Academic Language and Content Assessment: Measuring the Progress of English Language Learners (ELLs)

CSE Technical Report 552

Robin A. Stevens, Frances A. Butler,
and Martha Castellon-Wellington
CRESST / University of California, Los Angeles

December 2000

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

Project 1.2. Assessment in Action: Making a Difference in Practice–Validity for Language Minority Students. Frances A. Butler, Project Director, CRESST/UCLA.

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ACKNOWLEDGMENTS

The research reported here is the result of the cooperative efforts of educators and colleagues at UCLA and elsewhere who recognize the importance of a better understanding of the performance of English language learners in our schools. Without the support of all involved, this work would not have been possible. In particular, the project is indebted to the Burbank Unified School District and the Riverside Publishing Company for participating as partners in this research.

Caroline Brumm, Coordinator of Student and Program Evaluation in the Burbank Unified School District, was an enthusiastic supporter. She assisted in obtaining district approval for the study and served as the Center's initial contact with the schools. Principals, assistant principals, teachers, and students cheerfully accommodated the many classroom visits that were necessary for data collection.

Maureen Grazioli, Vice President of Marketing at Riverside Publishing, was instrumental in obtaining permission to use Forms K and L of the Iowa Tests of Basic Skills Social Studies Test for the 7th Grade and served as liaison between Riverside Publishing and CRESST. Riverside Publishing donated the test materials used in this study.

At UCLA, Judy Miyoshi, Paula Hidalgo, and Rosa Valdes assisted with data collection. Xiaoxia Ai provided assistance with the codebook and preliminary analyses. Mike Rinaldi and Motoko Ueyama provided assistance with data input. Nathan Carr and Seth Leon conducted the data analyses. Judy Miyoshi provided administrative support throughout the project, and Steven Acosta provided assistance with the preparation of this report.

Alison Bailey, Richard Durán, Joan Herman, and Ani Moughamian read earlier drafts of the report. Their comments and suggestions were invaluable.

A sincere thank you for the contributions of all who participated in helping to make this report possible.

ACADEMIC LANGUAGE AND CONTENT ASSESSMENT: MEASURING THE PROGRESS OF ENGLISH LANGUAGE LEARNERS (ELLs)[1]

Robin A. Stevens, Frances A. Butler, and Martha Castellon-Wellington
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/University of California, Los Angeles

Abstract

As the nation moves toward inclusion of *all* students in large-scale assessments for purposes of accountability, there is an urgent need to determine when English language learners (ELLs) are able to express what they know on a standardized content test in English. At stake is the validity and reliability of the scores for ELLs and the resulting educational decisions made on the basis of these scores. Because tests are used increasingly to make high-stakes educational decisions, a means for including ELLs in a fair and equitable way is needed. One approach to assuring validity of test scores is to determine at what point ELLs are fluent enough to express what they know on a content test in English. This study investigates the relationships between the language and performance of seventh-grade ELLs on two tests—a language proficiency test and a standardized achievement test. The language of the two tests is analyzed and compared, followed by analyses of concurrent performance on the same two tests. Language analyses indicate that the correspondence between the language of the two tests is limited. Data analyses of student performance show that, although there is a statistically significant relationship between performance on the two tests, the correlations are modest. An additional finding indicates that there are statistically significant within-group performance differences, providing evidence that ELLs are not a homogenous group. Furthermore, item-response analyses provide evidence that, for the highest performing group of ELLs in the study, language may be less of an issue than the content itself. Recommendations include the development of an academic language measure for the purpose of establishing test-taker readiness, further research that investigates the interaction between language and performance as well as within-group performance differences, and the impact of opportunity to learn on content test performance.

[1] The research presented in this report was conducted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to investigate ways in which English language learners (ELLs) can be included in large-scale testing with the goal of fair and equitable assessment for all students. This is the third in a series of reports meant to contribute to a better understanding of ELL performance on large-scale assessments. The conceptual groundwork for the study was laid by the first report (Butler & Stevens, 1997). The second report looked specifically at test-taker preference in the choice of accommodations on a standardized achievement test (Castellon-Wellington, 1999). This report looks at another dimension of the broad issue by comparing the language of two assessments—a language proficiency test and the standardized achievement test—and presenting analyses of concurrent student performance on the two tests.

Within the context of providing fair and equitable access to the curriculum and to assessment for all students in the United States, attention to students who do not speak English as their first language continues to grow. Schools across the country offer a range of curriculum interventions—bilingual classes, sheltered programs, and ESL pullout instruction—tailored to students' English language needs. These interventions are often challenging to implement, though they are necessary due to the considerable diversity among the students (Butler & Stevens, 1997a).

Equally challenging is the assessment dilemma presented when these same students are required to take standardized tests in English, regardless of their level of proficiency in the language. The assessments are used to measure the students' academic progress and also the progress of individual schools and school districts. All constituents, and especially the students themselves, need to know when and how much progress is being made. However, the data from such assessments can be misleading since students cannot reasonably be expected to demonstrate what they have learned in the content areas if they cannot read or write well in English. Furthermore, the tests may not be "norms appropriate," because tests that are used with a student group not represented in the norming studies may result in the instability of the normal distribution of student scores. This potential misuse may result in scores that are unreliable and possibly invalid (Davidson, 1994).

On the other hand, excluding students who are English language learners (ELLs) from accountability reporting is equally misleading, as the population's scores are not represented fully and thus programmatic decisions about the population may be inappropriate. What is critical, then, is determining valid indicators of students' academic progress in school, indicators that accurately reflect what students are learning, both in content and, when appropriate, in language. Within the overarching goal of monitoring achievement, both large-scale content assessments and measures of academic language proficiency have critical roles to play. The challenge is to determine how to utilize both effectively in the service of providing as comprehensive a picture of ELL student progress as possible.

One approach to inclusion research focuses on determining how proficient a student must be in English for an achievement test score to be a valid indicator of what that student has learned in the content areas. The goal is to be able to say, with a high degree of confidence, that for a student who achieves some criterion level of performance on a language test, weak performance on a content assessment is most likely not due to a low level of English reading proficiency. This report explores the relationship between the language assessed on a widely used language proficiency test and the language used on a standardized content assessment. To do so, the language tapped by the language assessment is compared, from an academic language perspective, to the language used on the content assessment, and concurrent student performance data on the two tests are examined. The research questions for the study are:

1. How does the language that is measured on a language proficiency test compare with the language that is used on a content assessment?

2. What is the relationship between student performance on these two types of tests? That is, what is the correspondence between ELL English reading proficiency and performance on a content assessment?

To answer the first question, the language features of each assessment are described and then a comparison of the features is presented. To answer the second question, analyses of student performance, including correlations between ELL performance on the two tests and item response analyses, are provided. In addition to overall student performance on the tests, closer inspection of individual item performance and distractor behavior sheds light on student strategies for selecting options.

Recent Related Research

Recent research by Abedi, Leon, and Mirocha (2000) compared students' performance on a content assessment with their language proficiency status within schools, districts, and states. Butler and Castellon-Wellington (2000) compared student content performance to concurrent performance on a language proficiency test. Both studies established a correlational relationship between English language proficiency and performance on standardized achievement tests in English. The nature of that relationship, however, remains unclear. There is a considerable range in the magnitude of the correlations, indicating that factors in addition to language are affecting performance. Thus, a next step in articulating the relationship between language proficiency and performance on large-scale assessments involves describing and comparing the language that is measured on widely used language tests and the language that is used on standardized content assessments. In addition to helping to clarify the linguistic relationship between the two, such a comparison would aid in the development or selection of language tests that identify students capable of processing the language on standardized content assessments.

Little research exists on the relationship between academic language, language proficiency tests, and performance on standardized content assessments. Ulibarri, Spencer, and Rivas (1981) compared the performance of 1st-, 3rd-, and 5th-grade Hispanic students on three English language tests with their achievement data for reading and math. They found that the language test data were not very useful in predicting achievement in reading and math. The authors suggested that the weak correlation between the two types of measures could be due to the proficiency categories defined by the language tests. More recently Butler and Castellon-Wellington (2000) compared the performance of 3rd- and 11th-grade ELL students on one of the language proficiency tests used by Ulibarri et al. to the students' performance on a standardized content achievement test. They found variation in the predictive ability between the language measure and content performance in reading and math at the 3rd-grade level. They also found predictive validity between the language measure and content performance in reading, math, science, and social science at the 11th-grade level. The language measure accounted for as much as 50% of the variance in content performance in some instances and as little as 16% in another, with 25% of the variance accounted for in most instances. The lack of item-level data prevented closer examination of student performance, which might have helped explain some of the variability.

In related research, Cunningham and Moore (1993) looked at the academic vocabulary used in comprehension questions on a standardized reading assessment with fourth-, fifth-, and sixth-grade students who were native speakers of English. They found that performance was significantly higher when the *test language*—specific jargon usually associated with test questions such as "Choose the best answer for each question"—was reduced. Bailey (2000) found that academic test language is a significant component of the language of standardized tests and hypothesized that it may contribute to reading difficulty for ELLs.

Implicit in the work reported here is the notion that academic language underlies the content matter constructs being tapped by the large-scale assessments or at least that academic language is the conduit of the concepts being tested. The degree to which the most commonly used language proficiency tests reflect academic language is therefore important. A working definition of academic language will facilitate discussions of the language on the two tests being analyzed. Thus, the exploration of the role of language in large-scale assessments begins with a discussion of academic language.

Academic Language

The term *academic language* commonly refers to the language used in the classroom or other academic contexts for the purpose of acquiring knowledge. While many would agree with this general description of academic language, the specifics are often contested because the concept has not been fully operationalized on the basis of empirical evidence. For this reason, researchers and educators frequently have differing views of what constitutes academic language.

