CRESST REPORT 738

Mikyung Kim Wolf
Joan L. Herman
Jinok Kim
Jamal Abedi
Seth Leon
Noelle Griffin
Patina L. Bachman
Sandy M. Chang
Tim Farnsworth
Hyekyung Jung
Julie Nollner
Hye Won Shin

PROVIDING VALIDITY EVIDENCE TO IMPROVE THE ASSESSMENT OF ENGLISH LANGUAGE LEARNERS

AUGUST, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Providing Validity Evidence to Improve the Assessment of English Language Learners

CRESST Report 738

Mikyung Kim Wolf, Joan L. Herman, Jinok Kim, Jamal Abedi,
Seth Leon, Noelle Griffin, Patina L. Bachman, Sandy M. Chang,
Tim Farnsworth, Hyekyung Jung, Julie Nollner, and Hye Won Shin
CRESST/University of California, Los Angeles

August 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the National Research and Development Centers or the U.S. Department of Education's Institute of Education Sciences (IES).

TABLE OF CONTENTS

PROVIDING VALIDITY EVIDENCE TO IMPROVE THE ASSESSMENT OF ENGLISH LANGUAGE LEARNERS[1]

Mikyung Kim Wolf, Joan L. Herman, Jinok Kim, Jamal Abedi,
Seth Leon, Noelle Griffin, Patina L. Bachman, Sandy M. Chang,
Tim Farnsworth, Hyekyung Jung, Julie Nollner, & Hye Won Shin
CRESST/University of California, Los Angeles

Abstract

This research project addresses the validity of assessments used to measure the performance of English language learners (ELLs), such as those mandated by the No Child Left Behind Act of 2001 (NCLB, 2002). The goals of the research are to help educators understand and improve ELL performance by investigating the validity of their current assessments, and to provide states with much needed guidance to improve the validity of their English language proficiency (ELP) and academic achievement assessments for ELL students. The research has three phases. In the first phase, the researchers analyze existing data and documents to understand the nature and validity of states' current practices and their priority needs. This first phase is exploratory in that the researchers identify key validity issues by examining the existing data and formulate research areas where further investigation is needed for the second phase. In the second phase of the research, the researchers will deepen their analysis of the areas identified from Phase I findings. In the third phase of the research, the researchers will develop specific guidelines on which states may base their ELL assessment policy and practice. The present report focuses on the researchers' Phase I research activities and results. The report also discusses preliminary implications and recommendations for improving ELL assessment systems.

[1] We would like to thank Lyle Bachman, Alison Bailey, Frances Butler, Diane August, and Guillermo Solano-Flores for their valuable comments on earlier drafts of this report. We are also very grateful to our three participating states for their willingness to share their data and support of our work.

INTRODUCTION

NCLB (2002) declares that states, districts, schools, and teachers must hold the same high standards for ELL students as they do for all other students, and that they are accountable for assuring that all students, including ELL students, meet high expectations. By mandating that ELL students be included in annual state assessments, subjected to annual assessments of English language development, and included in Adequate Yearly Progress (AYP) performance targets, the Federal legislation operationalizes attention to the needs and progress of ELL students in English proficiency and in school subject matters. The assessments serve as both levers and instruments of reform; as part of the accountability system, they establish goals and incentives for improvement, and as data, they provide educators with information for assessing the success of their programs, determining student needs, and planning subsequent instruction. It is axiomatic that both functions require reliable and valid measurement of the performance of ELL students. This research project thus aims to addresses ways to improve the validity of inferences drawn from assessments used to measure the performance of ELL students.

As summarized in a recent report by the U.S. Government Accountability Office (GAO, 2006), many states have moved ahead rapidly to develop needed ELL assessments. States have developed or adopted new measures of ELP, and they have adopted accommodation strategies for assessing ELL students' achievement of academic standards. Accommodation strategies are intended to reduce the confounding influence of ELL students' English proficiency on their performance on state test of academic content (e.g., mathematics). However, there are numerous technical challenges to the accurate assessment of ELL students (see GAO, 2006). The technical challenges are complicated by the diverse language, and cultural and demographic backgrounds reflected in this country's growing ELL population. Difficulties also exist in disentangling measures of academic content knowledge that are administered in English, such as math and science, from students' English language proficiency.

The rush to meet NCLB assessment requirements has left states without the expertise, time, or resources to systematically document or address fundamental, underlying validity issues (GAO, 2006). Although the validity of the assessment of ELL students has been a topic of research and expert recommendation (i.e., the Standards for Educational and Psychological Testing; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999), the actual conduct of validity studies for ELL students has been limited.

As a result, the decisions made on the basis of assessments may not be warranted. Not only is the validity of measures of ELP and content performance uncertain, other critical aspects of the ELL assessment process also need attention, such as the initial designation of students as ELL and the redesignation of the students upon achieving English language proficiency.

