January 2008

Education Policy Analysis Archives 16/01

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 16 Number 1 January 16, 2008 ISSN 1068–2341

Achievement Testing for English Language Learners, Ready or Not? [1]

Sau-Lim Tsang
ARC Associates

Anne Katz
School for International Training

Jim Stack
San Francisco Unified School District

Citation: Tsang, S.-L., Katz, A., & Stack, J. (2008). Achieving testing for English Language Learners, ready or not?. *Education Policy Analysis Archives, 16*(1). Retrieved [date] from http://epaa.asu.edu/epaa/v16n1/.

Abstract
School reform efforts across the US have focused on creating systems in which all students are expected to achieve to high standards. To ensure that students reach those standards and to document what students know and can do, schools collect assessment information on students' academic achievement. More information is needed, however, to find out when such assessments are appropriate for English learners and can provide meaningful information about what such learners know and can do. We describe and discuss a study that addresses the question of when it is appropriate to administer content area tests in English to English learners.

[1] This study was partially 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. However, any opinions, findings, conclusions or recommendations expressed in this article are those of the authors.

Drawing on the student database of San Francisco Unified School District, we examined the effect of language demands on the SAT/9 mathematics scores of Chinese-speaking and Spanish-speaking students. Our results showed that while the English language demands of the problem solving subscale affect all students, they have a larger effect on English learners' performance, thus rendering the tests inaccurate in measuring English learners' subject matter achievement. Our results also showed that this effect gradually decreases as students become more proficient in English, taking five to six years for students to reach parity with national norms. These results have important implications for the design of school accountability systems and policies with high-stakes consequences for English learners such as high-school graduation requirements based on standardized tests.
Keywords: bilingual students; high risk students; high stakes tests; language proficiency; limited English speaking; standardized tests.

Pruebas de Rendimiento Académico para los Estudiantes en Proceso de Aprendizaje del Idioma Inglés, Listos o No

Resumen

Los esfuerzos de reforma escolar en los Estados Unidos se han enfocado en la creación de sistemas en los cuales se espera que todos los estudiantes alcancen estándares elevados. Para asegurarse que los estudiantes logran dichos estándares y para documentar lo que los estudiantes saben y pueden hacer, las escuelas recogen información sobre la evaluación del logro académico de los estudiantes. Sin embargo, se necesita más información para saber cuándo dichas evaluaciones son apropiadas para los estudiantes en proceso de adquisición del idioma inglés, y que al mismo tiempo puedan proporcionar información significativa sobre lo que estos estudiantes saben y pueden hacer. En este trabajo, describimos y discutimos un estudio que trata sobre el asunto de cuándo es apropiado administrar pruebas de contenido de área en el idioma inglés a estudiantes en proceso de aprendizaje de dicho idioma. Usando el banco de datos de estudiantes del Distrito Escolar Unificado de San Francisco, examinamos el efecto de las exigencias lingüísticas en los resultados de matemáticas de la prueba SAT/9 administrada a alumnos con idioma materno chino y español. Nuestros resultados mostraron que mientras las demandas lingüísticas del inglés en la subescala Resolución de Problemas afectan a todos los estudiantes, su efecto es más pronunciado en el rendimiento de aquellos estudiantes que están en el proceso de aprendizaje del inglés, de tal forma, que estas pruebas se muestran inadecuadas para medir el rendimiento académico por materia de los estudiantes en proceso de aprendizaje del inglés. Nuestros resultados también mostraron que este efecto se reduce gradualmente conforme los estudiantes se vuelven diestros en el idioma inglés, tomando entre cinco y seis años para que los estudiantes alcancen paridad académica de acuerdo con las normas nacionales. Estos resultados tienen implicaciones importantes en el diseño de sistemas escolares de responsabilidad y en las políticas con consecuencias muy importantes para los estudiantes en proceso de aprendizaje del inglés, tales como los requisitos de graduación para la secundaria basados en pruebas estandarizadas.
Palabras clave: estudiantes bilingües; estudiantes en alto riesgo; pruebas de alto impacto; destreza lingüística; dominio limitado del inglés; pruebas estandarizadas.

Large-scale assessment has been used increasingly to guide state and local educational policy. Under the mandates of the federal government's No Child Left Behind (NCLB) Act of 2001, the revision of the Elementary and Secondary Education Act and, arguably, the most broad reaching educational intervention effort of the federal government up to this time, states have expanded the scope and frequency of student testing, revamped their accountability systems, and begun striving to demonstrate annual progress in raising the percentage of students proficient in reading and math. To encourage compliance, NCLB imposes a series of sanctions on schools and school districts that do not meet targeted benchmarks.

This legislation is intended to set high standards for the nation's schools and to compel schools to focus and improve their curriculum and instruction—in other words, to push schools to target teaching on what we expect students to learn (Heubert & Hauser, 1999). However, the use of tests to determine performance levels has spurred much debate. Supporters laud the use of objective measures as a means to raise academic standards, hold schools accountable for their curriculum and instruction, and provide parents with evidence of their children's academic performance. Others argue that basing high-stakes decisions about school performance solely on a limited set of test results often places schools serving immigrant and minority students at a disadvantage (Kim & Sunderman, 2005).

Such tests may also not serve the very populations they are designed to support. One ongoing controversy has been the use of standardized achievement tests written in English to assess the academic performance of English Language Learners (ELLs) (Abedi, Leon, & Mirocha, 2000; La Celle-Peterson & Rivera, 1994). Critics have argued that such tests do not provide an accurate estimate of these students' academic achievement because their limited proficiency in English interferes with their performance on the tests. With limited English proficiency, students may not understand test items or even how to answer the questions. If testing is to serve as an essential tool to inform the improvement of instructional practices and student learning, how can we have confidence in using test results that may not provide accurate information about how well a student is doing at school or in a subject matter?