Solomon and Rhodes (1995) identified two dominant theories of academic language in the literature. The first defines academic language primarily in terms of the language functions and corresponding structures that students must use in the classroom. The second model proposes a distinction between academic language and social language, with emphasis on context and cognitive difficulty. Most currently used definitions of academic language integrate these two views. Solomon and Rhodes proposed an alternative view, a sociolinguistic view that defines academic language in terms of register, with language features that vary according to context and task.

In the literature of the first dominant view, there is a focus on language functions, that is, on how individuals use language to accomplish specific tasks, such as to inform, analyze, and compare (Halliday, 1975; Wilkins, 1976). Chamot and O’Malley (1994) describe academic language primarily in terms of the language functions used by teachers and students for the purpose of acquiring new knowledge and skills. “In grade-level content classes, students need to be able to understand teachers’ explanations, discuss what is being learned, read for different purposes, and write about their learning” (Chamot & O’Malley, 1994, pp. 40-41). Academic language may be global, used commonly across a variety of content areas, or content-specific, used exclusively in a single content area (Hamayan & Perlman, 1990; O’Malley, 1992; Chamot & O’Malley, 1994). In addressing this distinction between academic language that is general and academic language that is content specific, Kinsella (1997) stated:

> Students must master the specialized terminology of various fields of study along with the discourse features characteristic of very different disciplines such as science and literature. Understanding the distinct expectations on assignments which stipulate directions such as *analyze, compare,* or *trace* and recognizing the critical shifts in focus when lecturers or writers employ transitional signals such as *moreover* or *nevertheless* are examples of the multiple forms of academic language proficiency necessary for success in secondary and higher education. (p. 49)

Cummins (1980) theorized the second prevalent view with his distinction between academic English, which he called Cognitive Academic Language Proficiency, and social or conversational language, known as Basic Interpersonal Communicative Skills. This distinction is largely based on the degree to which language is contextualized and the cognitive demand of a given task or situation. Cummins (1984) proposed conceptualizing language proficiency along two continua, the first addressing the range of contextual support available for the construction of meaning and the second addressing the degree to which a communicative task is either cognitively demanding or undemanding.

Finally, the sociolinguistic view Solomon and Rhodes (1995) proposed is that academic language is a register that includes task-specific stylistic registers. For example, discussing the results of a science experiment would involve a different register than discussing the results of a mathematical word problem, as well as different ways of presenting and articulating the subject matter. Others have taken a similar, integrative approach, defining academic language to include “speech registers related to each field of study” (Teachers of English to Speakers of Other Languages, 1997, p. 153).

The sociolinguistic view of academic language as a register is not a new one. Anthropologists, sociologists, and sociolinguists were responsible for some of the earliest writings on the differential nature of the language demands made in school. Basil Bernstein (1961), for example, claimed that there were at least two distinct varieties of language used in society, a restricted or public code and an elaborated or formal code. He suggested that while every speaker has access to the public code, “a language of implicit meaning,” not everyone has access to the formal code, language that can be used to form “a complex conceptual hierarchy for the organizing of experience” (p. 169).

Other researchers have focused on the sociocultural and sociolinguistic differences between home and the classroom and how these differences manifest themselves in terms of student academic performance (e.g., Heath, 1983; Philips, 1972). These researchers helped to point out how home culture and language-use influence many aspects of language and learning in the classroom, including pragmatics, language variety, register, and semantics.

In this report, academic language is defined to include all three perspectives; it is distinguished from social varieties of English in multiple ways, including register, pragmatics, grammar, and discourse. As with all language, academic language falls on a continuum of language use that depends on situation, task, and interlocutors. Academic language is used specifically for the purpose of communicating through written or oral expression on academic topics to complete a range of tasks in academic settings, though it is used outside these contexts as well. Reading a textbook, discussing a poem, or taking a standardized content assessment are all examples of academic tasks that may require variation in register, or the stylized registers discussed by Solomon and Rhodes (1995). In other words, these tasks may require the use of specific discourse, functions, vocabulary, and/or structures not used across other academic tasks.

The notion of academic language as defined above provides perspective for examining the language content of two assessments and student performance on those assessments. A description of the research methodology follows.

Methodology

This section of the report provides an overview of the study and analyses of the language of the two selected tests.

Research Design

This study is a small-scale experimental effort that takes both qualitative and quantitative approaches to analyzing the relationship between the language on a language proficiency test and a standardized content test. The rationale for this approach is an outgrowth of previous research (e.g., Butler & Castellon-Wellington, 2000, and Abedi et al., 2000) that established correlational relationships between the two types of tests but lacks the item-level data needed to investigate causes of variability in the magnitude of the correlations. Thus the goal of this study is to provide the groundwork for much-needed larger scale research that can more thoroughly address the variability of performance of ELLs on language and content tests.

Two types of analyses are provided in this report. The first is an analysis of the language of the two assessments used. Descriptions of the two instruments are followed by a comparison of the language used in those two instruments. The instruments have different purposes, so the discussions of the language in the two tests are organized in ways that best reflect the language used on each test. Thus, the sections used to describe each test in the report are not parallel; each has slightly different categories that do not necessarily overlap. The content of the language assessment is described according to the subsections of the test; the language of the content test is described in terms of the items themselves. The assessments are compared afterwards in terms of linguistic features.

The second type of analysis is data-based and includes descriptive statistics, correlational data, and item-response pattern analyses. Since the goal of the study is to examine differential student performance on a language proficiency test to help determine causes of variability in scores on a content assessment, emphasis is placed on analyzing the item-response patterns of two subgroups of ELLs—the highest and lowest scoring ELLs on the content test. A comparison follows below of the response patterns of the two ELL groups and the English-only (EO) students in the study. Response patterns may suggest whether the students were making logical choices when they did not know the answer to an item or whether they exhibited behavior that indicates a large amount of guesswork. Differences in item-response patterns may also help to determine whether students were having difficulties with language, content, or both. It was hypothesized that the response patterns of the top-scoring third and bottom-scoring third of the ELL sample would vary due to language proficiency differences, as determined by ELL performance on the language proficiency test, and that the EO group responses would be less random, or due to chance, than the responses of the other two groups.

Limitations of the Study

The study reported here is exploratory and, as indicated above, is intended to lay groundwork for larger scale research efforts. There were limitations to the design and data collection, which introduced confounding factors. First, the number of EO students is extremely small. Those EO students in the study were in sheltered classes where teachers were expected to provide both sheltered instruction for ELLs and regular instruction for EOs. Control EO classrooms were not included but should be incorporated in future research to allow for more balanced comparisons. Second, the time between administration of the language proficiency test and the content assessment was one month. This time gap may have affected student improvement in English. Because of these limitations, generalizations from the results must be made with caution.

Setting and Subjects

Data collection took place in 1997 in the Burbank Unified School District (BUSD) in southern California.[2] Of the 14,350 students enrolled, 6,713 spoke languages other than English at home; of those 6,713 students, 3,277 were designated limited English proficient (LEP).

Six 7th-grade social studies classrooms from three schools were selected by the district to participate in the study on the basis of their sheltered status.[3] The students in the study spoke 10 languages besides English, with Spanish and Armenian speakers being the largest groups. In all, the sample consisted of 102 seventh-grade ELLs and 19 EO students. The classrooms were composed primarily of ELLs whose English language proficiency was not high enough to allow them to be placed in mainstream English-only classrooms without language support.

Two classrooms consisted exclusively of ELLs who received sheltered instruction in social studies. Three additional classrooms were composed of a mix of ELL students receiving sheltered instruction and EO students. EO students from two of those classrooms participated; the third classroom did not want its EOs to participate. The sixth classroom was a designated English language development (ELD) class. The students in this class were newer to the United States than ELLs from the other classes. However, the social studies content in this classroom was also delivered using sheltered techniques.

Instruments

The Language Assessment Scales ([LAS] Duncan & De Avila, 1990) and the Iowa Tests of Basic Skills (ITBS) Social Studies Test for Seventh Grade (Level 13), Form L, (University of Iowa, 1993a) were used in the data collection and language analyses described below. The LAS Reading Component, Form 3A, was selected for analysis because it is a widely used test of language proficiency and specifically is used by BUSD for placement purposes. The ITBS was selected because it is one of the most widely used standardized test batteries in the United States for Grades 1-8, and its content is representative of the test materials ELLs must be able to process linguistically.

[2] Data were collected in the 1997–1998 school year.

[3] In sheltered classrooms, the teacher delivers the content using special techniques designed to make content more accessible to students whose English skills are still developing.

The LAS Reading Component and the ITBS Social Studies Test were administered to the same students within approximately one month of each other. Informal posttest focus groups with the students and interviews with the teachers were also conducted.

Language features of the LAS Reading Component. The LAS Reading Component (Form 3A) is part of the LAS Reading and Writing Assessment for Grades 7-9+.[4] The content areas for Form 3A include Vocabulary (Synonyms and Antonyms), Fluency, Reading for Information, and Mechanics and Usage. Table 1 below indicates the subsections and number of items included in Form 3A.

The items are all objective, multiple-choice items. Students fill in ovals on a separate answer sheet, not in the test booklet. The Examiner's Manual (1988) suggests that the LAS Reading Component be administered in a 45-55 minute session. LAS Reading is not intended to be a timed test, with the implication that the amount of time allotted is ample for most students to complete all items. A discussion of the language assessed by each subsection of LAS Reading follows.

Vocabulary. Vocabulary is difficult to classify. It is often described in terms of frequency and context of use. However, frequency and context of use are dependent on experiential factors such as age and education. Therefore, the context of use for any one vocabulary word may shift, particularly with age and grade level, and as individuals gain world knowledge and experience, uses of that word may increase.

