15.3)	51.7 (18.7)	52.0 (20.9)	50.4 (18.2)	46.8 (20.2)	55.2 (20.8)
LEP	41.6 (14.3)	38.9 (15.3)	38.5 (17.2)	37.5 (13.6)	33.2 (14.0)	42.5 (16.6)
RFEP	57.1 (13.1)	57.4 (17.2)	59.6 (19.0)	56.0 (17.0)	52.8 (19.6)	62.4 (19.5)
Spanish						
LEP+RFEP	36.8 (13.8)	36.0 (14.9)	35.1 (16.5)	39.1 (14.2)	31.7 (14.4)	41.5 (19.2)
LEP	32.1 (12.0)	30.8 (12.1)	30.1 (13.7)	34.1 (12.5)	28.1 (11.9)	35.5 (16.4)
RFEP	44.3 (13.3)	43.9 (15.1)	41.6 (17.6)	47.0 (13.2)	36.9 (16.1)	49.2 (19.7)

Table 7 shows that the Spanish students are performing lower than the national norm in Science and Social Science. But more importantly for this study, the low scores have led to the low variance of their distributions.

iv. Analyses of Data and Results: Correlations

Next, a series of correlations were calculated between the SAT/9 Reading scores and Math, Science, and Social Science scores.

Table 8: Reading x Math, 2001.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th	11 th
Chinese LEP + RFEP	.70	.69	.73	.71	.65	.68	.66	.65	.59	.56
Spanish LEP + RFEP	.64	.66	.66	.74	.68	.69	.72	.66	.60	.67
National	.73	.78	.78	.77	.81	.76	.75	.69	.65	.70

Table 9: Reading x Science and Social Science, 2001.

	R x Science			R x Social Science		
	9 th	10 th	11 th	9 th	10 th	11 th
Chinese LEP + RFEP	.70	.74	.74	.76	.76	.74
Spanish LEP + RFEP	.63	.65	.67	.59	.54	.66
National	.67	.69	.69	.70	.72	.71

Table 8 indicates that the correlations between Reading and Math for both the Chinese and Spanish students are lower than those of the national sample. For the Chinese students, the scores decrease from Grade 2 to Grade 11 while the correlations for the Spanish students did not show consistent variation.

For the correlations between Reading and Social Science and Science scores (available only for 9th to 11th grades), Table 9 shows the Chinese students have consistently higher correlations than the national sample suggesting that the demand for reading abilities/skills in those two tests might have affected these students more than the students in the national sample. However, this explanation does not hold true for the Spanish students who have lower correlations than the national sample. The low correlations may be the result of the Spanish students' general lower scores, which produce lower variance in the distribution of their scores (see Table 5 and 7 and related discussions).

For Math, we also calculated the correlations (2nd to 8th graders) of Reading scores with the two sub-scales: Procedures and Problem Solving. The Procedures sub-scales consist of items requiring fewer reading comprehension abilities/skills while the Problem Solving sub-scale consists of many word problems requiring more reading comprehension abilities/skills. We hypothesized that the correlation between the Reading and Math/Procedure would be lower than the correlation between Reading and Math/Problem Solving.

Table 10: Reading x Math sub-scales, 2001.

	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th
Chinese LEP + RFEP							
R x Math/Procedure	.50	.51	.61	61	57	59	53
R x Math/Problem Solving	.70	.71	72	70	64	67	68
Spanish LEP + RFEP							
R x Math/Procedure	.51	.52	53	66	57	56	61
R x Math/Problem Solving	.63	.68	68	72	69	70	70
National							
R x Math/Procedure	63	66	67	68	73	67	68
R x Math/Problem Solving	71	78	77	75	79	75	74

The results confirmed our hypothesis. The correlations between the Reading and Math/Problem Solving are consistently higher than those between Reading and Math/Procedure. An examination of the correlations for the national sample shows that the same holds true for these students, indicating that the need for higher reading abilities/skills when doing word problems is consistent across student populations. However, the differences in the correlations of the two Math sub-scales with reading are much larger for the Chinese and Spanish-speaking students than for students in the national sample. That is, the demands of reading comprehension abilities/skills have a greater effect on the Chinese and Spanish students' performance on the problem solving items.

To further investigate our research question, we looked at time in district programs. We identified a cohort of ELL students who entered the district in Kindergarten and had been continuously enrolled in SFUSD schools. Table 14 displays the Chinese and Spanish LEP students by the number of years in school by grade

levels. As discussed earlier we excluded first year students from our sample; thus, these data represent students with SAT/9 scores utilized in our analyses.

Table 11: Number of EL students by grade levels by years in SFUSD, 2001.

	2 nd gr	3 rd gr	4 th gr	5 th gr	6 th gr	7 th gr	8 th gr	9 th gr	10 th gr	11 th gr
2 Years										
Chinese	27	16	26	19	31	31	44	86	72	55
Spanish	56	31	37	38	26	35	34	48	54	37
3 Years										
Chinese	927	31	25	27	43	53	60	88	90	105
Spanish	641	45	25	28	22	17	22	34	34	29
4 Years										
Chinese	26	992	34	27	37	37	49	44	65	100
Spanish	41	635	27	22	18	20	11	18	15	22
5 Years										
Chinese	4	14	901	30	25	23	22	39	37	35
Spanish	4	50	573	28	23	20	13	19	9	18
6 Years										
Chinese	2	1	22	945	35	28	29	29	31	42
Spanish	3	4	54	575	32	30	26	28	17	7
7 Years										
Chinese	1	2	3	21	905	35	37	40	37	43
Spanish	0	1	3	26	390	22	21	25	23	30
≥ 8 Years										
Chinese	1	0	2	4	33	957	736	738	793	712
Spanish	4	3	8	4	52	398	392	368	376	315

An examination of Table 11 shows that the majority (1,568) of the EL students at 2nd grade had been in the school district for three years. These are the students who entered SFUSD at Kindergarten. Similarly the majority (1,627 and 1,474) of LEP students at 3rd and 4th grades had been in the district for four years and five years respectively. Across Table 11, cells with the largest numbers of students from grades 2 to 6 years were selected for our analyses since they were large enough to track year to year retroactively.

Once these sub-cohorts had been identified, we analyzed their SAT/9 data by the number of years the students had been enrolled in the district, which represented the number of years the students had been acquiring English language proficiency in

school. We calculated the correlations of Reading and Math subscale scores from 2nd to 5th grade.