The issue becomes critical when testing results are used for high-stakes decisions impacting individual students. Thus, for example, California students, beginning in the 2005–06 school year, cannot receive their high school graduation diploma if they do not pass the California High School Exit Examination (CAHSEE). Many ELLs who do not have adequate time to acquire the English language proficiency for passing the CAHSEE will not graduate from high schools. Statistics provided by the California State Department of Education showed that as of June 2006, for the class of 2006, 26% of the 18,565 ELLs had not passed the CAHSEE (California Department of Education, 2006). Similarly, ELLs are also penalized by their performance on standardized tests when applying for colleges and, in many cases, applying for jobs.

In light of the high-stakes and consequential nature of testing, some educators have suggested banning testing altogether for ELLs. However, this practice is short sighted since ELLs need to be included in school accountability schemes, and test results do provide us with information on students' performance, albeit through the filter of their limited English proficiency. Such test results can be used as part of a total portfolio of data to guide the improvement of curriculum and instruction. In addition, most educators agree that ELLs can be tested in English after they have been in US schools and acquired enough English proficiency to take the test. The critical issue is determining when students have acquired enough English proficiency to be tested in English.

To address this issue, a study was conducted from September 2001 until June 2004 to explore the development of ELL student performance on achievement tests in conjunction with developing English language proficiency. Supported in part by a grant from the U.S. Department of Education, the study utilized the extensive student assessment database of the San Francisco Unified School District. Specifically, the study asked: *When is it appropriate to administer standardized content area tests in English to ELLs?*

The question is more complex and difficult to answer than one may initially assume. The first difficulty is with the *when*. Does it refer to time, i.e., the number of years an ELL has been learning English or to an ELL's English proficiency acquisition status? In the latter case, educators have argued that different types of language proficiency have differential relevance to academic achievements. Related to this difficulty are the academic language demands of various content areas. Disciplines such as math, history, and science use specialized vocabulary, sentence structures, and genres in content-specific ways. In a testing situation within a specific area, ELLs may or may not have developed the relevant English proficiency for that content area. In the following section, we review a range of literature to gain more understanding of the issues underlying this research question.

Literature

In this section, we review recent literature and research relevant to the assessment of English language learners. This review seeks to clarify the research question of this study by exploring the complexity of defining English language proficiency, recognizing that this complexity has a direct impact on the creation of valid, equitable content area assessments for English language learners. We begin by exploring models and theories of language proficiency and then examine two strands of research. The first strand of research studies focuses upon the development of academic achievement of ELLs, and the second, upon research on the assessment of English language learners, with an emphasis on assessment in the content areas.

Models of language proficiency, language use and language ability

Over the past decades, notions of language proficiency have evolved from descriptions of listening, speaking, reading, and writing, rooted in the knowledge of systems such as phonology, syntax, and lexicon, to a focus on how language is used to create meaning, particularly within the context of specific settings (Lado, 1961; Canale & Swain, 1980; Chomsky, 1965; Hymes, 1972; Larsen-Freeman, 2003; North, 2000). We focus on three models that have influenced current thinking about school-based language proficiency.

Cummins' (1981b) model of second language acquisition and use provides an early framework for examining the multidimensionality of language used within an educational setting. According to Cummins' model, any communicative task can be described along two continua, according to its cognitive demand and to its contextual support. For example, some communication can be portrayed as 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 such as casual greetings and leave-takings. Other communication tasks can be cognitively demanding yet with minimal contextual support, accompanied only by linguistic cues. These descriptors characterize the language of schooling and literacy tasks. In Cummins' model, command of both basic communication skills and cognitive academic skills is necessary for language

proficiency and thus for language learners to be successful in school. Critics of Cummins' model argue that while his distinction between social and academic language is a useful one, his definitions ignore the way in which situations of language use can determine the level of cognitive complexity (Bailey, 2006). His distinctions also do not acknowledge the complexities of academic language or its development (Scarcella, 2003), and according to some, they promote a deficit theory of language use (Wiley, 1996). Nevertheless, his model has helped educators realize that the language proficiency needed for success in school is multidimensional and must encompass more than oral fluency and social uses of language.

Collier's (1995) model adds a new element to the discussion of language proficiency—sociocultural processes—and configures linguistic, academic and cognitive development as separate though inter-related factors in a multifaceted and complex model of the language acquisition process. 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 immigration status, limited economic resources, and societal attitudes towards immigrant groups, for example, cultural stereotyping and the subordinate status of a minority group. Like Cummins' model, Collier's argument helps policy makers and practitioners develop a clearer understanding of the complexities of language learning in schools and, thus, become more aware of the challenges inherent in language development as well as the time that it takes to become proficient.

In line with these complex and dynamic views of language, Bachman (1990, 2002) and Bachman and Palmer (1996) describe language ability within an interactional framework. Shifting away from the term "language proficiency," Bachman (1990, 2002) creates models of "language use" and "language ability" and applies them to language testing. Acknowledging the influence of a sociocultural context on the performance of the language learner, he notes that testing methods and the background characteristics of language learners influence scores as much as the students' language skills. This model also distinguishes between linguistic and academic development. In describing language use, 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 emphasize 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.

These three models are useful heuristics for developing a deeper understanding of language proficiency and the complexities inherent in developing proficiency. They also illustrate the divide between theoretical models of language use, proficiency and ability and the assessments designed for English language learners in school settings (Hakuta, Butler & Witt, 2000; North, 2000). While these models provide us with a multi-dimensional notion of language proficiency, they hardly address the intersection of language and content. In the next two sections, we examine some of the research that has contributed to the development of these complex models of language proficiency and to our understanding of language acquisition in school settings.