Table 1

Number of Items in LAS Reading Component Subsections for Form 3A

Subsections	No. of items
Vocabulary	
Synonyms	10
Antonyms	10
Fluency	10
Reading for Information	10
Mechanics and Usage	15
Total	55

[4] There is a separate test form for each of three grade clusters and an alternate form available for each level, as follows: Form 1 A / B (Grades 2–3), Form 2 A / B (Grades 4–6), and Form 3 A / B (Grades 7–9+).

In this report, we will refer to three categories of words: high-frequency general words, nonspecialized academic words, and specialized content area words.[5] High-frequency general words are used regularly in everyday contexts and are defined as "basic words that communicate ideas, feelings, and actions (for example, commonly used adverbs and nouns); proper nouns used throughout printed matter; and connective words used to join and express complex relationships among sentences" (Daines, 1982, p. 120). Nonspecialized academic words are a subcategory of Scarcella and Zimmerman's (1998) *academic words* classification and are defined as the academic words used across multiple content areas. For example, in the question "What do the stars on the American flag represent?" r*epresent* is an example of a nonspecialized academic word. Specialized content area words are defined as vocabulary unique to content areas. At the seventh-grade level, *antitrust* is a specialized word found in a social studies word list culled by Criscoe and Gee (1984) from a review of the glossaries of social studies texts for that grade level.

Vocabulary is divided into two sections on Form 3A of LAS Reading, Synonyms and Antonyms. Students must read one word and then select the synonym or antonym from four potential responses. These vocabulary items are isolated from any visual or textual context. Approximately half are conceptually concrete; that is, these words can be represented visually—for example, words that describe feelings or attributes, like *sad* or *fierce*.[6]

Generally, the vocabulary on LAS Reading consists of high-frequency general words, including common nouns, adjectives, verbs, and adverbs used in everyday contexts, and some nonspecialized academic words. Some words may be difficult for ELLs because they are isolated from context and are conceptually abstract words, like *amazing* and *sensible.* Additionally, the words sometimes have multiple meanings and forms. Words with multiple meanings also tend to be problematic because ELLs can have difficulty distinguishing between common and academic uses of such words (Scarcella & Zimmerman, 1998). An example of such a word is *sound*, which students will first learn to associate with hearing; however, it is used in academic contexts to indicate the quality of an argument or theory (e.g., his ideas are sound) and to refer to a long passage of water in geography. Table 2 provides

[5] There are other vocabulary categories beyond the scope of this report, specifically those used outside academia, such as technical vocabulary for specific work environments.

[6] Note that the examples used in this report are not actual items from LAS Reading. Rather, they are examples created by the authors to illustrate the types of items that appear on the test.

Table 2

Examples of Words With Everyday and Academic Uses

Word	Everyday usage	Academic usage
Stem	*stem* of a plant	*stems* from the belief
Point	don't *point* your finger	the author's *point* was clear
Exercise	*exercise* daily	*exercise* your stock options

additional examples of high-frequency general words that have both everyday and academic uses.

Finally, 83% of the vocabulary is multisyllabic. If a word is multisyllabic, yet morphologically transparent, such as *unfortunately*, students may be able to determine meaning through analysis of the parts. The more morphologically transparent a word is, the better the chance that students will be able to use a combination of word analysis skills, word knowledge, and context to guess the meaning of an unknown word (Nagy, Anderson, & Herman, 1987; Nagy & Scott, 1990). Few of the words in these two vocabulary sections can be considered transparent, indicating that students would not be able to use word attack skills to analyze the vocabulary.

Fluency. The Fluency section consists of 10 items. The Technical Report (Duncan & De Avila, 1988b) states that these items "measure overall language fluency and the ability to infer a missing word based on a knowledge of language usage and semantics" (p. 8). With the exception of one item that contains two sentences, all of the prompts consist of one sentence with a missing word. Students must select the word from a list of four alternatives that best completes the sentence. Parts of speech assessed include nouns, verbs, adjectives, and adverbs.

The topics in the sentences are general and, though some relate to school, they are not academic topics that might be found in textbooks or on a standardized content test. Vocabulary consists of high-frequency general words. Average sentence length is 11.3 words with a range of 7 to 15 words. As average sentence length increases, the sentences tend to become more linguistically complex, containing more embedded clauses.

In order to answer the questions correctly, students must be able to demonstrate an understanding of both syntactic and semantic relationships.[7] For example, in the sentence "The _____ is riding a bicycle," students must understand that the word following *the* is a noun and that it is most likely a boy, girl, man, or woman. The ability to make associations such as these is essential for students to accurately complete the items in this section.

Reading for Information. The Reading for Information section consists of a short passage followed by 10 comprehension questions. The LAS Examiner's Manual (Duncan & De Avila, 1988a) states that the items are intended to "measure the ability to identify information" (p. 3). All but one item, which requires students to make an inference, are *identify* items that require students to scan the text for information.

The passage is expository and includes the following rhetorical structures: enumeration, cause and effect, generalization-example, and chronological ordering. The passage contains five paragraphs, with sentences ranging in length from 11 to 27 words. Five of the longest sentences consist of two independent clauses joined by a coordinating conjunction. There are many proper nouns, including names of famous people and countries. The tenses used include simple present and past, present and past perfect, and mixed tenses within sentences. There are many embedded clauses throughout the passage, including numerous temporal clauses, relative clauses, and prepositional phrases. In general, the text reflects the type of complex language found in textbooks. However, the test question structures and vocabulary follow the text so closely that students are not required to manipulate the language.

Eight prompts are in sentence-completion format, and two prompts require the test taker to select the correct option to insert at the beginning of the sentence. Average prompt length is 8.3 words, with a range of 5 to 13 words. Response options range from 1 to 6 words in length. Four of the 10 items require students to identify the correct proper noun from the response options. These items are considered to be one-word response options. Sentence structures are similar to those in the passage, with two prompts paralleling the text word for word, and two others almost exactly identical to the text. Four prompts are partially paraphrased; two are completely paraphrased, one of which is the only inference item.

[7] Semantics is defined here as the "study of meaning and the systematic ways those meanings are expressed in language" (Hatch & Brown, 1995, p. 1).

Mechanics and Usage. The Mechanics and Usage section includes 6 mechanics items and 9 usage items. Students must read a sentence and then select the response that completes the sentence: a punctuation mark, capitalization choice, or usage choice. Usage items include morphology, syntax, and discrete grammar points, such as possessive pronouns, articles, subject-verb agreement, and coordinating conjunctions. Average sentence length is 5.5 words, with sentences that range in length from 3 to 7 words. Sentence structures are very simple and include the use of some short temporal and prepositional phrases.

Summary. It appears that LAS Reading measures student ability to identify information in a text, recognize a range of decontextualized vocabulary words, select semantically appropriate words to complete sentences, and identify the appropriate punctuation, capitalization, and word forms in short, simple questions and sentences. Although the Reading for Information text is similar to that found in social studies texts, students are required only to scan the text for most of the answers. Furthermore, the prompts are in sentence completion format and often share identical linguistic features and vocabulary with the sentences in the text that contain the answers. In other sections of the test, the vocabulary and subject matter consist of general topics and high-frequency vocabulary, with few exceptions. Some vocabulary words can be categorized as nonspecialized academic words. Taken together, these features appear to be representative of general language use. LAS Reading contains few features of academic language—such as a range of nonspecialized academic vocabulary, complex syntactic structures, or language functions—that would align it with the definitions of academic language discussed earlier.

Language features of the ITBS Social Studies Test. The ITBS Social Studies Test is part of a subject area achievement test battery designed to measure how well students have learned the basic knowledge and skills taught in school. There is a test for each grade level, from Grades 3 through 8. The Level 13 seventh-grade social studies test (Form L) was used in this study. It contains 44 multiple-choice questions. Students must choose among four options and record their answers on a separate answer form. The test measures a combination of social studies facts and skills in the following content areas: history, geography, economics, political science, sociology and anthropology, and related social sciences including ethics, law, and religion. Examples of the range of topics and time periods found on the social studies test include ancient Egypt and Greece, the Middle Ages, the Industrial Revolution, and world history since 1900.

Subject matter in the ITBS Social Studies Test is cumulative, meaning that it includes a range of social studies topics and skills taught from Grades K through 7, from antiquity to modern history, excluding state histories. Each state covers many of the same social studies topics but not necessarily in the same sequence from grade to grade. Therefore, the test may contain questions from topic areas that are covered by the seventh grade in some states but not others. For example, in California, American history from the early 18th century to the present is not covered until the eighth grade. Thus, items drawn from this time period may be difficult for seventh-grade students in California since they may not have been exposed to the test material for that topic.

In the remainder of this section, the language of the test will be characterized through a description of the items, including the features and syntax of the prompts, stems and response options, functions, and vocabulary.

Features of prompts and stems. The questions on the ITBS Social Studies Test are all structured in generally the same format with variation occurring in the prompt type and stem formats. There are three main prompt types: stem-only (64%), discourse-with-stem (9%), and a visual-with-stem (27%). A *stem-only* prompt is a prompt that consists of a question only, such as "Which of these inventions made it practical to construct tall office buildings?"[8] [Individual sample item provided by Riverside Publishing Company, Chicago, Illinois, Grade 6, Social Studies.]

A *discourse-with-stem* item consists of one or two sentences followed by a question. The discourse provides background information students must understand in order to answer the question appropriately. This type of item almost always requires students to make an inference. Although it is a science item, Sample Item # 1 below offers a clear example of a discourse-type prompt.

The last prompt type, the *visual with a stem*, is composed of a visual, such as a graph, a timeline, or a political cartoon, followed by three to five items. The number of questions asked ranges from three to five per visual. These visuals contain additional text, such as labels consisting of single words or sentence fragments

[8] The item samples used were provided by Riverside Publishing Company, Chicago, Illinois, as sample items and are not taken from the ITBS Social Studies Test.