Recent test[2] results of states with large percentages of ELL student populations provide dramatic examples of the urgent need to better understand and improve ELL students' performance. For example, on the basis of test scores in mathematics in the 2003–2004 school year across 48 states, the GAO reported that ELL students' math proficiency level averaged 20% lower than the overall population (GAO, 2006). For the 2005 National Assessment of Educational Progress (NAEP) in mathematics, 46% of Grade 4 ELL students scored Below Basic as compared to 18% of non-ELL students. In Grade 8, 71% of ELL students still scored Below Basic as compared to 30% of non-ELL students (Perie, Grigg, & Dion, 2005). What are the sources of these significant gaps? What role is played by the validity of the assessments for ELL students—for example, do the language demands of content assessments in English underestimate ELL students' accomplishments in subject matter fields? What role might language proficiency play in effective access to instruction and content assessment? What role do other background variables play? What is the role of ELL students' opportunities to learn the knowledge and skills measured on assessments? Effective policies and practices for reducing the gap in ELL learning first require the use of valid measures of ELL students' achievement, including both their English proficiency and academic content proficiency. Accurate assessment must undergird any credible analyses of the complex relationships between English proficiency, academic achievement, redesignation criteria, opportunity to learn (OTL), and academic learning, which are essential in understanding and improving ELL students' academic success. Unless these validity limitations of current ELL assessment practices are addressed, researchers' ability to trust and make decisions based on the results of ELL students' performance is sharply reduced.

General Research Goals

This research project aims to help educators understand and improve ELL students' performance by investigating the validity of their current assessments and to provide states with much needed guidance to improve the validity of their ELP and academic achievement assessments, particularly in math and science, for ELL students. The project entails three phases in achieving its research goals. In the first phase, in partnership with collaborating

[2] The terms test and assessment are used interchangeably throughout this report as both terms are frequently used with the same meaning in practice.

states, the researchers attempt to use existing documents and analyze available data to understand the nature and validity of states' current practices and their priority needs. In the second phase of the study, the researchers will deepen the analysis of the areas where further investigation is needed, in collaboration with one or two states. The specific nature of this phase of the project is being informed by the first phase results and formative feedback from both other experts in the field, as well as state stakeholders. Drawn from the findings of the first and second phases, the third phase of the study will aim to develop specific guidelines on which states may base their ELL assessment policy and practice. The present report focuses on the researchers' Phase I project activities and illustrates the researchers' preliminary guidelines to improve the valid use of ELL assessments.

Participating States and Data

Three states participated in the first phase of the project (These states will be referred to as State A, State B, and State C throughout the report). In order to accomplish the research goals for Phase I, the research team requested state assessment data sets from two AYP reporting grade levels (e.g., Grades 4 and 8) from each state. The variables that the researchers requested included students' demographic background (e.g., ethnicity, socio-economic status, mobility, home language), state test scores in reading, mathematics, and science at the student and item levels, ELP test scores at the student and item levels, accommodation types used at the student level, and students' ELL status (e.g., level of ELP and redesignation status). Although the content areas of primary interest for this project are math and science, students' scores in reading/English language arts also were requested as an additional variable to control for students' language ability. The research team also requested individual-level longitudinal data (2 to 3 years) in order to examine the progress of ELL students' performance over time. In addition, the researchers attempted to obtain actual test items to investigate the language demands and academic language characteristics included on those items.

It should be noted that different sets of data were obtained from each state due to the different assessments and data management systems across states. Specific available data analyzed in each aspect of Phase I will be described in later chapters of this report.

Research Foci in Phase I

Among a number of validity issues in assessing ELL students, the researchers' Phase I analyses focused on the following six areas:

1. the language demands exhibited on state content-area and ELP assessments,
2. the identification of items that function differentially for ELL subgroups and the characteristic of those items,
3. achievement gaps among the subgroups of ELL students (e.g., ELLs, redesignated ELLs) compared to non-ELL students,
4. the relationship between the ELP and content-area assessment scores,
5. the factors related to the redesignation status, and
6. accommodation practices.

The Phase I research in these areas consists of several different analytical and methodological approaches, which are addressed separately in each of the following chapters. The first chapter presents the results from the content analysis of assessment items in terms of their language demands. The second chapter summarizes the results from the differential item functioning (DIF) analysis and describes the characteristics of DIF items. The third chapter contains the analyses of the students' performance on the content-area and ELP assessments, performance gaps, redesignation decisions, and available data on accommodation practices (the research areas [3] – [6] mentioned earlier). Each chapter includes a set of specific research questions, and the related studies and background information about the addressed research questions in detail. Each chapter also describes the methods utilized for the specific research questions. The researchers focus on presenting the results in such a way that readers may understand the common trends and issues surrounding ELL assessment across the three states. Thus, Chapters 2 and 3, which involve rigorous statistical analyses, focus on describing and discussing their findings. Only a couple of statistical results are presented in tables and figures as examples.[3] The last chapter concludes this report with the integration of the Phase I findings and implications for improving the validity of the states' assessment systems.