Research on Developing Academic Achievement of ELLs

Some researchers have looked at the development of language proficiency within an academic context by exploring how long it takes students schooled only in the second language (English) to reach the average academic achievement level of native speakers. Cummins (1981a)

examined the time needed for immigrants to Canada to acquire English academic language proficiency when taught in that second language after arrival. He 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 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 might not be generalizable outside of the Canadian social context.

In the same vein, 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 an 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 very little or no previous exposure to English. Collier found that to approach the 50 NCE in reading, language, science, and social studies, the students needed 4 to 8 years. In this study, age on arrival had an effect in that students arriving between the ages of 8 and 11 were the fastest achievers. The cross sectional data on advantaged ELL 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.

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, Allen, Cummings, & Swain, 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 arrived at the age 6–7 to make more rapid progress toward grade level norms than those who arrived at older ages. They also found that when length of residence is controlled, there was a significant relationship in reading achievement between home language and English. 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.

While these studies focused on the number of years it took for students to approximate the achievement levels of English-only students, they did not answer the question of whether or when academic assessment data of ELLs are accurate measures of content achievement.

Research on the impact of academic language on the assessment of English Language Learners

The studies in this area have emphasized the importance of understanding the effect of the language demands inherent in academic tasks and specifically standardized test items on the performance of ELLs. Bailey (2000, 2006) describes language demands in terms of potential language difficulties faced by ELLs at the lexical, syntactic, and discourse levels during schooling

tasks, both in the classroom and when taking standardized tests. In analyzing a test item, for example, she identifies sources of difficulty for ELLs in the use of uncommon meanings of words and complex sentence constructions and also the need to make meaningful connections between new and old information within a stretch of discourse. In their analysis of test data comparing ELL and native speakers of English, Abedi, Leon, and Mirochi (2000) found that the achievement gap between ELLs and non-ELLs increased as the language demands of the assessment tools increased.

Other studies have identified a mismatch between the model of English language proficiency underlying standardized English language proficiency tests commonly used in school districts and the academic language required for successful performance on content tests. In a study designed to describe and compare the language and performance of 7th grade ELLs on tests of language proficiency and achievement (Stevens, Butler, & Castellon-Wellington, 2000), text analyses revealed limited correspondence between the two tests, suggesting that "competent performance" on a commonly-used language proficiency test such as the Language Assessment Scales (LAS) may not provide sufficient evidence for determining whether or not ELLs can handle the academic language demands of content assessments. Given this lack of congruence, 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. By creating such a framework, Bailey and Butler reason that we will be able to identify a "threshold level of proficiency," which up until now has been elusive because individual school districts and states have used such varying requirements for identifying English proficiency (p. 33).

This mismatch becomes acutely relevant in high-stakes arenas such as high school graduation requirements. Fillmore and Snow (2000) examined prototype test items for a high school graduation examination for one of the 23 states that has adopted this requirement. Their analysis reveals that the language used in the high school graduation tests is similar to that used in school textbooks and academic discussions about science, mathematics, literature, or social studies. Thus, students must have competence in academic English to do well on the test. Additional studies have begun to identify and describe academic language in the content areas (Bailey, Butler, LaFramenta & Ong, 2001; Schleppegrell, Achugar, & Oteiza, 2004), 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. While these studies have highlighted the issues around academic language in content areas particularly in light of generalized notions of English language proficiency, this study undertook the task of determining what effects this mismatch may have within actual student performances on tests over time.

Study Context

San Francisco Unified School District

We began our study in the spring of 2002, seeking to answer the following question: *When is it appropriate to administer standardized content area tests in English to ELLs?* As the literature review in the previous section shows, the research question is a complex one. To answer this question, we had the entire student database of the San Francisco Unified School District (SFUSD) at our disposal. SFUSD is a microcosm of increasing student diversity across the U.S. 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 primarily from Asia, Central America, and South America as

well as elsewhere. One of the most densely populated cities in the United States, its approximately 770,700 residents in 2000 lived within its 46.4 square-mile area (U.S. Census, 2000).

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 ELLs. These ELLs 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 identified (or re-designated) as Fluent English Proficiency (FEP) students with ongoing language and academic content area needs.

Many ELLs live below the poverty line in a city with a high cost of living aggravated by an influx of wealthy Silicon Valley/e-commerce professionals. In 1999, ELLs comprised 39% of students in the district's Title I program. Within this group, both Latino and Chinese students were overrepresented 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.

To understand the context for English language learners in San Francisco, one must understand the history of civil rights for language learners in San Francisco's public schools. In 1974, SFUSD was the setting for the Supreme Court Decision, *Lau v. Nichols*, which stated that SFUSD had "denie[d] [the Chinese-speaking minority students] a meaningful opportunity to participate in the educational program" (*Lau v. Nichols*, 1974, p. 567). In conjunction with the Equal Educational Opportunities Act of 1974, the decision 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 eliminated the requirements for bilingual programs in 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 ELLs 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, p. 1)

ELLs in SFUSD

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), when a student is from an English speaking background; Initial Fluent English Proficient (IFEP), when a student is from a non-English background but is proficient in English; and Limited English Proficient (LEP), when the student is from a non-English speaking background and is not proficient in English.

As an 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 into a fourth category as Re-designated Fluent English Proficient (RFEP). Thus, all students in SFUSD fall within four language

proficiency categories. For the purpose of this study, we defined ELLs to include both LEP and RFEP students to ensure that the full range of developing language proficiency was captured within our data sets.