Sample Item # 1

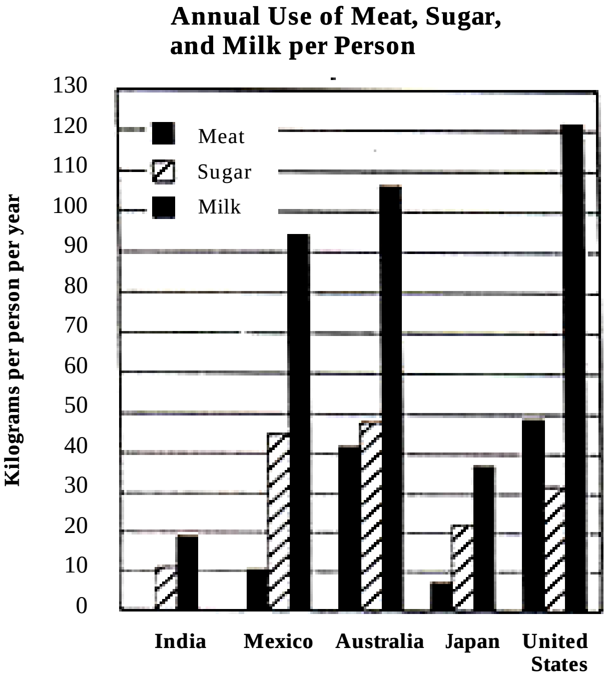
Two balls are the same size, but the white ball floats in water and the green ball sinks. What does this observation tell us?

a) The green ball is heavier than the white ball.[9]
b) The white ball has air in the middle of it.
c) White things float better than dark things.
d) Round things float better than square things.

(University of Iowa, 1996a, p. 15)

and/or dates. Students must use the visual and any accompanying text to answer the questions. In Sample Item # 2, students are given a bar graph to which they must refer in order to answer the question.

Sample Item # 2



Which two countries had about the same level of sugar usage per person?

a) India and Japan
b) India and the United States
c) Japan and the United States
d) Mexico and Australia

(University of Iowa, 1996b, p. 13)

[9] Answers to sample items are italicized.

Syntax of the prompts and response options. The social studies test contains prompts and response options with varying degrees of syntactic complexity. The syntax of the stems can range from simple constructions, such as “What was the Underground Railroad?” (University of Iowa, 1996b, p. 14), to complex constructions, such as “What is the fewest number of times Justin will need to shovel his neighbor’s walk and driveway to earn enough money to buy the skates?” (University of Iowa, 1996b, p. 12). The stems range in length from 5 to 20 words.

Items on the test may contain negation, such as “Which animal probably would *not* be found in a forest habitat?” (University of Iowa, 1996b, p. 15). Others may contain the passive voice, mixed tenses, conditionals, extended noun phrases, or extended and consecutive prepositional phrases. Some items contain multiple clauses and propositions,[10] such as the example in the last paragraph about shoveling the driveway. Multiple clauses and propositions can make both the reading and cognitive load more difficult because the item has multiple elements that must be processed both in terms of language and content. However, depending on the item, there may be fewer or more propositional elements.

Many items contain test language or jargon, such as “which of the following,” “which of the following is *not*,” “which statement best describes,” or “which is the best evidence.” An example of a question that contains test language is “Which of the following best explains why Chicago is a transportation center?” This feature may present additional challenges to ELLs, especially those students with lower levels of English proficiency, because they must focus on the phrase *best explains* to answer the question correctly. In other words, all of the distractors may hold some element of truth; however, students must understand that the task is to select the option that answers the question most precisely.

The social studies response options appear in three formats: (a) 1- or 2-word options, (b) sentence fragment options, or (c) complete sentences. The 1- or 2-word response options appear the least frequently (9% of the items) and usually consist of a noun or proper noun. Fragments appear most frequently (52% or more) and range from 2 to 14 words in length. Complete sentence response options are the second most frequent response format (39%). When full sentences are used as response options, they can be as long as 21 words. The length of the sentence distractors can vary within an item such that one response option may be only 5 words long and

[10] A proposition is defined here as “the content of the message the speaker wishes to express” (Celce-Murcia & Larsen-Freeman, 1983, p. 524).

another response option for the same item may be 10 words long. As with the prompts, the response options vary in degree of complexity. Options contain relative clauses, adjectival phrases, passive constructions, modals, complex noun phrases, and a variety of tenses. For some items, each response option contains a different structure and tense, adding complexity to the task of reading and selecting the best option.

Functions. "Language functions refer to how individuals use language to accomplish specific tasks" (Halliday, 1975; Wilkins, 1976). That is, language functions are what people do by means of language, such as order food in a restaurant, greet a colleague at work, or justify an opinion. Cognitive functions, on the other hand, involve cognition or thinking skills, such as drawing conclusions, comparing graphs, or evaluating statements, and range in difficulty from lower order functions, such as *identify*, to higher order functions, such as *synthesize* (see Bloom's *Taxonomy*, Bloom, Engleheart, Furst, Hill, & Krathwohl, 1956). Cognitive functions and language functions overlap when language is used to demonstrate cognition.[11] The important distinction is that language functions are the verbal or external expressions of internal thought processes. In large-scale assessments that utilize multiple-choice options, students are required to internally process language functions before selecting an answer. In contrast, in open-ended performance assessments, students are asked to construct their own responses and thus employ their ability to use a variety of expressive language functions.

In the ITBS Social Studies Test, the students must select the answer that is factually most accurate or the answer that is aligned most closely with the inferences arrived at through cognition. Although the test items require students to evaluate, identify, analyze, infer, compare and contrast, and classify, the majority of the items (55%) are factual and thus only require students to evaluate and identify options.

Vocabulary. The wide range of vocabulary used in the ITBS Social Studies Test falls into all three categories defined above in the description of the vocabulary on LAS Reading: high-frequency general words, nonspecialized academic words, and specialized content area words. Not surprisingly, there are many high-frequency general words on the ITBS. The ITBS also uses many nonspecialized academic words—words used across academic contexts and registers, such as *evidence* and *represent*—as well as specialized content words. Specialized content words are

[11] Note, however, that cognition can be demonstrated physically, verbally, or not at all, since it is an internal process.

words used in a specific content area; *feudalism* and *democracy* are examples used in social studies. Specialized content words also include scientific terms and compound words like *energy consumption*.

Many of these content words are conceptually abstract and critical to understanding in the content area. Short (1994) noted that ELLs require instruction that helps them make connections between such words and their own lives or familiar current events. The ITBS Social Studies Test contains a wider variety of words from the two latter word groups mentioned above than the LAS Reading, which consists primarily of words from the first category and some nonspecialized academic words. Many of the words on the LAS can be found on the ITBS. The converse, however, is not true.

Summary. Although the ITBS Social Studies Test does not contain any extended reading passages, it does contain a wide variety of sentence structures and vocabulary. Students must be able to process long prompts and response options with multiple embeddings. The items also include test language, which adds to sentence length and sometimes to complexity. Vocabulary is sophisticated and includes many nonspecialized academic words as well as specialized content area words. Students must use a variety of cognitive functions to answer items correctly. Questions appear similar in format, structure, and vocabulary to those found in textbooks, such as the Houghton Mifflin seventh-grade social studies text *Across the Centuries* (Armento, Nach, Salter, & Wixson, 1991), aligning the language of the ITBS with the definition of academic language noted earlier.

Comparison of the Language on the LAS Reading and the ITBS Social Studies Test

The language assessed on the LAS Reading (Form 3A) and the language used on the ITBS Social Studies Test (Form L) differ in multiple ways, including topics, the use of test language, discourse, functions, syntactic complexity, and vocabulary. The following is a brief summary of differences between the linguistic features of the two tests.

Test Topics

The subsections of the LAS Reading component contain general topics that may be used in common, everyday contexts, with one exception. The Reading for Information section contains an expository passage on a topic that might be found in a social studies text. In contrast, the ITBS Social Studies Test taps a range of social studies topics, including economics, geography, history, and political science. Some items require students to read additional text or use visuals like those found in textbooks, such as graphs, pictures, and timelines.

Test Language

Bailey (2000) states, “The test-taking routine is a conventional script with specific structures that need to be learned” (p. 89). Indeed, not only is it a conventional script, *test language* is a special linguistic register of academic language that is marked by the use of formal grammatical structures and tends to be formulaic; for example, it contains phrases such as “which of these best describes.” Its characteristics often add to the length and syntactical complexity of sentences without actually adding content. For example, the question “Which of the following best describes the traffic in Los Angeles?” can be translated into “Which sentence describes traffic in Los Angeles?” If some of the distractors are partially correct, then it is necessary to use a superlative such as *best*. In cases where there is only one correct option, using a superlative is unnecessary. Superlatives are frequently used on the ITBS, even when there is only one correct option and all others are completely incorrect.

LAS Reading contains neither the formal structures of test language nor the formulaic phrases found on the ITBS. In particular, the LAS Reading for Information section, which has the greatest potential for use of test language because of its emphasis on reading comprehension, seems to avoid test language phrases and constructions altogether. Instead, items are written in language similar to the text. Thus, students can scan the text for key words and structures and potentially answer some items without reading the entire text.

Discourse

The discourse styles of the two tests are very different. Except for the LAS Reading for Information passage, LAS Reading consists of prompts that are in either single word or sentence completion form. The Fluency and Mechanics sections sometimes require students to read two sentences of connected discourse in a single prompt. The Reading for Information section consists of an extended reading passage five paragraphs in length with single-sentence completion items. The distractors throughout the entire test are all sentence fragments because all the items are sentence completion prompts.

The ITBS Social Studies Test, on the other hand, has a variety of prompt and response types, sometimes requiring students to read two or three sentences of connected discourse or to use visuals in order to answer the questions. Although the ITBS does not contain extended text (i.e., a single paragraph or more), all prompts are written in complete sentences, as are many of the distractors.