[3] Detailed statistical results are available by request.

References

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

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: U.S. Department of Education, National Center for Education Statistics, U.S. Government Printing Office.

U.S. Government Accountability Office. (2006) No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency (GAO-06-815). Washington, DC: Retrieved July 20, 2006, from: www.gao.gov

CHAPTER 1:

AN INVESTIGATION OF THE LANGUAGE DEMANDS IN STATES' CONTENT-AREA AND ENGLISH LANGUAGE PROFICIENCY TESTS

Mikyung Kim Wolf, Sandy M. Chang, Hyekyung Jung, Tim Farnsworth,
Patina L. Bachman, Julie Nollner, and Hye Won Shin
CRESST/University of California, Los Angeles

The No Child Left Behind Act of 2001 (NCLB, 2002) has significantly influenced states' polices in assessing English Language Learner (ELL) students. States have moved rapidly to identify or develop appropriate assessments to account for ELL students' academic achievement in content areas (e.g., reading/English language-arts, mathematics, and science) as well as their English language proficiency (ELP), and to provide evidence that their assessment and accountability systems are valid (see Peer Review Guidance, U.S. Department of Education [DOE], 2004, for details of evidence to be collected). Of various sources of validity evidence, the nature of the language used in the assessments is fundamental in that it sheds light on the construct being measured. Further, understanding the language demands imposed by assessment is crucial to make appropriate inferences about test results.

In assessing ELL students' content knowledge and skills, their facility with the language of the assessment (English) confounds students' ability to show what they know and understand. For example, when a math test is administered in English, test results may be a function of ELL students' ability to understand the language of the test as well as their mathematics attainment. A serious validity concern is raised if test items contain unnecessary linguistic complexity that interferes with ELL students' ability to show their content knowledge. Measurement experts term variation in test performance that is caused by factors unrelated to the construct being measured "construct-irrelevant variance." In order to reduce the construct-irrelevant variance that the language of assessment may cause, it is important to examine the characteristics or demands of the language of the test items with which ELL students need to cope.

Understanding the nature of language in ELP tests is even more critical in order to make appropriate inferences from the test results. Much research has identified the need for ELP tests to assess ELL students' academic English proficiency (Bailey & Butler, 2003; Bailey, Butler, Stevens, & Lord, 2007; Stevens, Butler, & Castellon-Wellington, 2000).

Researchers assert that traditional ELP tests tend to focus primarily on social language, with little attention to the academic language skills the ELL students need for mainstream classroom readiness. The limitations of traditional ELP tests, coupled with NCLB requirements to create ELP tests that are aligned with states' ELP and academic standards, have led states to develop or adopt new ELP tests. Yet there has been little research to date on whether and how the newly developed ELP tests address previous limitations, including construct validity issues.

Thus the current study investigates the language characteristics that are used in items on states' standardized content-area and recently-developed ELP tests. Particularly, the study compares the language demands and the characteristics of academic language across states' math, science, and ELP tests.

Specifically, the researchers' analysis of the linguistic demands of test items addresses the following research questions:

1. What characteristics of academic English language are present in states' standardized math, science, and ELP tests?
2. To what extent do the characteristics of academic English language on the content-area tests correspond with those of the ELP tests?

The results from the content analysis reported here provide information about the nature of the test constructs and the extent to which the tests represent academic language demands. The findings provide a better understanding of the inferences made from the assessments. The analysis also suggests a systematic tool that practitioners can use to investigate the academic language characteristics of test items. As part of this work, the research team applied an evolving methodology drawn from previous research and refined it for analyzing the academic language characteristics of tests. The team hopes that the methodology proves useful for others who will also be analyzing tests for academic language characteristics as part of their validity evidence. The current analysis also provides a base for further investigation of the language demands placed on ELL students, for instance, examining the extent to which the academic language characteristics are related to students' performance on the tests.

Relevant Literature

In this section, the researchers focus their review of the literature on two areas: first, efforts to articulate the construct of academic English language from empirical bases and second, efforts to establish the importance of assessing the academic English proficiency of

ELL students with appropriate language tests. Previous work in these two areas provides points of departure for the work reported here.