In the 2000–01 school year, there were 18,624 ELLs enrolled in SFUSD. Table 1 shows the distribution of the ELLs according to the major language groups and grade levels (K–11). The data also show that Chinese-speaking students accounted for 44% and Spanish-speaking students accounted for 38% of the total ELLs in the district.[2] Given this distribution, we chose to focus on the largest language groups to maximize the meaning of within-subgroup analysis.

Table 1
Distribution of ELLs by language groups and grades (K–11), Spring 2001

	Grade level												
Language	K	1	2	3	4	5	6	7	8	9	10	11	Total
Cambodian	11	9	18	13	17	17	13	15	19	22	12	9	1755
Chinese	844	1120	1014	1019	803	650	446	413	363	437	435	502	8046
Filipino	50	58	69	78	79	71	61	92	77	71	74	68	848
Japanese	24	27	16	16	11	4	4		6	2	8	5	123
Korean	12	17	21	19	16	16	8	6	8	10	8	8	149
Spanish	725	750	799	801	729	642	471	432	389	455	366	351	6910
Vietnamese	59	65	60	51	49	37	40	34	32	27	32	45	531
Other	103	115	146	120	134	120	92	97	97	110	98	105	1337

Data

Achievement tests

This article focuses on data from the Stanford Achievement Test, Ninth Edition (SAT/9).[3] At the time of the study, SAT/9 was the California state-mandated achievement test administered to all students for grades 2 to 11. The test was required beginning in 1999. However, SFUSD filed a lawsuit against the State claiming the testing mandate was unfair to LEP students. 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 in the district for the first year. Testing for all students started in 2001. Thus, beginning in the 2000–01 school year SAT/9 results

[2] Chinese-speaking students represent a range of Chinese dialects. The largest dialect group in SFUSD is Cantonese, which constitutes 83% of the K–12 and 87% of the K–5 Chinese-speaking students who are the subjects of the analysis in our study. Another 8% of students are from different areas of Canton province with dialects that are mutually intelligible with Cantonese. In addition, staff of SFUSD suggested that because of San Francisco's overwhelming Cantonese environment, almost all non-Cantonese background students speak Cantonese as a second dialect. As discussed later in the paper, our analysis will only include students who entered SFUSD at the kindergarten level. None of these students had formal education in any Chinese languages/dialects. Thus, we included all Chinese-speaking students as one group in our study sample.

[3] In addition to SAT/9, SFUSD collected data from select students on the following tests: California English Language Development Test (CELDT), California Standards Test (CST), High School Exit Exam, Language and Literacy Assessment Rubric, Integrated Writing Assessment, and SABE 2.

were the only set of data that was collected on students of all language proficiency categories. We used the results from the tests administered in the last two weeks of April in 2001. The SAT/9 scale scores that were available are shown in Table 2.

Table 2
Available SAT/9 data

		Grade level cluster	
		2nd to 8th	9th to 11th
Reading	Vocabulary	√	√
	Comprehension	√	√
Language		√	√
Mathematics	Procedures	√	√
	Problem Solving	√	√
Science			√
Social Sciences			√

In addition, background information for all students is collected at their first enrollment in the school district and maintained and updated yearly in a centralized database, including birth date, birth place, year of entry into the U.S., parents' education background, home language, family income indicators, English language proficiency, GPA, and additional information.

Limits of data and exclusions

The data were constrained in five ways. First, item statistics were not available from California Department of Education or the testing contractor; we were unable to calculate the reliability of the test and subtest for our sample and sub-samples. In addition, SAT/9 was administered to students in grades 2 through 11. Therefore, our analyses were limited to those grade levels. Third, we did not use the science and social science data for high school students since scores in these content areas were not available for elementary and middle school students. Fourth, our preliminary analysis of SAT/9 reading and math data showed that a significant number of students scored at the first percentile (Table 3) 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 in the testing, undermining the validity of the scores. Thus, we excluded these students from our analyses.[4]

[4] Table 3 shows the number of students who scored at the1st 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.

Table 3
Number of LEP students scoring at the first percentile by grade, 2001

	Grade level										
Language	2	3	4	5	6	7	8	9	10	11	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

Finally, state policy at the time of the study allowed test accommodations for first-year limited English proficient students; students' teachers made the determination as to whether accommodations were appropriate. Accommodations included extra or extended administration time, the reading of test items or questions, translation of test directions, the use of bilingual dictionaries. Table 4 shows 51% of 1st year ELL students were provided accommodations during this testing.

Table 4
Number of first-year ELL students provided accommodations by language group, 2001

N	Chinese	Spanish	Other ELLs	All 1st-year ELLs
All tested	415	318	253	986
With accommodations	229	191	82	502
% with accommodations	55%	60%	32%	51%

Since accommodated test scores were not acceptable for the purpose of our study and the majority of the first-year LEP students were provided accommodations, the remaining LEP students would have composed a biased sample. Therefore, we excluded all first year LEP students in our study. After the exclusions, our study sample consisted of 9925 Chinese- and 4890 Spanish-speaking students in grades 2 through 11. Table 5 shows the distribution of the English-language learning students in the sample.

Table 5
Distribution of sample by language group, language status, and grade level

Language	Grade level										
status	2	3	4	5	6	7	8	9	10	11	Total
Chinese											
LEP	951	849	631	447	295	304	270	293	276	293	4609
RFEP	1	155	331	574	769	818	678	680	708	602	5316
Total	956	1004	962	1021	1064	1122	948	973	984	895	9925
Spanish											
LEP	592	624	543	431	369	322	265	245	202	147	3740
RFEP	0	17	58	180	130	145	174	160	159	127	1150
Total	592	641	601	611	499	467	439	405	361	274	4890

Analysis

At the descriptive level, as Tables 6 and 7 show in mean NCE SAT/9 scores in reading and math, Chinese-speaking students scored higher than the Spanish-speaking students, especially in mathematics, where the Chinese-speaking students achieved well above the national norm. The differences in achievement profiles of the two groups are consistent with other findings. The data also show consistently larger standard deviations for the Chinese students' math scores at all grade levels 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). The difference in achievement of the two language groups provides sample variation with some generalizing consequences.