Syntax and Grammar

The syntax and grammar of LAS Reading items are less complex than that of the ITBS. LAS Reading items are shorter and contain fewer embedded clauses, temporal markers, passive constructions, or other textual features found in the ITBS Social Studies Test items. Though the Reading for Information section contains the type of language found in the social studies content area, students are not required to manipulate it. As noted above, because the syntax of the items in the Reading for Information section is usually parallel to the syntax in the passage, students can scan the passage for the sentence that contains the answer.

The ITBS Social Studies Test includes items with syntactic features that add to the linguistic complexity of the items, such as multiple embeddings, temporal markers, causative structures, and comparatives. The formal test language discussed above adds to this complexity. Furthermore, the length and linguistic variety in many of the response options increase the amount of language processing needed. Although some of the longer responses are not linguistically more complex, others are. For example, some responses contain relative clauses, modals, mixed tenses, passive voice, and complex noun and prepositional phrases. Thus, in order to answer questions correctly, students must be able to process and understand a greater variety and quantity of language in both the prompts and response options than is necessary for LAS Reading items.

Functions

Since LAS Reading and the ITBS are not performance-based assessments, neither test requires students to accomplish expressive language tasks. Instead of actively using language functions to respond to the prompts, students must be able to process functional language in the test items, response options, and additional text and visuals in order to select the correct response.

There is a contrast, however, in the number of functions that appear in the items on the two tests. One item on LAS Reading requires students to make an inference while almost half of the ITBS items require students to infer, analyze, compare, classify, or reason in order to select the correct answer. As mentioned earlier, although the LAS Reading for Information passage includes more sophisticated structures and functional language content than the other sections of the test, 9 of the 10 items do not tap the understanding of this functional language. Thus, the ITBS, more than LAS Reading, requires students to process and show evidence of understanding a greater variety of functional language.

Vocabulary

The vocabulary that appears on LAS Reading consists of high-frequency general words and a few nonspecialized academic words. The vocabulary coverage on the ITBS Social Studies Test is expanded to include many nonspecialized academic words and specialized content area vocabulary used in the social studies disciplines. Therefore, though the LAS vocabulary can be found on the ITBS, many words on the ITBS are not on the LAS. The vocabulary tapped on the LAS is decontextualized, whereas the vocabulary used on the ITBS is contextualized. In addition, there are many long, multisyllabic words on the ITBS that have multiple meanings and specialized uses in academic contexts.

Summary

The language of the LAS is less complex, more discrete and decontextualized, and more limited in its range of grammatical constructions than the language of the ITBS. The ITBS Social Studies Test contains content-specific academic language reflective of the language that appears in textbooks, whereas the LAS contains more generic language, common to everyday contexts. Furthermore, because the ITBS includes a variety of item-response formats (e.g., complete sentences, sentence fragments) and prompts (e.g., question, question with visual), students are required to process a wider variety of language on the ITBS than on the LAS.

The correspondence between the language used on the LAS and the language used on the ITBS is thus limited. The vocabulary and linguistic structures used on the LAS are common in everyday life and for this reason are likely to appear on content-based assessments. However, the level of language measured by the LAS is not sufficient to indicate student ability to process the language of these assessments. The increase in level of syntactic complexity, variety of sentence structures, and the expanded vocabulary on the ITBS require a more sophisticated language associated with academic discourse.

Data Analyses and Results

The data analyses provide the following: (a) descriptive statistics for ELLs and EOs, (b) correlations for ELLs, (c) a description of section-level performance on the LAS for ELLs, and (d) a description and analysis of item-level performance on the ITBS for two ELL subgroups.

Descriptive Statistics

Table 3 shows the descriptive statistics for the ELLs and EOs on both the LAS and the ITBS. ELLs had a mean raw score of 37.82 (*SD* = 8.2) out of 55 possible on the LAS. The minimum and maximum scores were 19 and 55 respectively. The scaled score for the mean was 69, which falls in the middle of Reading Competency Level 2, the Limited Reader level**.** The mean LAS score for the EOs was 47.72 (*SD* = 6.33), which falls into the Competent Reader level, with minimum and maximum scores of 30 and 54 respectively. The range of LAS scores placed ELLs in Non-reader, Limited, and Competent Reader categories, and EOs into the Limited and Competent Reader categories. The LAS reading levels and their corresponding scores are presented in Table 4.

Table 3

Descriptive Statistics for ELL and EO Performance on the LAS and ITBS

	Test	*N*	No. of items	*M*	*SD*	Min	Max
ELLs	LAS	102	55	37.82	8.20	19	55
	ITBS	102	44	12.85	4.98	0	28
EOs	LAS	18	55	47.72	6.33	30	54
	ITBS	19	44	21.63	5.45	13	30

Table 4

LAS Reading Competency Levels

	Raw score range	Scaled score range
Non-reader	0–32	0–59
Limited Reader	33–43	60–79
Competent Reader	44–55	80–100

On the ITBS, ELLs had a mean raw score of 12.85 (*SD* = 4.98) out of 44 possible, with minimum and maximum scores of 0 and 28 respectively. The EOs' mean score on the ITBS was 21.63 (*SD* = 5.45) with scores ranging from 13 to 30. The percentile rankings for the two groups were 16th and 50th respectively. Percentile rankings were calculated using the midyear conversions (University of Iowa, 1993b). Again, note that there was a range of performance for both groups.

Test Correlations

Table 5 provides the Pearson Product-Moment Correlations for ELL total LAS Reading and subscale scores with the ITBS total score.

With the exception of the LAS Reading for Information subsection, all of the correlations for LAS subsections and the total LAS are significant at the .01 level for ELLs. However, the magnitude of the correlations is weak. For example, the r^2 for the strongest correlation, the total LAS with ITBS, is only .20, which means that performance on the LAS only accounts for 20% of the variance on ITBS scores. Thus, 80% of the variance is unexplained by LAS Reading.

Table 5

Pearson Product-Moment Correlations for ELL LAS Reading Scores (Total and Subsection) With ITBS Scores (*N* = 98)

	Total LAS	Synonyms	Fluency	Antonyms	Mechanics	Reading for information
ITBS	.4482	.4030	.3787	.3717	.2578	.1839
p	(.000)	(.000)	(.000)	(.000)	(.010)	(.070)

Test Reliability

Reliability coefficients were computed for the ELLs; the coefficients for the EOs were not computed because of the small *N* size. The reliability coefficient (coefficient alpha) was .864 for LAS Reading and .569 for the ITBS Social Studies Test. The published reliability for the ITBS Spring 1992 norming sample is .87. Note that only 0.9% of the seventh-grade norming sample was composed of LEP students. Another 1.2% were students in Migrant-Funded Programs and 0.8% were students in Bilingual or ESL Programs (University of Iowa and the Riverside Publishing Company, 1994). The small number of LEP students in the ITBS norming sample suggests that the ITBS is not intended for use with students whose English language skills are weak. The lower reliability for the ITBS with the ELLs in this study hints at inappropriate use of the test with these students.

Description of LAS Performance

As noted above, there was a range of performance for ELLs across subsections on LAS Reading, with the least amount of variation in performance on the Mechanics section. Minimum and maximum scores ranged from 0 to 10 on the Synonyms, Fluency, Antonyms, and Reading for Information sections; the range was 7 to 15 on the Mechanics section. The descriptive statistics for ELL performance on each subsection of LAS Reading are provided in Table 6.

Further analyses were performed to determine the difficulty of each section and of items within those sections. The Fluency section was the most difficult with an average *p* value of .57, followed by the Antonyms and Reading for Information sections with an average *p* value of .63, the Synonym section (*p* = .67), and lastly the Mechanics and Usage section (*p* = .85).

The Fluency section posed the greatest difficulty for the students overall. The *p* values for individual items varied from .24 to .86, indicating a considerable range of item difficulty. Although in most cases student responses were spread out evenly among the distractors, five items contained weak distractors that attracted only a small percentage of incorrect responses. It is evident that the students considered each option and selected the option that seemed most reasonable to them.

The only item ELLs did well on in this section as a group was an item constructed such that students could quickly eliminate distractors on the basis of one word in the sentence. Sample Item # 3 illustrates this type of item.

Table 6

Descriptive Statistics for ELL Performance on the Total LAS Reading and Subsections

	N	No. of items	*M*	*SD*	Min	Max
Total LAS	102	55	37.82	8.20	19	55
Synonyms	102	10	6.67	2.19	2	10
Fluency	102	10	5.77	2.33	1	10
Antonyms	102	10	6.36	2.32	1	10
Mechanics	102	15	12.72	1.89	7	15
Reading	102	10	6.30	2.71	0	10

Sample Item # 3

A __________ from the university spoke to our English class.

a) student
b) horse
c) confession
d) truck

(created by the authors)

To answer this item correctly, the student must recognize that the most critical word in the sentence for determining the answer is *spoke*. Understanding this verb disqualifies three of the possible responses. Although it would be grammatically correct to use *horse* or *truck*, neither is semantically possible, and *confession* is an inanimate noun. On many of the other items in this section, however, it appears that the students in our sample did not have the word knowledge or semantic knowledge needed to link key words to the answers.

In the Reading for Information section, the *p* values for individual items ranged from .36 to .83, again a considerable range of item difficulty. The most difficult item was the only inference item on the test. Responses were spread evenly among the distractors in this section, indicating that these items were performing better than items in other sections. Responses in the Reading for Information section indicated that students, in general, were able to scan a passage for information.

Performance on the Synonyms and Antonyms sections indicates that the ELL students sometimes had difficulty with conceptually abstract vocabulary, such as *sensible* and *expectation*. The *p* values on the Antonym section ranged from .42 to .81. On the Synonyms section, the *p* values ranged from .47 to .89 for individual items.