Defining Academic English Language

Defining academic English and investigating its characteristics is an evolving research area. On the surface, the construct of academic English seems straightforward. Following Chamot and O'Malley (1994), it is "the language that is used by teachers and students for the purpose of acquiring new knowledge and skills … imparting new information, describing abstract ideas, and developing students' conceptual understanding (p. 40)." However, for purposes of curriculum development, classroom teaching, and assessment, specificity is needed with regard to actual language use in a wide range of classroom settings across content areas to operationalize the construct. That is, curriculum developers, teachers, and language testers must be able to articulate the construct in specific details of vocabulary, grammar, and discourse. Over the years, a number of scholars have been looking at academic English from a variety of perspectives (Jordan, 1997; Flowerdew & Peacock, 2001; Douglas, 2000), and while a coherent picture of the construct across content areas and grade levels is not yet available, it is helpful to consider some of the approaches that have been taken.

Cummins (1981) in seminal work in this area distinguished academic language from social language, naming the former as Cognitive Academic Language Proficiency (CALP) and the latter as Basic Interpersonal Communicative Skills (BICS). In considering the cognitive and contextual demands of language use, he argued that the context-reduced communication of CALP often involves a high degree of lexical variety and syntactic sophistication. From this, CALP is considered as language that is more cognitively complex, has a reduced need for contextual cues, and most likely represents language learned in academic settings. Cummins' work has largely been interpreted as a dichotomous approach to language use that juxtaposes social language and academic language as being typically specific to different language use environments—one outside of school; the other school settings. As Cummins (2000) later noted, however, these two language types should not be viewed as dichotomous and separately acquired, but rather as simultaneous in their learning and acquisition with the distinction being in the degree of cognitive and contextual demands. Both are important in the development of a student's overall language proficiency.

Chamot and O'Malley (1994), Short (1993, 1994), and more recently Bailey, Butler, LaFramenta, and Ong (2001/2004) suggested a task-oriented approach to academic language that focused on language functions and in doing so helped capture very specific types of language use that cut across content areas. Examples of academic language functions include *analyzing, comparing, predicting, persuading, solving problems*, and *evaluating*. The notion of language functions—which can require specific lexical, grammatical, and discourse patterns—provides a possible organizing paradigm for addressing academic English across content areas.

Another approach to defining academic English is to consider the linguistic elements that make up the register of schooling. Schleppegrell (2001) provided such an analysis of school-based texts and labeled academic English as the "language of schooling." For example, Schleppegrell described the lexical choices of school-based texts as specific and technical rather than generic. With respect to grammatical features, she maintained that central grammatical features of school-based text, such as the use of lexical subjects and nominalizations, enable the language of schooling to present information in highly structured ways.

Recent efforts have encompassed multiple perspectives on academic English and have suggested frameworks to operationalize the characteristics in teaching and learning academic English. Scarcella (2003) proposed that academic language involves multiple linguistic, cognitive, and sociocultural/psychological dimensions comprised of integrated components. For instance, she describes discrete linguistic features (phonological, lexical, and grammatical components), language functions (sociolinguistic component), and stylistic register (discourse component) of academic English. As a basis of the framework, she adopted the "communicative competence" model proposed by Canale and Swain (1980) and Bachman (1990), which includes grammatical competence, sociolinguistic competence, discourse competence, and strategic competence. Scarcella's framework is a broad competence-based approach which includes the language functions and related discrete linguistic features discussed earlier. Based on a series of studies to investigate the nature of language in academic texts, tests, and discourse, Butler, Bailey, and their colleagues (Bailey & Butler, 2003; Bailey et al., 2001/2004; Butler, Bailey, Stevens, Huang, & Lord, 2004) also proposed a framework to characterize academic English at the lexical, grammatical, and discourse levels, as well as in language functions. An integration of these recent perspectives represents the approach utilized in the present study.

Importance of Assessing Academic English Language

As mentioned earlier, language testing experts and language researchers have highlighted the inadequacy of many traditional ELP tests in tapping the development of the academic English language skills students need to be successful in school settings (Stevens et al., 2000; Butler & Castellon-Wellington, 2000/2005). For example, Stevens et al. (2000)

examined the relationship between language proficiency and student performance on standardized content-area tests, comparing the type of language assessed on language tests and the language used on content tests. They examined the language in the Language Assessment Scales (LAS; De Avila & Duncan, 1990) and a standardized content-area test in social studies for the Grade 7. They found that the language assessment test had more general language, whereas the content test had more academic language characteristics. In the LAS, syntax was less complex, vocabulary consisted generally of everyday words, and discourse was less demanding to process. On the other hand, the content-area test had academically more demanding language including academic vocabulary, various syntactic structures, and more specific linguistic registers. Based on their findings on the mismatch of language between the two tests, the researchers argued that there is a need for the development of language proficiency assessments that measure students' academic language proficiency. These limitations in traditional language tests point to the importance of accurately assessing academic English language skills in order to determine whether students who are identified as competent are, in fact, proficient enough to succeed in an academic setting.