Table 6
SAT/9 reading NCE scores (standard deviations in parentheses), 2001

Language Status	Grade level 2	3	4	5	6	7	8	9	10	11
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 7
SAT/9 math NCE scores and standard deviations (in parentheses), 2001

Language Status	Grade level 2	3	4	5	6	7	8	9	10	11
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)

Table 8 shows correlations between SAT/9 reading and math scores. In line with previous research (Bailey, 2000) that suggests that reading items represent higher degrees of language difficulty than math items, we hypothesized that the correlations for the ELLs would be lower than the national norming sample because math content and test items have fewer English language demands. If that were true, ELLs' performance on the test would be less dependent on their English reading ability and test scores. Table 8 confirms 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 correlations decrease from grade 2 to grade 11 while the correlations for the Spanish students did not show a consistent trend by grade.

Table 8
Reading x math correlations, 2001

Language and status	Grade level 2	3	4	5	6	7	8	9	10	11
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 sample	.73	.78	.78	.77	.81	.76	.75	.69	.65	.70

With some evidence for this hypothesis, we examined the relationship of the reading scores with the two sub-scales of the math test: procedures and problem solving. The procedures sub-scales consist of items requiring fewer reading comprehension abilities/skills while the problem solving sub-scale includes word problems requiring more reading comprehension abilities/skills. We hypothesized that the correlation between reading and math/procedure would be lower than the correlation between reading and math/problem solving, as the test items in the procedure subscale have less (or no) reading in them compared to the test items in the problem solving subscale.

Table 9
Reading x math sub-scale correlations, 2001[5]

	Grade level						
Language and scales	2	3	4	5	6	7	8
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 norming sample							
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 of Table 9 confirmed our hypothesis. The correlations between reading and math/problem solving are consistently higher than those between reading and math/procedure. An examination of the correlations for the national norming 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 in our San Francisco sample than for students in the national sample. That is, the demands of reading abilities/skills have a greater effect on the Chinese and Spanish students' performance on the problem solving items.

To answer the *when* in our research question, we looked at the length of time an ELL was enrolled in the school district. Assuming that ELLs are learning English in school, the number of years in school can also be viewed as a proxy for the length of time an ELL has been learning English. We identified the cohort of ELLs who entered the district in Kindergarten and had been continuously enrolled in SFUSD schools. Table 10 shows 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 used in our analyses. An examination of Table 10 shows that the majority (1,568) of the ELLs at 2nd grade had been in the school district for three years. These are the students who entered SFUSD in kindergarten. Similarly, the majority (1,627 and 1,474) of LEP students in 3rd and 4th grades had been in the district for four and five years respectively. Across Table 10, cells with the largest numbers of students from grades 2 to 5 were selected for our analyses since they were large enough to track year to year retroactively.

[5] SAT/9 does not provide procedures and problem solving subscales for 9th, 10th, and 11th grades.

Table 10
Number of ELLs by grade levels by years by language group in SFUSD, 2001

	Grade level									
Years ELL and language group	2	3	4	5	6	7	8	9	10	11
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
$\geq$ 8 Years										
Chinese	1	0	2	4	33	**957**	736	738	793	712
Spanish	4	3	8	4	52	**398**	392	368	376	315

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 grades 2 through 5.[6] Further, we examined the difference of the correlations between reading x math/procedure and reading x math/problem solving. This difference allows us to operationalize hypothesized language demands of the tests. Table 11 shows that the differences decrease as the students' years in SFUSD increase (as they move up in their grade levels). The decreases were consistent for both the Chinese and Spanish ELLs.

To identify the pattern across student populations, we plotted the differences in correlations, comparing those of the Chinese and Spanish-speaking students with the national norming sample. Figure 1 shows that the intercorrelational differences for all three groups converge as they spend more time in school. The Chinese ELLs in San Francisco converge with the national sample at fourth grade (modally their fifth year in SFUSD) while the Spanish ELLs converge at fifth grade (modally their sixth year).

Table 11
2001 ELLs' reading x math/procedure correlation and math/problem solving correlation by years in SFUSD vs. national norming sample

Language group and test combination	Length in ELL			
	3 years	4 years	5 years	6 years
Chinese ELLs				
(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 ELLs				
(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 norming sample				
(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

[6] We have also calculated and plotted differences in correlations for 6th graders. The results are similar to those of the 5th graders. However, there are significant dropouts (or transfers) among the ELLs from the district starting at the middle grade levels. We did not include the calculations from those grade levels because we do not know if the middle school population is comparable to the elementary school population.