The *p* values were .57 or higher on all items in the Mechanics and Usage section. The range was .57 to .99 with an average of .85. In fact, the *p* values were above .80 on 11 of the 15 items. There were only three items on which ELL students did not perform as well, and these were all usage items. A high score on this section does not necessarily indicate student ability to process text since the questions focus on discrete usage and punctuation. Thus, the limitations of the content coverage in the Mechanics and Usage section combined with high *p* values suggest that performance on this section may actually be inflating overall test scores and could be misleading in terms of assessing reading proficiency.

Analysis of ITBS Performance

Overall, the EOs in this sample outperformed the ELLs on the ITBS Social Studies Test. As reported in Table 2 above, the mean score for EOs was 21.63 (*SD* = 5.45), and the mean score for all ELLs was 12.85 out of 44 (*SD* = 4.98).

EOs outperformed ELLs on 36 of the 44 test items. Their performance was equal on two items, and ELLs outperformed EOs on six items. Both groups did poorly on these six items. In comparing the overall performance of the two groups, it is difficult to determine the exact cause of the differential performance—language proficiency or opportunity to learn (OTL). There is a range of sentence length and complexity as well as topic areas covered across the 36 items, but there was no apparent pattern in the performance of ELLs as a group.

In an effort to determine which items were most problematic for ELLs and why, the students were divided into three groups according to their performance on the ITBS, and the item response patterns of ELLs in the highest scoring group (the top third ELLs) and the lowest scoring group (the bottom third ELLs) were analyzed.[12] The mean for the top third ELLs was 18.11 (*SD* = 2.98), and the mean for the bottom third was 7.71 (*SD* = 2.42). Table 7 below shows the descriptive statistics and percentile rankings for ITBS performance for EOs, the total ELLs, and the two subgroups of ELLs.

The mean LAS scores for the two groups of ELLs were 41.97 (*SD* = 8.16) for the top third and 34.42 (*SD* = 7.50) for the bottom third. The mean score for the top third ELLs falls into the top of the Limited Reader range, and the mean score for the bottom third ELLs falls into the bottom of the Limited Reader range. *T* tests were

Table 7

Descriptive Statistics and Percentile Rankings for Performance on the ITBS

Group	*M*	*SD*	Range	*N*	Percentile rank
English Only (EO)	21.63	5.45	13–30	17	50th
ELLs—All	12.85	4.98	0–28	102	16th
ELLs—Top 1/3	18.11	2.98	15–28	37	34th
ELLs—Bottom 1/3	7.71	2.42	0–10	35	10th

[12] See Sax (1989) for a discussion of item-analysis procedures, from which part of our methodology for analyzing item response patterns was taken.

performed to determine whether differences between the means on both tests for the two ELL groups were statistically significant. The differences between means for both LAS Reading (*t* = 3.99, *df* = 67, *p* = .000) and the ITBS (*t* = -16.19, *df* = 70, *p* = .000) were statistically significant, indicating that these two subgroups performed differently on both tests despite being categorized into the same level of language ability according to LAS Reading. This clear difference in ELL subgroup performance indicates that grouping ELLs according to existing LAS categories such as Limited Reader misses meaningful differences in performance within groups.

The next logical step would be to perform *t* tests to see if differences between the means of the top third ELL group and the EOs are statistically significant. However, due to the small EO sample size, this analysis is inappropriate, so the performance differences between all three groups are discussed qualitatively instead.

Although the EOs outperformed the top third ELLs on 28 items, compared to outperforming all ELLs as a group as noted above, the top third ELLs outperformed EOs on 12 items and performance was the same on four items. The percentile rank for the top ELLs (34th) is also much higher than the percentile rank for the ELLs as a group (16th) or the bottom third ELLs (10th), indicating that these students have different characteristics than the bottom third ELLs.

The performance of the top third ELLs was similar to the EO group in at least two ways. First, their overall performance on many items was comparable. On 26 of the 44 items (59%), the *p* values for the top third ELLs either exceeded the EOs' *p* values or were only 0-5% lower. Table 8 shows the top ELLs' performance compared to the EOs.'

Secondly, the item response patterns of this ELL group were similar to the EOs' response patterns. When the top third ELLs and EOs responded incorrectly, these ELLs tended to choose the same distractors as the EOs. By contrast, the bottom third

Table 8

Comparison of Top ELLs' Performance With EO Students' Performance

Top ELLs' *p* values	
Higher than EOs'	12 items (27%)
0-5% lower than EOs'	14 items (32%)
15-20% lower than EOs'	11 items (25%)
>20% lower than EOs'	7 items (16%)
Total	44 items

ELLs often spread their answers among the distractors and showed no clear pattern that paralleled either the EOs or the top ELLs, indicating that they may have been guessing. In fact, the majority of their answers are in complete opposition to the other two groups.

Performance characteristics of the top third ELLs. As a group, the top performing ELLs had *p* values of .35 or higher on 29 test items and less than .35 on the other 15 items.[13] All groups (EOs, top third ELLs, and bottom third ELLs) had *p* values below .35 on 6 of those 15 items, indicating that the 6 items were the most difficult items on the test for all students in the sample.

Of the 12 items on which the top ELLs scored best (*p* values ranged from .51 to .84), 6 were basic factual items and 6 were discourse or visual-with-stem prompts that required cognitive analyses. Stems and responses ranged in length from 11 to 28 words. Five of the items contained full-sentence response options and seven contained fragments. The longest stem was 17 words, and the longest single-response option was 16 words. Three items each were from the related social sciences, economics, and history content areas. Two were from sociology and anthropology, and one was from geography.

Of the 15 items that were most difficult for top performing ELLs, 9 were basic factual items that required students to identify the correct response, and 6 were items that required cognitive analyses. The stems and responses for these items ranged in length from 11 to 30 words. Note, however, that ELLs outperformed EOs on items that ranged in length from 14 to 28 words. Length itself was not a problem, but the longer test items tended to be linguistically more complex. For example, the five prompts that were over 15 words in length contained test language such as “which of the following” and “which is the best evidence,” and features such as extended noun phrases, comparative constructions, prepositional phrases, or ellipses. The combination of a higher number of embedded phrases and more propositional elements than in the shorter sentences lends to their linguistic complexity.

Only one item clearly posed a linguistic problem to the top performing ELLs, as well as the other ELLs and the EOs. In order to answer this particular item correctly, students must understand that the linguistic construction of the options is parallel to

[13] The *p* value of .35 was used to divide student performance into categories for analysis because .35 provided a natural break in the data.

the linguistic construction of the prompt. Otherwise students will not be clear about the referent for *it*. Sample Item # 4 provides an illustration.

> Sample Item # 4
>
> How does the size of California compare with the size of Washington state?
> (students look at map provided to them)
>
> a) It is about the same.
> b) It is about half the size.
> *c) It is about twice the size.*
> d) It is about four times the size.
>
> (created by the authors)

If students followed the same pattern they exhibited on the ITBS, between 72.7% and 88.6% of the ELLs and the EOs would split their answers between *b* and *c.* Based on student performance in this study, if the distractors were worded more explicitly, such as "California is about half the size" or "California is about twice the size," students likely would have had fewer problems with the language, especially ELLs whose performance was affected the most.

In terms of topic area, it is notable that one third of the 15 items most difficult for the top performing ELLs were from the economics content area, and three were from political science. Seven of the items focus on the United States, and seven others on recent history since 1900. These are potentially difficult topic areas for California students due to the curriculum sequencing in California.

Thus, three factors suggest that for this group of ELLs, OTL proved to be more problematic than the language of the test. First, the ELLs were able to process the language needed to outperform EOs on 12 of the 44 test items and to perform comparably with EOs on another 14 items. These 26 items varied in length and linguistic complexity. Second, they responded to test items in a pattern similar to that of the EOs, indicating that they could read and understand the test as well as the EOs but may not have had the background knowledge needed to answer correctly. Third, as noted above, the items that proved most difficult for the top ELLs were clustered in a limited number of content areas.

Performance characteristics of the bottom third ELLs. In contrast to the top third ELLs, the bottom third ELLs had *p* values of .35 or higher on only 7 items on the test. Their *p* values ranged from .10 to .15 on 26 items and were less than .10 on 11 items. In general their performance is much different from the other two groups.

The top third ELLs and EOs often exhibited a rationale in choosing their answers, even when they were incorrect. The responses of the bottom third ELLs, on the other hand, were random and contradictory to the patterns exhibited by the other two groups. They sometimes appeared to adopt a strategy of avoiding distractors with unfamiliar vocabulary and looked to the stem for guidance in choosing their answers, choosing answers with vocabulary words similar to those in the stem. They also seemed to alternate between a strategy of either selecting or avoiding the longest and shortest distractors, a strategy that worked to their disadvantage.

Of the seven items on which they performed best, four were basic factual or identify items, and three required cognitive analyses. Three of the seven items were from the history content area, two were from related social sciences, and one each was from sociology and anthropology, and geography. Note that three items related to two topic areas often taught in greater length than other topic areas in the seventh grade (Butler & Stevens, 1997b), and 2 other items related to famous people who are covered in the context of an American holiday. The average sentence length for the seven prompts is 13 words, indicating that students can process longer sentences than appear on LAS Reading if they are not too complex. Only one of these items called for students to process discourse, but the discourse was not complex, consisting of simple past structures and one dependent fronted clause. The other prompts contained enumeration, a prepositional time phrase, and chunks of test language, such as “best describes” and “which of the following.”