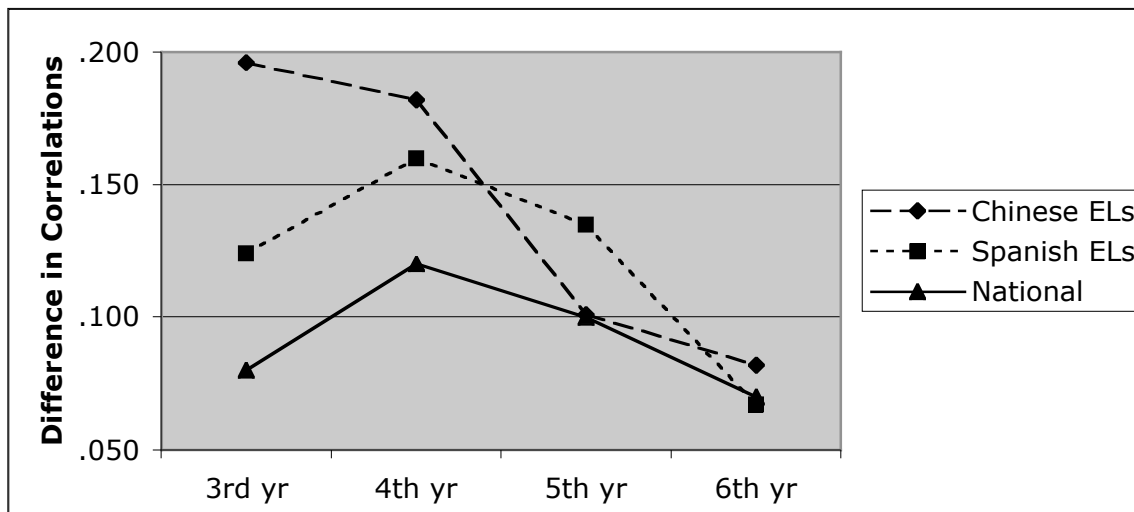
Table 12: 2001 EL students' Reading x Math/procedure and Math/Problem Solving x years in SFUSD vs. national sample.

	3 years 2 nd grade	4 years 3 rd grade	5 years 4 th grade	6 years 5 th grade
Chinese ELs				
(1) Read. x Math/Procedure	.50	.53	.61	.60
(2) Read. x Math/Problem Solving	.70	.71	.72	.69
(2) – (1)	.20	.18	.10	.08
Spanish ELs				
(1) Read. x Math/Procedure	.51	.51	.66	.66
(2) Read. x Math/Problem Solving	.64	.67	.52	.72
(2) – (1)	.13	.16	.14	.07
National				
(1) Read. x Math/Procedure	.63	.66	.67	.68
(2) Read. x Math/Problem Solving	.71	.78	.77	.75
(2) – (1)	.08	.12	.10	.07

We calculated the difference of the correlations between Reading x Math/Procedure and Reading x Math/Problem Solving. We are naming this difference Language Demand Index (LDI). Table 12 shows that the LDIs decrease as the students' years in SFUSD increase (as they move up in their grade levels). The decreases were consistent for the Chinese and Spanish EL students.

To more clearly see the patterning across student populations, we plotted the

Table 13: Comparison of 2001 SFUSD EL students vs. National Sample: (Reading x Math/Problem Solving) – (Reading x Math/Procedure)



LDIs (Table 13) in a graph comparing those of the Chinese and Spanish-speaking students with the national sample. The graph shows that the LDIs for all three groups converge as they advance in their grade levels. The Chinese EL students converge with the national sample at fourth grade (fifth year in school) while the Spanish EL students converge at fifth grade (sixth year in school).

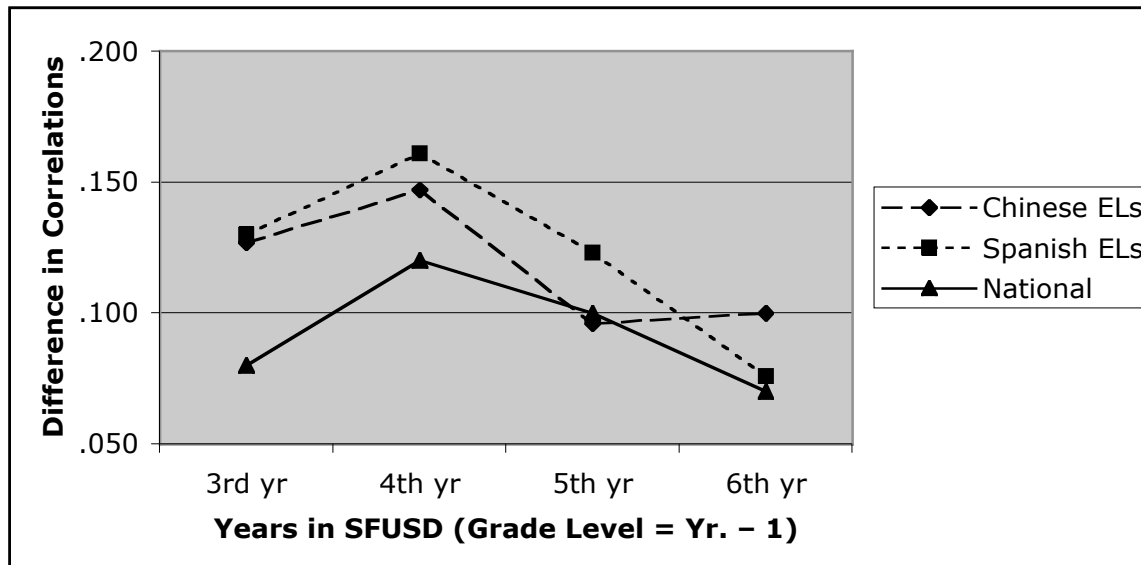
These results suggest that the language demand of reading comprehension abilities/skills for word problems is the same for the Chinese EL students as for students in the national sample in their fifth year of acquiring English proficiency. Similarly, the language demand of reading comprehension abilities/skills for word problems is the same for the Spanish EL students as for students in the national sample in their sixth year of acquiring English proficiency.

To validate our findings, we replicated the analysis using the data from the 2002 testing. We again selected the EL students from 2nd to 5th grades who had entered SFUSD in kindergarten. This cohort of students was completely different from the cohort of the 2001 data. Tables 14 and 15 on the following page show the results.

Table 14: 2002 EL students' Reading x Math/procedure and Math/Problem Solving x years in SFUSD vs. national sample

	3 years 2 nd grade	4 years 3 rd grade	5 years 4 th grade	6 years 5 th grade
Chinese ELs				
(1) Read. x Math/Procedure	.53	.53	.59	.58
(2) Read. x Math/Problem Solving	.66	.67	.69	.69
(2) – (1)	.13	.15	.10	.10
Spanish ELs				
(1) Read. x Math/Procedure	.50	.56	.60	.57
(2) Read. x Math/Problem Solving	.63	.72	.72	.65
(2) – (1)	.13	.16	.12	.08
National				
(1) Read. x Math/Procedure	.63	.66	.67	.68
(2) Read. x Math/Problem Solving	.71	.78	.77	.75
(2) – (1)	.08	.12	.10	.07

Table 15: Comparison of 2002 SFUSD EL students vs. National Sample: (Reading x Math/Problem Solving) – (Reading x Math/Procedure)



The above two tables showed results similar to the 2001 data. The Chinese EL students' differences in correlations converge with the national sample at fourth grade (fifth year in school) while the Spanish EL students converge at fifth grade (sixth year in school).