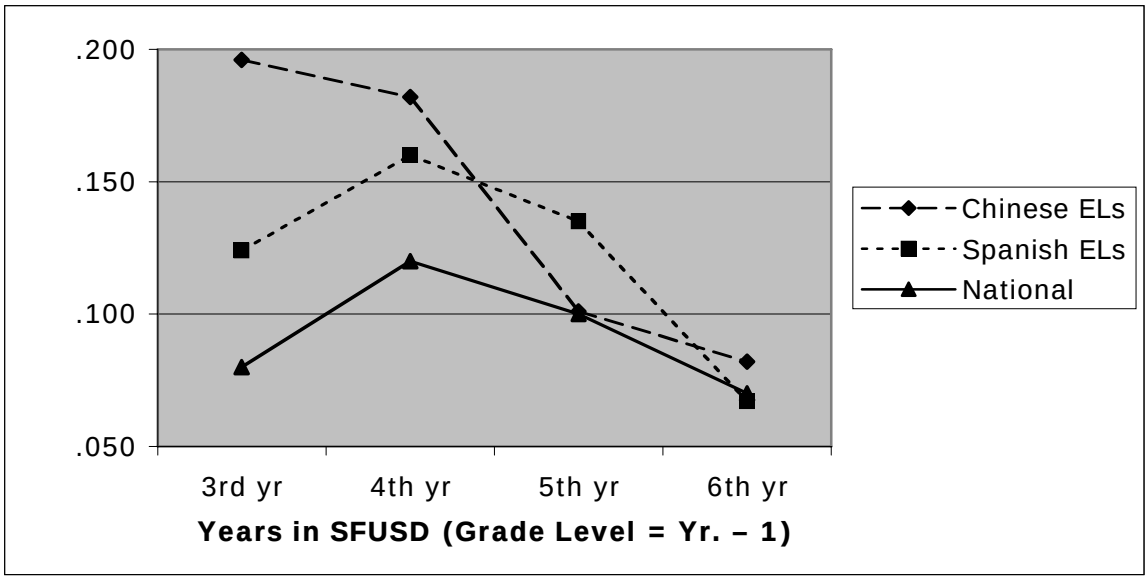

Figure 1. Comparison of 2001 SFUSD ELLs with the national norming sample (reading x math/problem solving) - (reading x math/procedure). See Table 11 for calculations.

These results suggest that when Chinese ELLs are in their fifth year of acquiring English proficiency, the language demand of reading comprehension abilities/skills for word problems is the same for them as for students in the national sample. Similarly, when Spanish ELLs are in their sixth year of acquiring English proficiency, the language demand of reading comprehension abilities/skills for word problems is the same for them as for students in the national sample. To confirm our findings, we replicated the analysis using the data from the spring 2002 test data when they became available.

Table 12
2002 ELLs" reading x math/procedure correlation and math/problem solving correlation by years in SFUSD vs. national sample

Language group and test combination	Length in ELL 3 years	4 years	5 years	6 years
Chinese ELLs				
(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 ELLs				
(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 norming sample				
(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

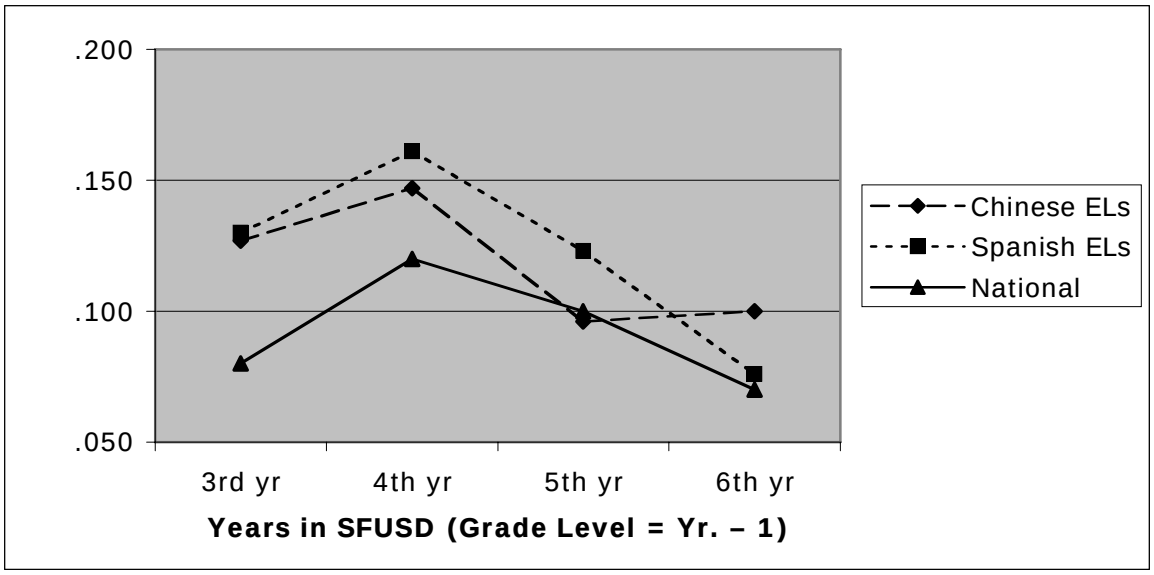

Figure 2. Comparison of 2002 SFUSD ELLs with the national norming sample (reading x math/problem solving) — (reading x math/procedure). See Table 11 for calculations.

We again selected the ELLs 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. Table 12 and Figure 2 show the results. The table and figure showed results similar to the 2001 data. The Chinese ELLs' differences in correlations converge with the national sample at fourth grade (fifth year in school) while the Spanish ELLs 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 ELLs' difference diverges from the national sample after converging at 4th grade. Further analysis is necessary to understand this difference.

Discussion

This article suggests a way of capturing the language demands on testing and the difficulties faced by English language learners. One could view the difference in correlations between reading and math/procedures scores, on the one hand, and reading and math/problem solving scores, on the other as a measure of the language demands of a test. We tentatively label this difference as a Language Demand Index (LDI). One should keep in mind that language demands on testing affect all students, not just ELLs. Our results showed that the Language Demand Index lies between .07 and .12 for the national sample of 2nd to 5th grade students. This positive range suggests that the English language demands of the word problems in the SAT/9 problem solving subscale affect fluent English speakers, too. However, the English language demands of word problems have a larger effect on ELLs' performance on the SAT/9 problem solving subscale at earlier grades. This result confirms what many educators have argued, that the English language demands of standardized tests render the tests inaccurate in measuring ELLs' achievement of subject matter.