The goal of the work reported here is to examine the academic language characteristics of math and science tests that states utilize and compare those results to the academic language characteristics of a new ELP test at the same grade level. The comparison will help determine if the newer generation of language proficiency tests captures the type of language being used in the content tests in the study. The following section provides a description of the methodology used to address the issue.

Methods

This section describes the data sources, the instrument used to analyze the data, and the procedures implemented to analyze the data.

Data

For the purpose of the content analysis of test items, the accountability instruments from two states—referred to hereafter as State A and State B—were obtained. State A provided one of the multiple forms of its standardized math and science tests from Grades 5 and 8 for the school year 2005–2006[4] (see Table 1). State A requested that the research team analyze the common items across the multiple forms since only those items were counted toward a student's total score. The common items comprised about 72% of each form of the

[4] State A was unable to provide its ELP test due to a state-specific confidentiality issue.

tests. State B provided multiple forms of its standardized math tests from Grades 4, 7, and 8[5] (see Table 2). Due to time constraints, the research team only analyzed one form for Grades 4 and 7, and two forms for Grade 8, which included some common items across all forms of the tests. State B also provided their ELP test from Grade bands 3–5 and 6–8 that the state used for the same academic year (see Table 3). The ELP test that State B provided is newly developed to be in compliance with NCLB. The ELP test developers explicitly mention in their blueprint that the test constructs include both social and academic language that students need for school settings. The following tables summarize the compositions of the assessments analyzed.

Table 1

The Composition of State A Content Tests

	Math	Science
Grades	5, 8	5, 8
Forms	8 forms	8 forms
Topics covered (from standards)	Numbers, number sense, & computation; patterns, functions, & algebra; measurement; spatial relationships, & geometry; data analysis	Science inquiry; science, technology, & society; atmospheric process & water cycle; solar system & universe; earth's composition & structure; matter; forces, & motion; energy
Number of items	46 (Grade 5); 51 (Grade 8)	44 (Grade 5); 42 (Grade 8)
Format	Multiple-choice, constructed response[a]	Multiple-choice, constructed response[a]

[a]Although the tests included constructed response items, only multiple-choice items were analyzed to have comparable data with the other state tests.

Table 2

The Composition of State B Content Tests

	Math
Grades	4, 7, 8
Forms	Grade 4, 6 forms; Grade 7, 6 forms; Grade 8, 4 forms
Topics covered (from standards)	Numbers & operations; measurement; geometry; data analysis & probability; algebra
Number of items	50 (Grade 4); 50 (Grade 7); 60 (Grade 8)
Format	Multiple-choice

[5] State B had not implemented its standardized science test at the time of the researchers' request for data and was thus not available to us.

Table 3

The Composition of State B ELP Test

	Listening	Speaking	Reading	Writing
Grades	3–5, 6–8	3–5, 6–8	3–5, 6–8	3–5, 6–8
Forms[a]	2 forms	2 forms	2 forms	2 forms
Sections	Classroom directions, mathematical language, classroom dialogue, academic lectures	Vocabulary, interpreting/ describing scenes, mathematical language, graph interpretation, express opinions	Vocabulary, graph interpretation, academic texts	Conventions, grammar, descriptive and narrative writing
Number of items	24 (Grades 3–5, 6–8)	19 (Grades 3–5, 6–8)	26 (Grades 3–5, 6–8)	19 (Grades 3–5, 6–8)
Format	Multiple-choice	Constructed response	Multiple-choice	Multiple-choice, constructed response

[a] Although the test publisher has two forms of this test, the state only administered one form to each grade.

Content Analysis Protocol for Language Demands

Based on previous research, a content analysis protocol and rater guidelines were developed in order to characterize the nature of language on the tests (see Appendix CH1 [pp. 47–53] at the end of this chapter for an abridged version of the protocol).[6] The content analysis protocol was mainly adapted from two sources: An academic English framework developed by CRESST researchers (Butler et al., 2004; Bailey et al., 2007) and a rating instrument to analyze language proficiency test items developed by Bachman and his colleagues (Bachman & Carr, 1999; Bachman, Davidson, Ryan, & Choi, 1995). These two sources, which provided a comprehensive foundation for the researchers' current work, dealt with various linguistic dimensions in the analysis of test items and texts.