However, the graph of 2002 data does show one discrepancy with that of the 2001 data. The Chinese EL students' LDIs diverges from the national sample after converging at 4th grade. Further analysis is necessary to understand this difference.

v. Findings

The following are the summary of findings of our analysis of the quantitative data.

- Language loading seems to affect all students, not just EL students.
- However, language loading affects EL students more than EO.
- The effect of language loading on Chinese EL students reduces markedly after 4 years and becomes the same as for EO students in their 5th year of instruction in SFUSD schools
- The effect of language loading on Spanish EL students reduces markedly after 5 years and becomes the same as for EO students in their sixth year of instruction in SFUSD schools

B. COMPONENT 2: Classroom Observations

The findings from our quantitative data suggested that while the test performances of both EO and EL students were affected by the language demands of standardized content-area testing, language impacted the performance of EL students even more markedly. More importantly, our findings indicated that since the patterns of achievement of ELs and EOs converged only after five years of instruction in SFUSD schools, the effect of language loading on EL performance persisted for some time.

The length of time issue raised questions about the academic achievement of ELs in the classroom, particularly the achievement of RFEP students who had been reclassified after less than four years. In our design, we had included RFEP students as a subgroup within our population of EL students so that we could represent the full spectrum of proficiency levels in our sample. Accordingly, we expected these students' performance on the standardized tests to pattern similarly to that of EOs. Yet on the Math section, the data revealed the reverse. The RFEP pattern of achievement was the same as that of ELs, not EOs. According to the testing data, it took four years to reduce

the effect of language loading, and five years for the pattern of RFEP math performance to match that of EOs. If language affected testing performance, what would RFEP student performance look like in an English only classroom, particularly if students were reclassified after only two or three years of instruction?

To get a sense of their academic classroom performance, we conducted a small-scale, qualitative study of RFEP students reclassified after three years and placed in EO classrooms. The study was designed to answer the following research questions:

1. What patterns of participation in academic tasks can we identify for RFEP students in mainstream classrooms?
2. What patterns of achievement can we describe?
3. How do RFEP students deal with instructional tasks? How do EO students in the same class perform?

i. Selection of schools, classrooms and students

After meeting with district personnel familiar with the distribution of EL populations in schools throughout the district, we selected four elementary and four middle schools. Half of the schools (two elementary and two middle) served EL populations that were primarily Chinese and half served EL populations with large numbers of Spanish speakers. The following table shows the school selection plan:

4 Elementary Schools	4 Middle Schools
<ul style="list-style-type: none"> • 2 schools with Chinese speaking RFEP students • 2 schools with Spanish speaking RFEP students 	<ul style="list-style-type: none"> • 2 schools with Chinese speaking RFEP students • 2 schools with Spanish speaking RFEP students

Once schools were identified, students and classrooms were selected. District records provided us with a set of elementary and middle school RFEP students within each school who met our requirement of early redesignation (less than four years).

1.. Elementary Schools

The following table provides an overview of the data collection focus for each elementary school.

4 elementary schools	8 classrooms	25 students
<ul style="list-style-type: none">• 2 schools with Chinese speaking RFEP students• 2 schools with Spanish speaking RFEP students	<ul style="list-style-type: none">• 1-3 third grade classrooms at each school with a minimum of 2 students reclassified RFEP after 3 years	<ul style="list-style-type: none">• 14 RFEP students• 11 EO students

We included both RFEP and EO students in our sampling plan so that we could compare interaction and achievement patterns of both sets of students. Using standardized math test scores, we attempted as much as possible to match academically EO students and RFEP students in each classroom. Observations were scheduled, as much as possible, during Language Arts and Social Studies segments of each class.

2. Middle Schools

The selection process for middle school classrooms differed from that used for elementary classrooms. Since middle school students move from classroom to classroom, we selected teachers and classrooms based on students' class schedule. To maximize our efforts to observe student interaction and academic achievement, we attempted to observe students in content area classrooms that we anticipated would have heavy language loading. Thus, the data set includes observations in Social Studies, Math/Science and GATE classrooms.

While we attempted to match our RFEP sample with EO classmates, we were able to secure consent forms for EO students in only three of the seven classrooms. The following table shows the focus of our design at the middle school level:

4 middle schools	7 classrooms	11 students
<ul style="list-style-type: none"> • 2 schools with Chinese speaking RFEP students • 2 schools with Spanish speaking RFEP students 	<ul style="list-style-type: none"> • 5 social studies classroom at each school • 1 Math/Science • 1 GATE 	<ul style="list-style-type: none"> • 8 RFEP • 3 EO

ii. Data Sources

The selection process for identifying schools, classrooms and students took place during the Fall, 2002, academic year. In Spring, 2003, once parent consent forms had been turned in, data collection began. At each site, the team gathered a range of both observation and interview data.

Classroom observations: Using observation protocols, team members observed patterns of interaction of targeted RFEP and EO students within the classroom. As a baseline for understanding the context of classroom interaction, the first observation at each school in the study focused on the teacher and how the teacher delineated the instructional features of that classroom. Subsequent observations then focused on targeted students and how they worked within that instructional shell. Specifically, we observed student-student and student-teacher interactions and student behavior on classroom tasks.

Principal interviews: At the elementary level, the principal of each school was interviewed at the start of data collection. Principal interviews served dual purposes. They provided an overview of the school context and an opportunity to review program and school services offered to EL students at that site.

Teacher interviews: All teachers were interviewed. The purpose of the interview was to gather teachers' assessments of the level of participation and academic achievement of RFEP students.

Student interviews: Middle school students were interviewed to get their assessment of their level of participation in instruction; they were also asked to identify areas of challenge and school services that were helpful to them.

The following table provides an overview of the interview and observation data collected at both elementary and middle school sites:

Elementary School Data:

	Principal Interviews	No. of Teachers	Teacher Interviews	Number of Students	Observations
Spanish—2 schools					
School A	1	2	2	2 RFEP 2EO	6 (3 per classroom)
School B	1	2	4 (2 per teacher)	3 RFEP 3 EO	6 (3 per classroom)
Chinese—2 schools					
School C	1	3	3	6 RFEP 3 EO	9 (3 per classroom)
School D	1	1	2 (2 per teacher)	3 RFEP 3 EO	3
Totals	4	8	11	14 RFEP 11 EO	24

Middle School Data:

	No. of Teachers	Teacher Interviews	No. of Students	Student Interviews	Observations
Spanish—2 schools					
School F	1 Soc. S.	1	1 RFEP		2
School G	1 Soc. S.	2	1 RFEP 1 EO	2	3
Chinese—2 schools					
School H	1 Soc. S. 1 GATE	4 (2 per teacher)	2 RFEP	2	6 (3 each)
School J	1 Math/Sci. 2 Soc. S.	6 (2 per teacher)	4 RFEP 2 EO	6	9 (3 each)
Totals	7	13	11	10	20

The district provided additional data on the academic achievement of targeted students at each participating school. These data included school grades and scores on standardized tests.

iii. Findings from observation and interview data

Team members developed school profiles for their sites using data collected from observations, interviews and school records. In team meetings, these data from across the eight sites provided a basis for systematically developing a set of preliminary findings. We catalogued the findings under four headings:

- a. Student achievement
- b. Classroom/Academic environment
- c. Family factors
- d. Peer factors

A discussion of each set of findings under these categories follows.

a. Student achievement

While the test data suggested that RFEP students underperformed academically compared to EO students, RFEP students turned out to be high achievers in the classroom context. They were among the best students in the classes we observed. This finding was confirmed through both observations and interviews with teachers.