Our results also showed that the effects of the English language demands of the problem solving subscale on ELLs gradually decrease as ELLs accumulate more years of schooling and move up the grades. If we assume that ELLs are acquiring English language proficiency in school and their years of schooling can serve as a proxy for their English proficiency, our results imply that the

effects of the English language demands of the word problems decrease as students became more English proficient. The effects of the English language demands on ELLs reduce gradually and eventually become the same as for the students in the national norm sample. This trend is true for both the Chinese-speaking and Spanish-speaking ELLs even though they have very different achievement profiles on the SAT/9 (Tables 5 and 6), with both groups performing significantly different from each other and from the national sample.

Nevertheless, the Chinese-speaking students took less time to reach parity with the national norm sample in reducing the effect of English language demands of the SAT/9 math/problem solving subscale. They reached parity with the national norm sample at 4th grade or after five years of schooling in SFUSD. As a population, the Spanish-speaking students required one more year in reducing the effect of English language demands of the SAT/9 math/Problem solving subscale and reached parity with the national norm at 5th grade or after six years of schooling in SFUSD

The different patterns of correlations on the two SAT/9 math subscales suggest that while both tapped students' math abilities, 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. The data from our study suggest that it took 5 to 6 years of instruction for ELLs to overcome the language demands of mathematics word problems in standardized achievement test. This result is true for both the Chinese and Spanish-speaking groups which have very different achievement profiles with the former achieving significantly above the national norm and the latter significantly lower than the national norm in mathematics.

It is important to note that our results showed that the Chinese-speaking ELLs were affected by the language interference of the word problems even when they were achieving significantly above the national norm in mathematics. However, our results also show that the Chinese-speaking ELLs took one less year than the Spanish-speaking students to reach parity with students in the national norm. We do not know if this is related to the different achievement profiles of the two groups. Does the higher achievement of the Chinese-speaking ELLs in mathematics allow them to overcome the language demands of the word problems in a shorter time? Or is the difference a function of the different education programs for Chinese- and Spanish-speaking ELLs enrolled in SFUSD? In interpreting these findings, the research setting, SFUSD, should be recognized as unique in state and national contexts for being a longtime proponent of English as a Second Language programs and bilingual education. Yet another possibility may be that Chinese-speaking ELLs attend more closely to the formalistic nature of standardized tests than Spanish-speaking ELLs, thus illustrating Bachman's (1990, 2002) contention of an interaction between testing method and language learner background.

Our findings also have two important limitations. First, since the findings are based only on mathematics achievement tests, we do not know the effect of language demands of achievement tests in other content areas. The academic language of many content areas is different than that of mathematics. Some might argue that the academic language of other content areas is more difficult, and ELLs might need additional time to overcome the language demands of these subject areas. Second, our study is based on data of students who entered kindergarten in SFUSD and started acquiring English proficiency at an early age. We do not know if the findings are the same for ELLs who enter US schools at later ages. Additional analyses of students entering at other grade levels are also needed.

This study's findings support the body of research which has shown that English learners need five to seven years before they can attain the academic literacy necessary to negotiate in mainstream classrooms, but 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 with

different content area tests. In addition, the achievement levels of the population of English learners that does not reclassify quickly needs to be traced by grade level.

The findings of our study also suggest looking more closely at 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 importance of examining interaction strategies of English learners as well as the impact of their socio-cultural backgrounds on their ability to learn and engage in the various discourse communities of the classroom and school, whether it is at the elementary, middle or high school level, since these participant structures can provide an avenue to academic engagement in content areas. In addition, Bailey and Butler's (2003) 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.

Schools need to be held accountable to ensure ELLs are receiving appropriate services. The results of this study support recommendations for creating more flexible approaches in accountability systems to determine the achievement of ELLs (e.g., Butler & Stevens, 2001; Gottlieb, 2003; Kim & Sunderman, 2005; McKay, 2005). To comply with reporting requirements of NCLB, policies need to be adapted to permit accountability systems to use multiple indicators and both state-wide and local assessments keyed to the same set of academic standards. For example, Gottlieb (2003) advocates the use of multiple indicators, suggesting that schools use teacher-based assessment as part of a system of large-scale testing, particularly for ELLs beginning to learn English. Such a school-based approach would require setting up standard testing conditions to ensure equity and rigor such as making standard prompts available to teachers and collecting content-related language samples on a regular basis.

McKay (2005) echoes Gottlieb's focus on teacher-based assessment in her recommendations for assessing elementary-level English learners. In articulating the advantages for this approach, she notes that a teacher assessor is familiar with the range of abilities students have demonstrated over time, and both teacher and students are provided with immediate feedback, thus providing opportunities for looping the assessment information back into instruction and program design in a more timely way. As with Gottlieb's proposal, such an approach would need to adopt assessment procedures ensuring the production of "trustworthy data" (p. 349) for use at various schooling levels.

This study also supports calls for policy changes in the formulation of annual yearly progress targets for schools under NCLB (Forum on Educational Accountability, 2007). Today, states establish targets indicating the percent of students that should reach academic proficiency each year. These "percent-proficient targets" increase yearly, and schools and district are judged as successful if the percent of students performing at or above the targets is equal to or greater than the target that year. Schools and districts with large populations of ELLs are being labeled as unsuccessful because many students in the process of acquiring English are not able to meet the percent- proficient targets. NCLB creates a structure that judges schools not to be meeting Adequate Yearly Progress even when the schools are offering viable, comprehensive programs in which students are making substantial gains from year to year toward the target. Our finding is that it takes time for ELLs to acquire the English necessary to perform meaningful on the standardized achievement tests, and this finding supports accountability models that include growth measures so the schools get credit for the progress ELLs and others make over time.