The bottom third ELLs had *p* values of less than .10 on 11 items. Note that the top third ELLs had *p* values lower than .10 on only 1 item. These 11 difficult items ranged in length from 6 to 17 words, with an average length of 11.45 words. This contrasts slightly with the top ELLs whose most difficult items had an average length of 13.6 words. Four of the 11 problem items contained specialized content area vocabulary, such as *exchange rate*. All students, including EOs, had difficulty with 3 out of 4 of these items. However, the bottom third ELLs exhibited different behavior than the other two groups, strongly avoiding the correct responses (*p* values ranged from .00 to .08), perhaps as a strategy of avoiding vocabulary they did not recognize.

Regarding topic areas, 4 of the 11 most difficult items were from the economics content area; 4 were from political science, and only 1 each from history, sociology and anthropology, and geography. The performance of the bottom third ELLs corresponds with the top third ELLs, who also had the greatest difficulty with economics and political science, indicating a weakness in these two areas for all ELLs in this study.

Discussion

At the beginning of this report, the following two research questions were presented:

1. How does the language that is measured on a language proficiency test compare with the language that is used on a content assessment?
2. What is the relationship between student performance on these two types of tests? That is, what is the correspondence between ELL English reading proficiency and performance on a content assessment?

In answer to the first question, the comparison of the language measured on the LAS and the language used on the ITBS indicates that the two tests overlap in terms of high-frequency general vocabulary and grammatical structures. However, the language of the ITBS is more complex in vocabulary and sentence structure, aligning its language to the definition of academic language provided at the beginning of this report. The ITBS includes many nonspecialized academic words, such as *represent* and *examine*, and specialized content words, such as *democracy* and *feudalism*. In terms of syntactic complexity, the ITBS contains more embeddings, such as prepositional phrases, and uses a specialized academic register referred to in this report as test language (e.g., expressions such as “which of the following”). Though the use of test language may not actually make an item more difficult, it increases the number of embeddings and sentence length, often contributing to complexity. These differences suggest that while LAS Reading may provide an indication of student ability to function in the mainstream classroom and to perform a limited range of reading tasks, the language it measures is not parallel to the language of standardized content assessments such as the ITBS. For this reason, LAS Reading alone may not be adequate to establish ELL readiness to take content assessments in English.

In answer to the second question, the analyses of the concurrent performance of the ELLs and EOs in our study produced several important findings. First, although as expected, correlational relationships do exist between ELL performance on the LAS and the ITBS, the magnitude of the correlations is weak, explaining only 20% of the variance between total LAS and ITBS. This indicates that other factors account for as much as 80% of the variance between the two tests.

Second, although overall ELL performance on the ITBS was weak, there were statistically significant performance differences between two subgroups of ELLs. Students who had the highest scores on the ITBS had higher mean scores on LAS reading than students who performed less well on the ITBS. However, even though the differences between the mean LAS scores for these two groups were statistically significant, both groups fell into the Limited Reader range. The mean for top third ELLs was at the top of the Limited Reader range, and the mean for the bottom third ELLs was at the bottom of the Limited Reader range.

Third, again as expected, the EOs in the study had higher mean scores than the ELLs as a group on both the LAS and ITBS. However, one group of ELLs performed similarly to the EOs. The ELLs who had the highest scores on the ITBS performed better than the other ELLs in the sample, and their response patterns were similar to the EOs in the study. In fact, response patterns of the top ELLs indicated that they were able to process the language of the test, but lacked the content knowledge to select the correct response. Conversely, the ELLs who did poorly on the ITBS responded to the prompts differently than the highest scoring ELLs and EOs. Thus, for the highest scoring ELLs in this study, a large part of observable performance on the ITBS may have been due to a lack of exposure to the material being tested. For the lowest scoring ELLs, both factors appear to have played a substantial role.

Taken together, the statistically significant correlations between student performance on the two tests and the significant differences in performance for the two ELL subgroups suggest that despite the linguistic limitations of using LAS Reading for the purpose of determining readiness to take content assessments, it does have some predictive ability. That ability, though, seems to be limited to predicting that low scores on the LAS will correspond to low scores on tests such as the ITBS. Conversely, higher scores on the LAS (potentially beginning with the ceiling of the LAS Limited Reader level) may indicate potential for higher scores on the ITBS. Additionally, item response analyses suggest that low student performance on both the LAS and the ITBS Social Studies Test may be due mostly to language, whereas higher concurrent student performance on the LAS and the ITBS may reflect content knowledge. Content test scores for students such as the high group in this study, for example, may be valid since there is evidence that they could process the language of the test and were, in fact, having difficulties similar to the EOs. More research is needed to investigate these preliminary findings.

Finally, the reliability coefficient for ELLs on the ITBS was low compared to the reliability of the norming group. As stated above, the underrepresentation of ELLs in the norming group suggests that the test may not be intended for students whose English language proficiency is weak. Certainly, if the test is to be used with ELL students, larger numbers of those students should be included in future norming studies.

Conclusion

Two key issues raised in this study, the effects of OTL on content test performance and the need for a measure of the academic language used in content assessment, are discussed in this section. Additionally, directions for future research and development are suggested.

Opportunity to Learn

OTL appears to have played an important role for all students in this study. The EOs and top third ELLs could process some of the most linguistically difficult items on the test and choose the correct answer. However, they were unable to choose the correct answer on less complex items on different topics, indicating that OTL was a likely problem. In particular, as noted above, there were weaknesses for all students in the political science content area. EOs did poorly on 3 out of 4 political science items, and the top ELLs had difficulty with all four. ELLs also encountered problems with items from an economics subsection, scoring poorly on 3 of the 4 items in the subsection. EOs, on the other hand, had particular difficulty with items from a History subsection, performing poorly on 4 of the 5 items in the subsection.

Neither ELLs nor EOs seemed to have difficulty with skill-based items—those items that involved the use of visual and higher order cognitive skills—with one exception, a political cartoon. The ability to interpret this cartoon depended on knowledge of American government. EOs may have had a better chance of selecting the correct response because they were more likely to have background knowledge about taxation and the structure of American government. They may have been exposed to discussions on these topics at home. Without such exposure or classroom instruction, students are likely to interpret political cartoons or other pieces of social commentary literally, as they did in this study, which would lead to misinterpretation.

OTL may have been an issue for all of the students in the study since the ITBS includes items from topic areas and time periods that California seventh graders would not yet have been exposed to. For example, the seventh-grade ITBS Social Studies Test includes items on American history that are covered in the eighth grade in California. The general topic areas covered at each grade level in California, from Grades 4 through 8, are shown in Table 9.

Additionally, earlier research on topic coverage in the seventh grade has revealed that there is tremendous variance in the length of time teachers spend on social studies topics (Butler & Stevens, 1997b). Teachers were surveyed at seven year-round schools in the Los Angeles Unified School District. They reported spending the most amount of time on medieval studies; however, note that although the average time spent was 4 weeks, there was a wide range in amount of time spent, from 1 to 7 weeks. This variance in the amount of time spent, combined with the slower pace of study many ELLs may have encountered in sheltered and ESL classes in earlier grades, may contribute to deficits in background knowledge. Since standardized assessments are often cumulative in nature, students who are struggling to learn English and content simultaneously are at a disadvantage because language (Collier, 1987) and concepts become increasingly more complex from one grade level to the next.

Other research and observations in the social studies content area (Butler, Stevens, & Castellon-Wellington, 1999), as well as interviews with both the teachers and students in this study, indicate that many students, ELLs and EOs alike, do not receive the entire California social studies curriculum as outlined in the chart above. Teachers cite difficulties using the state-mandated textbooks with both ELLs and EOs because they are linguistically too complex. They often feel the need to

Table 9

California Social Studies Curriculum

Grade	General topics Grades 4 through 8
4	California history
5	United States history and geography (to 1850)
6	World history and geography (Ancient Civilizations)
7	World history and geography (the Fall of Rome to the Enlightenment)
8	United States history and geography (the Constitutional Convention to the 1900s)

Note. Source: History-Social Science Curriculum Framework and Criteria Committee, 1997.

condense and simplify material in the texts and frequently use supplementary materials instead. Some teachers add a unit on geography at the beginning of the year because students are often weak in this critical area. One teacher of an Advanced Placement social studies class said that she “struggles to get through the curriculum” because it is so broad (C. Anderson, personal communication, March 10, 1997). On a positive note, teachers in this study reported the frequent use of visuals in class, which may help build the skills needed to do well on some of the items on the ITBS.

Finally, students in this study expressed frustration with the content of the ITBS during the informal posttest focus groups, saying that they had not studied most of the material. They commented that they wished they could have prepared for the test. Additionally, when they asked questions during the administration of the test, the questions were usually about vocabulary words, which indicate that vocabulary was a major issue for these students.

Academic Language Proficiency

Understanding what part of student performance is related to academic language and what part is related to content knowledge requires the use of a language assessment that measures the academic language used on content assessments. According to the analyses above, LAS Reading provides a measure of high-frequency general vocabulary, the ability to scan a passage for information, and the ability to recognize basic punctuation, capitalization, and discrete grammar points. It does not measure the type of language that corresponds to the language students are required to process when taking tests like the ITBS. Since the purpose of the LAS is primarily to identify students who are ready to participate in mainstream English-only classrooms, there are implied limitations in the uses of the LAS for other purposes. LAS Reading provides a picture of an ELL’s basic language processing skills. It is not designed to measure academic language proficiency. Indeed, few commercially available language tests, if any, are designed to measure academic language as defined in the beginning of this report. That is, the language they measure provides little information about the extent to which students are capable of processing the more complex language found in achievement tests or other contexts in which academic language is present. Therefore, a language test created for the purpose of determining readiness for standardized content assessments is needed, which will in turn help to assure the validity of those assessments with ELLs.

Recommendations for Research and Development

Our recommendations are threefold. First, based on the finding that the LAS does not measure the academic language found on standardized tests, we recommend research and development on a language measure for that purpose. This has been recommended elsewhere (Butler & Castellon-Wellington, 2000), and earlier work was begun at CRESST towards operationalizing academic language proficiency and developing task-based prototype items in social studies (Butler, Stevens, & Castellon-Wellington, 1999).