That said, we do not mean to imply that all RFEP students are alike. Clearly, RFEP students can be characterized not only by their English language proficiency, but by a myriad of other factors. Some RFEPs we observed, for example, were introverts who followed classroom rules and completed tasks according to the teacher's directions. Others were less compliant in the classroom although they completed assignments and only sometimes followed classroom rules. Both kinds of students, however, were successful academically. The following findings illustrate some of the complexity in the face of this achievement.

1. RFEP students varied in their degree of oral language participation. At the middle school level, for example, RFEP students did not speak much in class. This outcome may have been affected as much by their level of introversion/extroversion as by their RFEP/EO status.

2. RFEP students tended to be more focused on academic tasks than other students even in disruptive classroom environments. In our observations, particularly at the middle school level, RFEP students did not tend to engage in off-task behavior.
3. As an off-shoot of good classroom behavior, RFEP GATE students may be placed in GATE classes based more on their behavior, motivation, and determination to succeed than on their ability in subject matter.
4. Lastly, there was a suggestion that some of the Asian students observed may appear to have learned something or to be engaged in a learning task, but they may not in fact be learning the material in depth. Our analysis of individual academic achievement data should provide more clarity on this issue.

An additional caveat must be mentioned about the achievement of RFEP students who have been redesignated within less than three years. Their classroom success may be an artifact of the reclassification process itself and may be localized to districts like San Francisco Unified who base redesignation on more than just test scores. In this district, redesignation criteria include teachers' assessment of students' ability to function in the classroom.

b.Classroom/academic environment

RFEP student success was achieved across varying classroom and academic environments. That is, RFEP students had to be able to identify local norms of classroom behavior and academic achievement in order to succeed. The following points illustrate this finding

1. In the data from our observations, it became apparent that what was considered appropriate or acceptable literate behavior in the classroom varied across classrooms.
2. Related to this notion of classrooms variation, the curriculum demands of several of our primary classrooms were at odds with classroom organization norms. For example,

the literacy activities required by a program like Open Court did not always match the classroom management expectations of the same classroom, for example, the practice of seating students in groups.

3. Different classroom cultures and structures led to different, classroom-specific definitions of “success.” Yet RFEP students adapted to these criteria and expectations and were still successful within these locally-developed notions of academic success.

4. Different classroom expectations and structures led to varying degrees of student participation. In some classes, there was a great deal of student talk; other classes were more teacher-fronted. This raises the issue of the relationship between student participation and academic achievement.

5. Teachers used the same language/ strategies for teaching content with EO and RFEP students. This is not to say that language was ignored. In fact, in both elementary and middle school classrooms we observed some language development work taking place, perhaps underscoring the need of both EO and RFEP students for language development activities as suggested by our quantitative results.

c. Family factors

Parents and, more broadly families, seemed to play a role in motivating RFEP students characterized as high achieving to do well in school. This finding holds across ethnicity but other factors may play a role.

1. For African American and Latino students, parents were educated at the college level whether here in the US or back in the home country. This wasn't necessarily the case with Asian students.

2. Asian RFEP student got a lot more pressure to do well from their parents than did EO Asian students. Consonant with previous findings, our data suggest that as families become more Westernized, they take on Westernized notions about education—that

education is not just about getting the highest score but about developing greater understanding.

3. Increased amounts of English in the home context provided additional support to students; students who were younger siblings were supported by older siblings to learn English faster; older sibling also provided English language role models.

d. Peer factors

For middle school students, performance in schools may also be influenced by the pressure they feel from peers to learn English. Our data suggest that RFEP students learned how to communicate in English even if they did not understand the rules of English. This finding was particularly salient if their peers, more often than not, spoke only English.

iv. A closer look at the language production of RFEP students compared to EO students

In looking at the instructional task participation of the target third grade RFEP students, we became interested in creating a more nuanced picture of these students by looking at additional data from their cumulative folders including: report card evaluations, teacher comments on report cards and other forms, and student scores on the California English Language Development Test (CELDT).

In addition, we analyzed the data from selected classroom observations, looking specifically at the language production of one target RFEP student at each of two schools compared to their EO counterpart in the same classroom. These two target RFEP students were chosen because they earned the highest test scores on the SAT-9.

In both cases, the RFEP student remained on task for a greater percentage of time, reading more and producing more written work than their EO counterpart. In both cases, however, the RFEP student produced less oral language, interacting less with peers and teachers, reinforcing our initial findings that while most RFEP students are

frequently viewed as “good” students, working hard and following classroom rules well, they lack opportunities to practice and improve their spoken English.

In the case of Target Student A, this lag in oral language production does not affect his redesignation. Note in the chart of data below taken from his cumulative folder that while his listening/speaking were assessed as intermediate, he is designated “fluent English” on the CELDT. Note also that on the other assessments used in redesignation in SFUSD, as noted by an asterik, there is more weight given to reading and writing than to speaking.

Target Student A: Redesignated Fluent English Proficient (RFEP) Third Grader

Place of Birth	San Francisco
Parent educational background	High school graduate
CELDT	Reading: early advanced Writing: early advanced Listening/Speaking: intermediate OVERALL: Fluent English
Redesignation Date	September 2002
Standardized Test Achievement Scores*	Math: 60 th percentile Reading: 70 th percentile
LALAR (Oral Section)*	Advanced
Writing*	CELDT

Target Student A’s third grade teacher reinforced the gap between the students’ oral English proficiency and his English reading and writing in his teacher interview and in his comments on the student’s report card:

“_____ is a very respectful and hard worker....continues to work diligently and conscientiously....He is a very quiet, usually quiet and focused, but he is starting to talk more because of his peers....He can express himself but he keeps things closeted....He is reading at grade level or above.”

Due to limitations of time and resources, we were unable to analyze the classroom observations further.

C. COMPONENT 3: Analysis of California English Language Development Test (CELDT)

The first component of this study uses the SAT-9 as an indicator of students' academic content achievement. Therefore the study assumes that SAT-9 is a valid indicator of language proficiency and academic achievement. To operationalize this assumption, we conducted an ancillary study examining the correlation between the California English Language Development Test (CELDT), the examination officially designated in California public schools to identify ELLs and assess their proficiency in English and student achievement on the SAT-9. All the analyses presented in this study are drawn from SFUSD data.

i. Background of the CELDT

The CELDT is currently used throughout California public schools to: 1) identify entering students as English Language Learners (ELL), 2) monitor the progress of ELLs annually, and 3) reclassify students from LEP to FEP. In May 2001, the California State Board of Education approved cut scores for five proficiency levels on the CELDT: Beginning, Early Intermediate, Intermediate, Early Advanced, Advanced. Students receive both an overall score identifying their overall English proficiency and skill area scores identifying their proficiency levels for each test component: speaking/listening, reading, writing.