Note that although the researchers' content analysis protocol was drawn from these previous works, considerable modifications were made in defining and operationalizing each category for the purpose of analyzing the specific content-area and ELP tests. Butler and her colleagues' (2004) work focused on extended textbook analysis, which had larger quantities of text to analyze than the tests the current research examined. The research team modified the categories to apply to a much smaller amount of language on the test items and refined the descriptions of the types of academic vocabulary to reflect the different language

[6] This summary version of the protocol includes the definition and one or two simple examples to describe each category. The unabridged content analysis protocol, which includes definitions, extensive examples, and coding or scoring procedures, is too long to be included in this report. The complete protocol is available by request.

demands that ELL students might experience. Bachman and Carr’s (1999) instrument was intended to rate the characteristics of each language feature (e.g., grammar, vocabulary) on score-scales to quantify the language demands of tests and to be utilized to assist the test development process. For example, their instrument includes a rating for cultural and topical knowledge embedded in the passages and items of a language test. For the team’s purposes of examining the linguistic complexity of test items, the researchers undertook Butler et al.’s (2004) method of counting all the observations of academic language features. The research team adapted some of the score scales from Bachman and Carr’s instrument to holistically rate an item for its language demands. Table 4 presents the categories included in the researchers’ content analysis protocol. Although the majority of the categories in the protocol apply to both content-area and ELP tests, some categories apply only to ELP tests due to the more extensive language present in passages of these tests. The application of the protocol yielded a linguistic profile for each test item.

Table 4

Content Analysis Protocol for Language Demands

Category	Subcategories/score scales
Linguistic features	
Length[a]	Total # of words (token), total # of unique words (type), total # of sentences, total # of words per sentence
Academic vocabulary[a]	General academic, context-specific, technical
Grammatical features[a]	Passive voice, modals, nominalizations, conditional clauses, relative clauses
Cohesion	Reference, substitution, adversative, causal, temporal, lexical
Sentence type	Simple, complex, compound, compound-complex
Non-linguistic features	
Form of presentation	The proportion of language to non-language (score scale 0–3)
Visual features	The amount of language in visuals (score scale 0–3)
Reliance on language (Content test only)	The extent to which the test taker needs language knowledge in order to answer an item (score scale 0–4)
Directness of information (ELP test only)	The extent to which the item may be answered correctly based on the information directly provided in the corresponding passage (score scale 0–2)
Scope of information (ELP test only)	The amount of information from a given passage needed to correctly answer an item (score scale 0–3)
Degree of academic specificity (ELP test only)	The degree to which items are subject-specific (score scale 0–2)
Academic language functions (ELP test: reading and listening passages only)	
Rhetorical mode	Argument, description, exposition, narration
Organizational features	Analysis, argument, comparison/contrast, definition, description, evaluation, exemplification, explanation, generalization, organization, prediction, question, summary
Thinking skills (ELP test: reading and listening items only)	
Type of comprehension	Literal, interpretive, inference, application

[a] There are more subcategories than the ones reported in this table. The complete list is included in the unabridged version of the content analysis protocol.

The categories of the content analysis protocol are elaborated in the following text. Definitions of each are also presented in Appendix CH1 found at the end of this chapter.

Linguistic Features

Length. The *total number of words* (*token*) and the *total number of unique words* (*type*) are counted to indicate the length of an item. In addition, the *total number of unique content words by type* (e.g., nouns, verbs, adjectives, adverbs) are counted with the assumption that the content of the test was primarily conveyed through these content words instead of functional words (e.g., articles, copulas, prepositions).

Academic vocabulary. Academic vocabulary was rated in three categories; *general academic*, *context-specific*, and *technical* vocabulary. General academic vocabulary is typically used in academic settings across multiple disciplines, such as *consequently* or *based on* (Coxhead, 2000). Academic vocabulary that is specific to disciplines is divided into two categories, generally following Chung and Nation's (2003) taxonomy. Context-specific vocabulary consists of words which may be heard in daily life, but they are also used in particular disciplines with specific meanings and are a part of academic content. Context-specific vocabulary is highly context-dependent in that it cannot be classified as discipline-specific without context. For instance, words such as *gas*, *liquid*, and *sound* can often be used in daily conversations, but the words are academic when the context of scientific properties is involved. Technical vocabulary consists of words which are discipline-specific and are seldom used outside of specific content-area classes. Examples of technical words are *hypotenuse, square root,* and *quasar*. One cannot use these types of words without knowing their discipline-specific meanings.

Grammatical features. The researchers coded five features that are commonly recognized as academic English features, as identified in previous literature (Butler et al., 2004; Scarcella, 2003; Schleppegrell, 2001). They include *passive voice phrases*, *modals*, *nominalizations*, and *conditional* and *relative clauses*. For example, nominalizations are frequently used in academic texts to concisely articulate complex concepts (Schleppegrell, 2001).

Cohesion. Among discourse features, the researchers focused on cohesive devices, which have explicit linguistic forms. Cohesive devices are typically used in academic written texts in order to connect text within or across clauses. For example, cohesive devices such as references (i.e., pronouns like he, hers, it) may connect the subject of a story across clauses. Cohesive devices identified in this analysis were *reference*, *substitution*, *adversative*, *causal*, *temporal*, and *lexical*.