In addition to changing the approach to assessment for accountability, the results suggest that policies with high-stakes consequences for ELLs such as high-school graduation requirements based on standardized tests alone should be re-examined. Alternative and multiple measures that

take into account students' level of English language proficiency may be more appropriate for determining whether or not ELLs are meeting expected high levels of achievement in content areas.

Finally, our difficulty in answering our original research question illustrates the complexity of understanding the academic achievements of ELLs. In light of this complexity, we need to consider better ways to convey important but abstruse educational findings to policy and curriculum makers, the media and even the general public. Understanding how to communicate with major stakeholders about the body of knowledge that exists in the educational research community with regard to how ELLs acquire academic literacy and how best to use standardized tests to assess the achievement of ELLs is a critical step in ensuring that ELLs have an opportunity to participate equitably in our educational systems.

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 (Ed.), *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).

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

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 (Ed.), *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. (Ed.). (2006). *The language demands of school*. New Haven, CT: Yale University Press.

Bailey, A. L., & 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., & 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).

Butler, F. & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K–12: Current trends and new dilemmas. *Language Testing, 19*(4), 409–427.

California Department of Education (2006). *Schools chief Jack O'connell announces more student success on California high school exit exam* [news release]. Sacramento: Author. Retrieved August 5, 2006, from http://www.cde.ca.gov/nr/ne/yr06/yr06rel61.asp.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaching to second language teaching and testing. *Applied Linguistics, 1,* 1–47.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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.: National Clearinghouse for Bilingual Education. Retrieved January 17, 2003 from http://www.ncbe.gwu.edu/ ncbepubs/directions/04.htm.

Collier, V., & 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.

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., & 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. (Eds.), *The development of bilingual proficiency: Final report volume III: Social context and age* (pp. 183–202). Toronto, Canada: Modern Language Centre, Ontario Institute for Studies in Education. (ERIC Document Reproduction Service No. ED291248.)

Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice*. Cambridge: Cambridge University Press.

Fillmore, L. W., & Snow, C. E. (2000). *What teachers need to know about language*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction No. ED447722.)

Forum on Educational Accountability, Expert Panel on Assessment. (2007). *Assessment and accountability for improving schools and learning: Principles and recommendations for federal law and state and local assessments*. Cambridge, MA: Author. Retrieved June 29, 2007, from http://www.edaccountability.org/AssessmentFullReportJUNE07.pdf.

Gottlieb, M. (2003). *Large-scale assessment of English Language Learners. Addressing educational accountability in K–12 settings*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Hakuta, K., Butler, Y. K., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Report 2000–1). Santa Barbara, CA: University of California Linguistic Minority Research Institute (Policy Report 2000–1). (ERIC Document Reproduction No. ED443275.)

Harley, B., Allen, P., Cummings, J., & Swain, M. (1990). *The development of second language proficiency*. New York: Cambridge University Press.

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.). (1999). *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. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35–71). New York: Holt, Reinhart & Winston.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher 34*(8), 3–13.

La Celle-Peterson, M. W., & 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.

Lado, R. (1961). *Language testing.* New York: McGraw-Hill.

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring.* Boston, MA: Thomson Heinle.

Lau v. Nichols, 414 U.S. 563 (1974).

McKay, P. (2006). *Assessing young language learners.* Cambridge: Cambridge University Press.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing.

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

San Francisco Unified School District. (2002). *The report of the San Francisco Unified School District Bilingual Education Task Force*. San Francisco: Author. Retrieved January 14, 2008, from http://portal.sfusd.edu/data/language_academy/draftH.pdf.

Scarcella, R. (2003). *Academic English: A conceptual framework*. Technical Report 2003–1. Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Schleppegrell, M. J., Achugar, M., & Oteiza, T. (2004). The grammar of history: Enhancing content-based instruction through a functional focus on language. *TESOL Quarterly 38*(1), 67–93.

Stevens, R. A., Butler, F. A., & 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).

Tsang, S. L. (1993). Asian American education and the national education goals. In *The National education goal 3: The issues of language and culture*. Proceedings of the Symposium convened by Center for Applied Linguistics. Washington, DC: Center for Applied Linguistics.

Wiley, T. G. (1996). *Literacy and language diversity in the United States.* Washington, DC: Center for Applied Linguistics and Delta Systems.

About the Author

Sau-Lim Tsang
ARC Associates

Anne Katz
School for International Training

Jim Stack
San Francisco Unified School District

Email: stsang@arcassociates.org

Sau-Lim Tsang is the executive director of ARC Associates, a non-profit organization that focuses on improving the education of diverse student groups. He is also the executive director of Oakland Unity High School, a charter school in Oakland, California.

Anne Katz has worked for over twenty years as a researcher and evaluator with educational projects involving linguistically and culturally diverse students. As a lecturer at the School for International Training in Brattleboro, Vermont, she teaches courses in curriculum, assessment, and evaluation.

Jim Stack is the former Director of Achievement Assessments for the San Francisco Unified School District. Dr. Stack was the 2003 President of the California Educational Research Association and is currently serving a three-year term (2006–2009) on the TESOL Board of Directors.

EDUCATION POLICY ANALYSIS ARCHIVES **http://epaa.asu.edu**

EDUCATION POLICY ANALYSIS ARCHIVES **http://epaa.asu.edu**