Table 16 below compares the results of the Fall 2002 administration of the CELDT in SFUSD and the whole state of California.

Table 16: Distribution of Fall 2002 CELDT scores.

CELDT Proficiency Levels	California	San Francisco
Beginning	11%	8%
Early Intermediate	23%	22%
Intermediate	40%	40%
Early Advanced	21%	21%
Advanced	4%	9%

As Table 16 indicates, students in SFUSD perform comparably to students in the rest of the state. Accordingly, reclassification rates are also similar: in 2002, 25% of the students in SFUSD met the review criteria for reclassification based on the CELDT compared with 24% of students in California.

In SFUSD, students are considered for reclassification as FEP on the basis of the following criteria:

1. an overall proficiency level of Early Advanced or above AND
2. proficiency levels of intermediate or above in all three test components (listening/speaking, reading, writing).

Table 17 below shows the percentages of students by grade level meeting the criteria for reclassification.

Table 17: Percentage of SFUSD students meeting English proficiency criteria for reclassification

Grade level	Percentage
Gr 1	18%
Gr 2	12%
Gr 3	8%
Gr 4	17%
Gr 5	27%
Gr 6	23%
Gr 7	31%
Gr 8	37%
Gr 9	37%
Gr 10	43%
Gr 11	47%
Gr 12	49%

The grade level results in Table 17 show an increasing percentage of students meeting the review criteria for reclassification. In Grades 9-12, the percentages move into the area between 40% and 50%, thus leading to the impression that a large number of students should be considered for reclassification. However, in the ancillary study that grew out of our investigation of our research question, the data suggest that there should be some caution with using CELDT as the sole criterion for identification, especially with secondary students. Specifically, we conducted additional analyses to

explore whether these students were ready to be reclassified as proficient in English and, thus, able to benefit from instruction in an English only classroom.

While CELDT scores by themselves might indicate that large numbers of students may be ready for reclassification to English-only classrooms, the examination of additional test data suggests a more complex picture of student readiness. To gather a more complete picture of the CELDT and what it purports to measure, we undertook the following steps:

1. we examined the CELDT scores themselves to see how subsection and overall scores were related;
2. we examined the CELDT scores in relation to academic achievement scores from the SAT 9 in reading and math; and
3. we also examined the academic achievement of the students meeting the probable reclassification criteria.

ii. Step 1: Examining the CELDT scores themselves.

The structure of the CELDT addresses the four traditional language skills, listening, speaking, reading and writing. The listening and speaking are measured together as one component and the reading and the writing are considered separate components. Thus, there are three testing components measuring the four language skills. After taking the test, each student receives a scale score and a performance level for each of the three components. The student also receives an overall scale score and performance level compiled from the three components. Because there are four language skills, each skill receives a weight of 25%. Thus, in calculating the overall score, the listening speaking test component contributes 50% to the overall score. The reading component contributes 25% to the overall score as does writing component.

In order to understand the sense of the overall score, this study examined all of the component test parts. Correlations were run between the three test components to see the relationship of each component to one other. Table 18 below displays the results of these correlations.

Table 18: Correlation among CELDT components.

	Listening	Reading	Writing
Listening	1.00	.47	.48
Reading		1.00	.67
Writing			1.00

As Table 18 shows, the correlation between reading and writing is higher than the correlations between reading writing and listening/speaking. Whereas the reading and writing correlation, at .67, is moderate to strong, the relationship between reading and listening and between writing and listening, at .47 and .48 respectively, are weak.

The difference between these correlations is of concern because the listening and speaking, which has lower correlations with the other components, contributes half of the over all score. Yet our review of literature provides evidence that the reading and writing dimensions of language proficiency are more salient for students' academic needs in classroom contexts.

iii. Step 2: Examining the CELDT scores in relation to academic achievement scores.

This analysis of the components of the CELDT led us to the question of the relationship between an overall score on the CELDT and academic achievement. To address the question of the relationship between language proficiency as operationalized by the CELDT and academic achievement, we looked at correlations between the overall scores of students on the CELDT and their scores on the reading and math portions of the SAT 9. We predicted that the relationship between language proficiency and academic achievement in literacy would be stronger than the relationship between language proficiency and mathematics.

Table 19 provides a dramatic illustration of the lack of correlation between the CELDT and math scores; again, the strength of the correlation diminishes as grade level increases.

Table 19: Correlation of Overall CELDT and Stanford 9 Math.

Elementary			
Gr 2	Gr 3	Gr 4	Gr5
.58	.53	.46	.39
Middle			
Gr 6	Gr 7	Gr 8	
.34	.23	.25	
High			
Gr 9	Gr 10	Gr 11	
.27	.16	.27	

These data support our prediction that the relationship between language proficiency and math would be low, suggesting that the language load in math content areas is lower than in other more literacy-laden content areas.

Table 20 presents a more complex picture of the relationship between language proficiency and academic achievement than our prediction would suggest. Generally, the correlations are moderate across grade levels. However, one would expect that the correlation between English language proficiency and academic literacy would be stronger as English language proficiency increases. We would expect a more linear pattern than demonstrated in our data set.

Table 20: Correlation of overall CELDT and Stanford 9 Reading.

Elementary			
Gr 2	Gr 3	Gr 4	Gr5
.71	.62	.56	.55
Middle			
Gr 6	Gr 7	Gr 8	
.55	.52	.56	
High			
Gr 9	Gr 10	Gr 11	
.53	.44	.56	

iv. Step 3: Examining the academic achievement of the students meeting the probable reclassification criteria.

To complete our analysis of the CELDT data, we decided to look at those students who received overall scores on the CELDT as early advanced or advanced and at least intermediate in all skill areas. These are the students whose CELDT scores

would lead them to be considered English proficient and eligible for review for reclassification.

For these students, we examined their national percentile (NP) scores on the SAT 9 reading section. We looked at the percentage of students scoring below two cut points: the 50th NP and the 30th NP. For the norming group of the SAT 9 reading portion, 50% scored below the 50th NP and 30% scored below the 30th NP. Thus, the number of our students scoring below these cut points was viewed in relation to this norm group.

Table 21: Overall Early Advanced or Advanced and at least Intermediate in all areas.

	Stanford 9 Reading		
	Number of students	% Below 50th NP	% Below 30th NP
Elementary	1743	43.1	16.5
Middle	659	83.6	54.3
High	661	93.9	79.7

At the elementary school level, the students compare well with the national norm group. In fact, only 16% of these students scored below the 30th NP. This would seem to indicate that at the elementary level, students who met the criteria for English proficiency are capable of handling the literacy demands of an academic achievement test. However, when we examined students at the secondary level, at middle and high school, the percentages of students scoring below the 50th percentile are 84% and 94% respectively. Moreover, more than half the students at the middle school level and three fourths of the students at the high school level scored below the 30th percentile level. The data suggest that for secondary students, there is a mismatch between the criteria for English proficiency as defined by the CELDT and language expectations for grade level academic achievement as defined by standardized tests such as the SAT 9.