Sentence type. Each sentence that appeared in the tests was classified as one of the following four basic sentence types: *simple*, *complex*, *compound*, and *compound-complex*.

Previous literature suggests that understanding and utilizing various sentence structures is a key academic language feature (Butler et al., 2004; Schleppegrell, 2001).

Non-linguistic Features

Non-linguistic features include categories that deal with how the text is presented, as well as the amount of text and visuals in the test items, prompts, or passages. The ratings in this section are more holistic compared to the earlier mentioned counts of linguistic features (e.g., length, academic vocabulary, sentence type). These categories are adapted from Bachman and Carr (1999; see Appendix CH1 [pp. 47–53] for complete score descriptions, definitions, and examples for all of the following categories).

Form of presentation (Form).[7] On a scale from 0 to 3, the rating for Form indicates the proportion of language to non-language in an item, including the presence of non-language. A score of 0 indicates that the item is composed entirely of non-language, while a score of 3 indicates the item is composed entirely of language.

Visual features (Visuals). This category rates the amount of language included in the visual feature (e.g., graphs, tables, diagrams, pictures). Each visual feature in item stems and responses are scored on a scale from 0 to 3. A score of 0 denotes no visual feature presence. A score of 1 indicates a visual feature with no language, while a score of 3 indicates that the visual feature contains more extensive language.

Reliance on language (Reliance). This rating describes the extent to which the test taker needs language knowledge (e.g., vocabulary knowledge, knowledge of textual relationship) in order to answer a test item. Each item is scored on a scale from 0 to 4, where a score of 0 indicates that language knowledge is not needed to arrive at the correct answer (i.e., there is no reliance), and a score of 4 indicates that in order to correctly complete the item the test taker must rely on all of the language in the item.

Directness of information (Directness). This category is only applied to passage-based items (i.e., reading and listening sections of an ELP test). It captures the extent to which the item may be answered correctly based on information directly provided in the corresponding passage. Scores are rated on a scale from 0 to 2, where a 0 indicates that the item may be answered correctly with only topical information (not directly from the given passage), and score of 2 shows that information from the passage is imperative to responding to the item correctly.

[7] Henceforth in the report, the shortened name of the feature given in parenthesis will be used to refer to the category.

Scope of information (Scope). This category is also only applied to passage-based items of an ELP test. It aims to capture the amount of information from a given passage needed to correctly answer an item. An item with a narrow scope requires a test taker to find specific information contained in the prompt, such as a vocabulary term. Broad scope items require the test taker to process the entire prompt and usually come to a conclusion, such as a "main idea" question in a reading test. Items are rated on a scale from 0 to 3; a rating of 0 indicates that an item requires no information to be extracted from the given passage and can be answered without reference to the passage. A rating of 1 indicates that to answer correctly, the scope the information is only contained at the word or isolated sentence level. A rating of 2 indicates that the scope of information is contained at the paragraph level. The highest rating, 3, indicates that to correctly answer the item requires information from the entire passage.

Degree of academic specificity. This category identifies the subject areas addressed in the ELP test items and passages, and the degree to which those items and passages are subject-specific. Items and passages are scored on a scale from 0 to 2, with a score of 0 being not academically specific, and a score of 2 being highly specific to a discipline. Additionally, the subject areas in the given items and passages are categorized into at least one of four subject areas (math, science, social studies, and English/language arts).

Academic Language Functions

Although academic language functions are concerned with the tasks students encounter in academic content areas (Chamot & O'Malley, 1994), the research team adapted this category to characterize the structure of texts and the authors' purposes for writing the passages found in the reading and listening sections of an ELP test. Butler et al.'s (2004) taxonomy is used to identify the specific types of academic language functions.

Rhetorical mode. As the first step in identifying academic language functions in a given text, the rhetorical mode of the entire text is identified. The subcategories include *argument, description, exposition*, and *narration*.

Organizational features. After the rhetorical mode is identified, the types of organizational features that the authors utilize to convey his or her purposes for writing are classified. These features include *explanation*, *comparison*, and *critique*, for example.

Thinking Skills

As Chamot and O'Malley (1994) noted, the level of thinking skills required are useful in identifying characteristics of academic language. For instance, higher-order thinking skills involve more complex language, which leads to more academic language demands. In examining the language demands in an ELP test, the researchers categorized the types of comprehension skills that a reading or listening item intends to measure.

Type of comprehension. To describe the specific types of comprehension skills measured through test items, categorical scores were recorded to indicate that the item required literal (score 1), interpretive (score 2), inference (score 3), or application (score 4) skills. These specific types are adapted from previous literature (Herber, 1978; Smith & Barrett, 1974). For example, an item may ask a test taker to identify a specific event in a given passage. This type of item is intended to measure a test taker's literal comprehension.