Second, we recommend research that controls for the impact of opportunity to learn on ELLs' performance on standardized content assessments. Here, and in other recent research (Abedi, Leon, & Mirocha, 2000; Butler & Castellon-Wellington, 2000; Castellon-Wellington, 1999), a relationship between language proficiency and performance on such assessments has been confirmed. However, while language is an important factor, it does not explain all of the variation in performance. In this study, through interviews with the teachers, informal focus groups with the students, and careful examination of item response patterns, we found that opportunity to learn played a critical role. This confounded the research, making it difficult to determine how much of a role language and OTL played in ELLs' overall performance. Thus, research is needed that controls for content and highlights the impact of language.

Third, replication of this research is recommended since findings are preliminary and have limitations noted earlier in the report. Perhaps the most important finding in this research is that subgroups of ELLs who are aggregated as LEP students can perform statistically differently from each other, a trend that may be muted in large-scale research in which little attention is focused on item-level analyses for subgroups of students. Within-group differences in this study suggest that it may be inappropriate to aggregate all ELLs into one group for research and analyses, even when they are all classified as LEP. The investigation of ELL subgroup performance across content areas is needed to identify the point at which performance on a content assessment is meaningful and to prove that observed language difficulties are more a function of opportunity to learn in the content area (e.g., specialized content area words and structures that all students must master) than overall language processing difficulties.

Final Remarks

This study sheds light on several important issues. The first is that emphasis on aggregating all ELLs for research and intervention is misleading and inappropriate due to the diversity within ELLs as a group. In this study, disaggregating ELLs according to their performance on the ITBS and performing item-response analyses revealed that suspected differences in the performance of the two ELL groups is real. Poor ELL performance on standardized content assessments is often attributed to language difficulties. However, the weak performance of even the highest scoring ELLs in the study is not primarily due to language difficulties; it may, in fact, be a result of limitations in OTL or other academic factors. This potentially impacts how interventions for these students should be selected. For example, for the highest scoring ELLs in this study, providing test accommodations, an approach currently being used to mitigate potential language problems on standardized assessments (see Abedi, Hofstetter, Baker, & Lord, 1998; Castellon-Wellington, 1999), would not be helpful since it is evident that they could already process the language of the assessment but simply did not have the content knowledge necessary to do well. Test scores for these students and others like them may actually reflect true gaps in knowledge, and thus their performance on content assessments may be valid indicators of their content knowledge.

Second, most existing commercially developed language assessments are not appropriate measures of academic language. Therefore, they are not appropriate for assessing readiness for taking standardized assessments in English. Using them for this purpose would be a misuse of the tests, potentially leading to incorrect assumptions or decisions about ELLs. If students are expected to take standardized assessments when they are not ready linguistically, the resulting test scores will not be valid. An appropriate measure is needed that will help determine to what degree language demands interfere with content performance and to establish test-taker readiness.

Third, due to the low test reliability on the ITBS for the ELL group as a whole and the chance performance levels of the lowest scoring ELL group, further research is needed to establish the appropriateness of standardized assessments such as the ITBS Social Studies Test for ELLs. This may be accomplished partially by including more ELLs in future norming studies.

Clearly, with standardized tests becoming such an integral part of high stakes educational decisions at both local and state levels, a way must be found to meet the needs of ELL students more effectively. Without valid and reliable assessment data that indicate *which* ELLs are learning *what* in our nation’s classrooms, erroneous conclusions about ELL performance may be drawn. Additional research that helps identify the criterion level of language proficiency needed to take standardized assessments in English is a necessary step toward the valid and reliable evaluation of ELL progress in school and the evaluation of the programs that serve them.

References

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (1998). *NAEP math performance and test accommodations: Interactions with student language background* (Draft Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Leon, S. (1999, December). *Impact of students' language background on content-based performance: Analyses of extant data* (Final Deliverable to OERI, Contract No. R305B60002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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 *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002; pp. 3-49). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Armento, B. J., Nash, G. B., Salter, C. L., & Wixson, K. K. (1991). *Across the centuries* (Teacher’s ed.). Boston: Houghton Mifflin Company.

Bailey, A. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002; pp. 85-105). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bernstein, B. (1961). Social structure, language and learning. *Educational Research, 3*, 163-76.

Bloom, B. S., Engleheart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives.* New York: David McKay.

Butler, F. A., & Castellon-Wellington, M. (2000). Students’ concurrent performance on tests of English language proficiency and academic achievement. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002; pp. 51-83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Butler, F. A., & Stevens, R. (1997b). *In-house summary report, history/social science topics teacher questionnaire* (Summary Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Butler, F. A., Stevens, R., & Castellon-Wellington, M. (1999). *Academic language proficiency task development process* (Final Deliverable to OERI, Contract No. R305B60002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Castellon-Wellington, M. (1999). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests* (CSE Tech. Rep. No. 524). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book: An ESL/EFL teacher's course*. Boston: Heinle & Heinle.

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

Collier, V. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly, 21*, 617-641.

Criscoe, B. L., & Gee, T. C. (1984). *Content reading: A diagnostic/prescriptive approach.* Englewood Cliffs, NJ: Prentice-Hall.

Cummins, J. (1980). The construct of proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics* (pp. 81-103). Washington, DC: Georgetown University Press.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. San Diego, CA: College-Hill Press.

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

Daines, D. (1982). *Reading in the content areas: Strategies for teachers*. Glenview, IL: Scott, Foresman.

Davidson, F. (1994). Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. *Language Testing, 11,* 83-95.

Duncan, S. E., & De Avila, E. A. (1988a). *Language assessment scales (LAS®) reading and writing examiner's manual, Forms 1A and B, Forms 2A and B, and Forms 3A and B.* Monterey, CA: CTB / McGraw-Hill.

Duncan, S. E., & De Avila, E. A. (1988b). *Language assessment scales (LAS®) reading and writing technical report: Validity and reliability, Forms 1, 2, and 3*. Monterey, CA: CTB / McGraw-Hill.

Duncan, S. E., & De Avila, E. A. (1990). *Language assessment scales (LAS®) reading component, Forms 1A, 2A, and 3A*. Monterey, CA: CTB / McGraw-Hill.

Halliday, M. A. K. (1975). *Learning how to mean: Explorations in the development of language*. London: Edward Arnold.

Hamayan, E. V., & Perlman, R. (1990). *Helping language minority students after they exit from bilingual/ESL programs: A handbook for teachers*. Washington, DC: National Clearinghouse for Bilingual Education.

Hatch, E., & Brown, C. (1995). *Vocabulary, semantics, and language education*. Cambridge: Cambridge University Press.

Heath, S. B. (1983). *Ways with words*. Cambridge: Cambridge University Press.

History-Social Science Curriculum Framework and Criteria Committee. (1997). *History-social science framework for California public schools*. Sacramento, CA: California Department of Education.

Kinsella, K. (1997). Moving from comprehensible input to "learning to learn" in content-based instruction. In M. A. Snow & D. M. Britton (Eds.), *Perspectives on integrating language and content* (pp. 46-68). White Plains, NY: Addison-Wesley Longman.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*, 237-270.

Nagy, W. E., & Scott, J. A. (1990). Word schemas: expectations about the form and meaning of new words. *Cognition and Instruction, 7*, 105-127.

O'Malley, J. M. (1992). Looking for academic language proficiency. In Office of Bilingual Education and Minority Languages Affairs (Ed.), *Proceedings of the second research symposium on limited English proficient students' issues* (pp. 173-182). Washington, DC: U.S. Government Printing Office.

Philips, S. U. (1972). Participant structures and communicative competence: Warm Springs children in community and classroom. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 370-394). New York: Teachers College Press.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation* (3rd ed.). Belmont, CA: Wadsworth.

Scarcella, R., & Zimmerman, C. (1998). Academic words and gender: ESL student performance on a test of academic lexicon. *Studies in Second Language Acquisition, 20*, 27-49.

Short, D. (1994,). Study examines the role of academic language in social studies content-ESL classes. *Forum, 17*(3). (National Clearinghouse for Bilingual

Education Newletter). Retrieved December 7, 2001, from www.ncbe.gwu.edu/ncbepubs/forum/1703.htm

Solomon, J., & Rhodes, N. (1995). *Conceptualizing academic language* (Research Rep. No. 15). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning.

Teachers of English to Speakers of Other Languages, Inc. (1997). *ESL standards for pre-K-12 students.* Alexandria, VA: Author.

Ulibarri, D. M., Spencer, M. L., & Rivas, G. A. (1981). Language proficiency and academic achievement: a study of language proficiency tests and their relationship to school ratings as predictors of academic achievement. *NABE Journal, 5,* 47-80.

University of Iowa. (1993a). *Iowa Tests of Basic Skills® complete battery, Level 13, Form L.* Chicago, IL: Riverside Publishing.

University of Iowa. (1993b). *Iowa Tests of Basic Skills ® norms and score conversions, Form L, complete and core batteries.* Chicago, IL: Riverside Publishing.

University of Iowa. (1996a). *Iowa Tests of Basic Skills® with integrated writing skills test. Practice tests, Levels 9-11, Form M.* Chicago, IL: Riverside Publishing Company.

University of Iowa. (1996b). *Iowa Tests of Basic Skills® with integrated writing skills test. Practice tests, Levels 12-14, Form M.* Chicago, IL: Riverside.

University of Iowa and The Riverside Publishing Company. (1994). *Integrated assessment program, technical summary I Riverside 2000.* Chicago, IL: Riverside Publishing.

Wilkins, D. A. (1976). *Notional syllabuses.* Oxford: Oxford University Press.