To examine this further, we redefined English proficiency with more stringent criteria. The following two tables illustrate the results of our analyses. In Table 22, the

criteria were overall early advanced or advanced and at least early advanced in all areas. And in Table 23, the criteria were even more stringent, overall advanced and at least early advanced in all areas. As both tables illustrate, there are still large numbers of students scoring below the 50th NP and in Table 23, the most stringent case, 43% of the middle school students and 58% of the high school students are scoring below the 30th NP. These data clearly illustrate that high level CELDT scores do not seem to indicate an adequate level of academic language proficiency needed by secondary students to be successful in content area standardized tests.

Table 22: Overall Early Advanced or Advanced and at least Early Advanced in all areas.

	Stanford 9 Reading		
	Number of students	% Below 50th NP	% Below 30th NP
Middle	274	78.8	46.4
High	342	90.9	73.4

Table 23: Overall Advanced and at least Early Advanced in all areas.

	Stanford 9 Reading		
	Number of students	% Below 50th NP	% Below 30th NP
Middle	61	67.2	42.6
High	110	83.6	58.2

This now leads to the question of how the literacy demands of the CELDT compared to the literacy demands of the SAT 9 and whether there is a mismatch. We examined the students who were early advanced in CELDT reading and their achievement on the SAT 9 and, separately, the students who were advanced in CELDT reading and their achievement on the SAT 9. Tables 24 and 25 below show these results.

Table 24: Early Advanced in CELDT reading.

	Stanford 9 Reading		
	Number of students	% Below 50th NP	% Below 30th NP
Elementary	1633	43.0	16.8
Middle	1084	88.3	59.7
High	1065	95.5	85.3

Table 25: Advanced in CELDT reading.

	Stanford 9 Reading		
	Number of students	% Below 50th NP	% Below 30th NP
Elementary	520	24.4	8.3
Middle	357	73.7	36.4
High	661	91.6	75.3

The elementary students who were early advanced in CELDT reading are capable of handling the literacy demands of the SAT 9 reading at higher levels than the norm group. However, at the middle and high school levels, once again, we find that 60% of the students at the middle school level are scoring below the 30th NP and 85 % at the high school level are scoring below the 30th NP.

Speculation would lead us to suppose that either the tests are measuring different constructs of reading or they may be measuring the same construct but at different levels of difficulty. Further research is needed to explore the underlying causes for the disparate results reported above—whether it's different constructs or degree of difficulty within the same construct.

VI. DISCUSSION OF FINDINGS

The different pattern of correlations on the two SAT/9 Math tests suggests that more situated notions of English language proficiency are needed to enable educators to make reasoned decisions as to when students can move into English-only instruction

and English-only assessment. While both tests tapped students' math abilities, the results indicate that an analysis of test items from math problem solving and math procedures would probably show that math problem solving requires students to engage in more extensive linguistic processing than when completing math procedures. Task demands both in instruction and assessment will affect student performances, and thus, student outcomes. As students progress in school to classrooms requiring cognitively complex operations, the language demands will also increase.

The data from our study suggest that it took 5 to 6 years of instruction for EL students to overcome the language demands of mathematics word problems in standardized achievement test. The results of this study support the previous work of Cummins (1981a), Collier (1987) and Hakuta, Butler and Witt (2000) which have shown that ELL students need five to seven years of support and instruction in English before they acquire sufficient academic English proficiency to succeed in a mainstream classroom and *to perform similarly to English only students*, particularly at higher grade levels. Our findings suggest that EL students require time to develop levels of academic language proficiency even when students receive high quality language support services such as those delivered by SFUSD. In interpreting these findings, the research setting, SFUSD, should be recognized as unique in state and national contexts for being a long time proponent of English as a Second Language programs and bilingual education. In addition, these findings on the effect of language loading were found to be similar for the SFUSD ELL student population from both Spanish and Chinese speaking backgrounds.

While these preliminary analyses are promising, we need to expand our inquiry to other content areas to see if this hypothesis will stand. We will also need to explore whether this pattern reflects number of years of instruction or is influenced by grade level. Since this cohort consists of students who entered at Kindergarten and remained in the district, additional analyses of students entering at other grade levels are also needed.

A final issue speaks to the issue of mobility. Since many populations of EL students reflect high mobility rates, our results stem from data collected under the best conditions of a stable population of students.

In our ancillary study of the CELDT, we realize that the CELDT was not designed to be aligned with the content area assessments of the SAT-9. It is only meant to assess language proficiency levels and to determine which students need special assistance. Indeed, our research shows that the CELDT is not a valid measure of language proficiency in the academic content areas. Our study of the CELDT inspired by the findings of this project suggest that the CELDT has a ceiling effect when used to indicate language proficiency, placing more weight on social rather than academic language, thus becoming less predictive at higher grade levels where content demand is higher.

Finally, the results of the qualitative component of this study illustrate that testing results do not accurately reflect how ELL students are functioning in classroom settings, even those who have been identified as above average learners of English. In particular, the study results show redesignated ELL students demonstrate less overall oral language participation in class than their EO counterparts.

VII. POLICY IMPLICATIONS

The results of the three components of this study support creating provisions in the current No Child Left Behind legislation allowing ELLs to be exempted from testing for at least three years while being provided with appropriate language support. In addition, the results of this study show that traditional standardized testing scores, which capture a “snapshot” of a student’s English and academic language proficiency at one point in time, cannot fully capture student’s achievement and progress and imply the need for a multidimensional framework of language proficiency, including teacher assessment as in the case of the LALAR, which is a requirement for redesignation of ELLs in SFUSD.

The results of the third component of our study, the study of the CELDT, have implications for the system by which ELLs are reclassified into mainstream English only classrooms. Reclassification of English Learners in California is based upon four criteria:

1) Assessment of English language proficiency, 2) Teacher evaluation, 3) Parent opinion and consultation and 4) Comparison of performance in basic skills (California Education Code, Section 313). Since May 2001, the CELDT has been the mandated assessment for English language proficiency.

In light of what this project has revealed about the CELDT, it seems clear that the CELDT, alone, cannot be used to reclassify students as Fluent English Proficient (FEP), particularly when attempting to measure the academic language proficiency of upper grade students.

In addition, while Grissom (2004) has pointed out that the multiple criteria for reclassification have been designed to prevent ELLs from being exited too early from programs from which they could benefit, he observes that the fourth requirement: performance on tests such as the SAT-9 has proven to be the most difficult barrier, preventing a population of English learner students from being redesignated for long periods of time.

Therefore, discussions about accountability and reclassification of English learners at state and local levels need to include careful count of English learner populations by grade level and length of time in English Learner status in order to determine the efficacy of the testing requirement.

VIII. RESEARCH IMPLICATIONS AND NEXT STEPS

While this study's findings supports the body of research which has shown that English learners need five to seven years before they can attain the academic literacy necessary to succeed in mainstream classrooms, there remains a need to look more closely at specific sub-populations of English learners. Achievement scores of quickly reclassified English learners need to be traced on various content area tests. In addition, the achievement levels of population of English learners that does not reclassify quickly needs to be traced by grade level.

The findings of our study also demonstrate a discontinuity between ELL students' testing results and their performance in the classroom. Therefore, there is a need for a closer examination of how English learners are functioning in classrooms, particularly

with regard to their oral language interaction and participation with peers and teachers. Hawkins (2004) articulates the need for such research, emphasizing the need to look at interaction strategies of English learners as well as the impact of their socio-cultural backgrounds on their ability to learn and engage in the “discourse community” of the classroom and school, whether it be at the elementary, middle or high school level. (p. 21). In addition, Bailey and Butler’s work towards creating a common framework for assessing academic language proficiency (ALP) incorporating understanding of school language demands, standards and testing requirements gets at the need for a broader and more equitable definition of evaluating ELL students’ English language and academic language proficiency.

Finally, further research about how the body of knowledge that exists in the educational research community with regard to how English learners acquire academic literacy and how best to use standardized tests to assess the achievement of English learners can be conveyed in some meaningful and practical way to policy and curriculum makers, the media and even the general public sorely needs to be conducted. Grissom (2004), for example, has identified the misunderstanding or misinterpretation of LEP Reclassification data on the part of the media and others in the case of California’s Proposition 227 passage. Because of the changeover from Title VII funding to the Title III Language Instruction for LEP and Immigrant Students under the auspices of the federal No Child Left Behind legislation, there is an additional need to document how those allocating funding at the state and local levels are using current research in their policy and decision making practices.

IX. REFERENCES

- Abedi, J., Leon, S. & Mirocha J. (2000). Examining ELL and non-ELL student performance differences and their relationship to background factors: continued analyses of extant data. In E.L. Baker (Principal Investigator), *The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives* (pp. 3-49). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Arizona State Department of Education (1977). Task force findings specifying remedies available for eliminating past educational practices ruled unlawful under *Lau versus Nichols*. (ERIC Document Reproduction Service No. ED346082).
- August, D. & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: a research agenda*. Washington, D.C.: National Academy of Sciences, National Research Council, Board on Children, Youth and Families.
- Bachman, L.F. (1990) *Fundamental considerations in language testing*. New York: Oxford University Press.
- Bachman, L.F. (1991) What does language testing have to offer? *TESOL Quarterly*, 25, (4), 671-704.
- Bachman, L.F. (2002). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement*, 21(3) 5-18.
- Bachman, L.F. & Palmer, A.S. (1996). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bailey, A.L. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In E.L. Baker (Principal Investigator), *The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives* (pp. 85-106). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

- Bailey, A.L. and Butler, F.A. (2003). An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document. CSE Report 611. Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Bailey, A.L., Butler F.A., LaFramenta, C. and Ong, C. (2001). Towards the characterization of academic language in upper elementary science classrooms. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Baral, D. (1979). Academic achievement of recent immigrants from Mexico. NABE Journal, 3(3), 1-13.
- Biegel, S. (July, 2003). San Francisco Unified School District Desegregation independent review (Report No. 20). San Francisco, CA: United States District Court for the Northern District of California.
- Butler, F.A. and Castellon-Washington, M. (2000). Students' concurrent performance on tests of English Language Proficiency and Academic Achievement. In E.L. Baker (Principal Investigator), The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives (pp. 51-83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A. and Stevens, R. (1997). Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations (CSE Technical Report 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A. and Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. Language Testing, 18(4), 409-428.
- Butler, Y.G., Orr J., Gutierrez, M. and Hakuta, K. (2000). Inadequate conclusions from an inadequate assessment: What can SAT-9 scores tell us about the impact of Proposition 227 in California? Bilingual Research Journal, 24(1&3), 141-154.

- Carroll, J.B. (1980). Testing communicative competence. Oxford: Pergamon Press.
- Cervantes, R. and Nakano, P. (1979). Oral language tests: The Language Assessment Scales and Bilingual Syntax Measure. Sacramento, CA: California State Department of Education.
- Collier, V. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, 21 (4), 617-641.
- Collier V. (1995). Acquiring a second language for school. *Directions in Language and Education*, 1:4. Washington, D.C.: The National Clearinghouse for Bilingual Education. Retrieved January 17, 2003 from <http://www.ncbe.gwu.edu/ncbepubs/directions/04.htm>
- Collier, V., and Thomas, W.P. (1988). Acquisition of cognitive academic language proficiency: A six-year study. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.
- Coltrane, B. (November 2002). English language learners and high stakes tests: An overview of the issues (EDO-FL-02-07) [Electronic Version]. ERIC Clearinghouse on Languages and Linguistics Digest. Retrieved January 17, 2003 from <http://www.cal.org/ericcll/digest/0207coltrane.html>
- Commins, N. and Miramontes, O. (1989). Perceived and actual linguistic competence: A descriptive study of four low-achieving Hispanic bilingual students. *American Educational Research Journal* 26 (4), 443-472.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49 (2), 222-251.
- Cummins, J. (1980). The entry and exit fallacy in bilingual education. *NABE Journal*, 4, 25-60.
- Cummins, J. (1981a). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, 2(2), 132 -149.
- Cummins, J. (1981b). The role of primary language development in promoting educational success for language minority students. In California State Department of Education, *Schooling and language minority students: A theoretical framework*(pp. 3-49). Los Angeles: California State University, Los Angeles, Evaluation, Dissemination, and Assessment Center.

- Cummins, J., and Nakajima, K. (1987). Age of arrival, length of residence, and interdependence of literacy skills among Japanese immigrant students. In Harley, B., Allen, P., Cummings, J., and Swain, M. The development of bilingual proficiency: Final Report Volume III: Social Context and Age (pp. 183-202). Toronto, Canada: Modern Language Centre, The Ontario Institute for Studies in Education. [ERIC Document Reproduction Service No. ED 291 248]
- DeAvila, E., Cervantes, R., and Duncan, S. (1978). Bilingual program exit criteria. Paper submitted to the Office of Program Evaluation and Research, California State Department of Education.
- Dieterich, T., Freeman, C., Crandall, J. (1979). A linguistic analysis of some English proficiency tests. *TESOL Quarterly*, 13 (4), 535 -550.
- Dolson, D. (1985). Bilingualism and scholastic performance: The literature revisited. *NABE Journal*, 10 (1), 1-28.
- Fern, V. D. (2002, February). The report of the San Francisco Unified School District Bilingual Education Task Force. San Francisco, CA: Multilingual Programs, San Francisco Unified School District.
- Fetter, M. (1983). Reading achievement and English language proficiency in the California Assessment Program Sixth Grade Test. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec. [ERIC Document Reproduction Service No. ED 230-599]
- Gillmore, G., and Dickerson, A. (1979). The relationship between instruments used for identifying children of limited English speaking ability in Texas. Houston: Region IV Service Center. [ERIC Document Reproduction Service No. ED 191-907]
- Gold, N. and Tempes, F. (1987). A state agency partnership with schools to improve bilingual education. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C.
- Graham, C. and Acosta, S. (1979). Comparison of state-adopted oral language tests. Paper presented at the National Association for Bilingual Education, 8th Annual Conference, Seattle, Washington.
- Grissom, J.B. (2004). Reclassification of English learners. *Education Policy Analysis Archives* 12(36), pp. 1-36.

- Hakuta, K., Butler, Y.K., and Witt, D. (2000). How long does it take English learners to attain proficiency? (Policy Report 2000-1). Santa Barbara, California: The University of California Linguistic Minority Research Institute (Policy Report 2000-1). [ERIC Document Reproduction Service No. ED 443 275]
- Hakuta, K., and Snow, C. (1986). The role of research in policy decisions about bilingual education. *NABE News*, 9 (3), 1, 18-21.
- Harley, B., Allen, P., Cummings, J., and Swain, M. (1990). The development of second language proficiency. New York: Cambridge University Press.
- Hayes, Z. A. (1981). Limited language proficiency: a problem in the definition and measurement of bilingualism. Paper presented at the Language Proficiency Assessment Symposium, Airlie House, Virginia. [ERIC Document Reproduction Service No. ED 228 -859].
- Hawkins, M.R. (2004). Researching English language and literacy development in schools. *Educational Researcher* 33(3), 14-25.
- Heubert, J.P. & Hauser, R.M. (Eds.). (1998). High stakes: Testing for tracking, promotion and retention. Washington, D.C.: National Academies Press. Retrieved January 17, 2003 from <http://www.nap.edu/books/0309062802/html/index.html>
- Hymes, D. (1972). On communicative competence. In J. B. Pride and J. Holmes (Eds.), *Sociolinguistics*, Harmondsworth, England: Penguin.
- Johnson, W., & Packer, A. (1987). *Workforce 2000: Work and Workers for the 21st Century*. Indianapolis, IN: Hudson Institute.
- Keating, H. (1984) An assessment of the Carpenteria preschool Spanish immersion program. *Teacher Education Quarterly*, 11(3), 80-94.
- Kirp, D.L. (1976). Race, politics and the courts: school desegregation in San Francisco. *Harvard Education Review*, 46(4), 572-611.
- Kopriva, R. (2000). Ensuring accuracy in testing for LEP Students: A practical guide for assessment development. Washington, D.C.: Council of Chief State School Officers.
- Krashen, S. and Biber, D. (1988). On course: Bilingual education's success in California. Sacramento, CA: California Association for Bilingual Education.

- Kuhlman, N. and Nadeau, A. (1999). English language development standards: The California model. *The CATESOL Journal*, 11(1), 143-160.
- LaCelle-Peterson, M.W. and 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, E. W. (1985). The academic achievement of Chinese-American fluent English proficient and non-minority background intermediate grade students. Unpublished Ed.D. dissertation. Stockton, CA: University of the Pacific.
- McGroarty, M. (1982). English language tests, school language use, and academic achievement in Spanish-speaking high school students. Unpublished Ph.D. dissertation. Palo Alto, CA: Stanford University.
- Miramontes, O.B., Nadeau, A. and Commins, N.L. (1997). *Restructuring schools for linguistic diversity*. New York: Teachers College Press.
- National Clearinghouse for Bilingual Education (October, 1997). High stakes assessment: A research agenda for English language learners symposium summary. Washington, D.C.: The George Washington University. Retrieved January 17, 2003, from <http://www.ncela.gwu.edu/ncbepubs/reports/highstakes/symposium.htm>
- North, Brian (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing.
- Numbers and Needs. (March 1993). Numbers of school-agers with spoken English difficulty increase by 83%. Washington, DC: Numbers and Needs, 3 (2), 2.
- Oller, J. (1979). *Language tests at school: A pragmatic approach*. London: Longman.
- Oller, J. (1980). A language factor deeper than speech: More data and theory for bilingual assessment. In James Alatis (Ed.), *Current issues in bilingual education: Georgetown University Roundtable on Language and Linguistics 1980* (pp. 14-30) Washington, DC: Georgetown University Press.
- Olsen, R. W-B. (1994). LEP enrollment statistics. *TESOL Matters*, 4 (1), 12.
- Page, A. and Ramirez, A. G. (1986). Some effects of bilingual home environments on academic performance and linguistic skills. In E. Garcia and B. Flores (Eds.), *Language and literacy research in bilingual education* (pp 51-66), Tempe AZ: Center

for Bilingual Education, Arizona State University.

Politizer, R., and Ramirez, A. (1981). Linguistic and communicative competence of students in a Spanish-English bilingual high school program. *NABE Journal*, 5 (3), 81-104.

President's Advisory Commission on Educational Excellence for Hispanic Americans. (2000). *Testing Hispanic students in the United States: Technical and policy issues*. Washington, D.C.: U.S. Government Printing Office.

Ramsey, C. and Wright, E. (1974). Age and second language learning. *The Journal of Social Psychology*, 94, 115-121.

Revilla, A.T. and Asato, J. (2002). The implementation of Proposition 227 in California schools: a critical analysis of the effect on teacher beliefs and classroom practices. *Equity and Excellence in Education*, 35(2), 108-18.

Rivera, C. (1999). Policies and practices related to inclusion and accommodation of limited English proficient students in state and district assessment systems. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Rivera, C., and Simich, C. (1989). Issues in the assessment of language proficiency of language minority students. *NABE Journal*, 6 (1), 19-39.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly*, 18 (2), 199-220.

Skutnabb-Kangas, T. (1979). *Language in the process of cultural assimilation and structural incorporation of linguistic minorities*. Rosslyn, VA: National Clearinghouse for Bilingual Education.

Stevens, R.A., Butler, F.A. and Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English language learners (ELLs) (CSE Technical Report 552)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Tregar, B. and Wong, B. F. (1981). The relationship between native and second language reading comprehension and second language oral ability. Paper presented at the Language Proficiency Assessment Symposium, Airlie House, Virginia. [ERIC Document Reproduction Service No. ED 228-857].

Ulibarri, D., Spencer, M., and Rivas, G. (1981). Language proficiency tests and their relationships to school ratings as predictors of academic achievement. *NABE Journal*, 5 (3), 47-80.

US Department of Labor, Secretary's Commission for Achieving Necessary Skills. (1992). *Learning a living: A blueprint for high performance*. Washington, DC: US Department of Labor.

Waggoner, D. (1988). Language minorities in the United States in the 1980's: The evidence from the 1980 census. In S. McKay & S. C. Wong (Eds.), *Language diversity: Problem or resource?*, (pp. 69-108). Boston: Heinle & Heinle.