Procedure

A total of six raters were trained to use the content analysis protocol on the states' content-area tests and an ELP test. The raters included applied linguists, ESL teachers, and a former elementary school teacher. At least two raters were assigned to each category to analyze the items. On the content-area tests, the unit of analysis was at the item level. In the researchers' analysis, an item is the combination of a stem and its four responses (A, B, C, D). The term *item level* refers to an analysis that includes both the stem and responses. Items on the ELP test were also rated at the item level. In addition, the ELP test had prompts and passages that were rated separately. The researchers defined passages as a portion of a written work or speech that is related to a specific topic. The reading and listening sections typically have a set of items associated with a passage. The research team used Bachman and Palmer's (1996) definition of prompts as input in the form of a directive in which the purpose is to elicit an extended production response, as opposed to a multiple-choice answer, from the student.

The raters individually analyzed the items, passages, and prompts and discussed their ratings to reach consensus scores. For the categories of length, vocabulary, grammar, cohesion, sentence type, and organizational features, the ratings were simple counts of the occurrences of each feature. For non-linguistic features and type of comprehension, score scales were applied. The average number of occurrences for each category was computed at the item level. Because most of the items were comprised of single sentences, computing the occurrence of each category per sentence—as opposed to per item—did not provide meaningful information. In addition, considering the research team's future research purpose of comparing the language demands and the students' performance on each item, the analysis was conducted at the item level.

The reliability of the raters' scores was examined by computing the percentage of exact agreement on the score. Disagreements were resolved through a discussion among raters, and the resulting consensus scores were used to compute descriptive statistics for the language characteristics and demands of the items. In order to statistically compare the mean scores across the tests, content areas, grade levels, and states, the research team also examined the effect size of the mean differences using Cohen's *d* as effect size measures (Cohen, 1988). The research team considered the mean differences statistically significant when the effect size was greater than 0.5.

Results

The results are reported in three sections: (a) interrater reliability, (b) language characteristics in the content-area tests, and (c) language characteristics in the ELP test. The language characteristics in the ELP test are further divided into two sub-sections pertaining to (1) items and (2) passages and prompts.

Interrater Reliability[8]

Exact agreement for each category between two to three raters ranged from 62.5% to 99%. Overall, high agreement was obtained for counting linguistic forms such as vocabulary words, grammatical features, cohesion, and sentence type, as recorded in Table 5. The lowest reliability was found in the category of Reliance (62.5%). Although reliability for Reliance was low on average, increasing reliability was obtained as the rating progressed. The last round of rating yielded a rater agreement of 85.4%.

[8] Reliability estimates of Coefficient alpha or generalizability (G)-theory index could not be produced because there was not enough variability within the researchers' rating scores. As indicated earlier, the majority of the researchers' ratings were counts, and many categories examining the language demands had a count of zero, which produced little variability; thus, only exact agreement between the raters was computed as a reliability index.

Table 5

Percentages of Exact Agreement across Categories

Category	% of exact agreement	Number of items rated
Academic vocabulary	94.0	655
Grammatical features	97.0	655
Cohesion	96.4	655
Sentence type	96.6	655
Form of presentation	87.7	647
Visual features	97.0	655
Reliance on language	62.5	411
Directness of a passage	99.3	130
Scope of information	91.7	130
Degree of academic specificity	86.2	212
Organizational features	69.7	54
Type of comprehension	98.0	100

Language Characteristics in Content-Area Tests: Mathematics and Science

Based on consensus scores, descriptive statistics were computed and compared across grades, content areas, and states. The majority of the results were qualitatively described. Statistically significant differences based on the Cohen's effect size measure were noted when applicable. The following is a summary of overall findings across tests.

Length. Table 6 provides the mean of the total number of words (token) per item for math tests combined in two states, ranging from 25.5 to 36.2 words. The total number of unique words (type) was higher in the science than it was in the math tests, ranging from 16.9 to 23.8 words (d = 0.64 at Grade 5). Overall, the upper grades had more number of unique words (type) in both content areas, except for the Math Grade 8 Form 2 test in State B. However, statistical significance was detected only from State A math tests (d = 0.79). These trends were the same for the number of words per sentence and the number of unique words (type) per item.

Table 6

Mean Number of Words, Sentences, and Number of Unique Words per Item and Standard Deviations (*SD*) for Each Test by State, Subject, and Grade

Test	Measure	Number of words (token)	Number of words per sentence	Number of unique words (type)
State A				
Math G5	Mean	25.5	10.7	16.9
(*n* = 46)	(*SD*)	(11.6)	(3.3)	(8.7)
Math G8	Mean	27.0	10.4	17.7
(*n* = 51)	(*SD*)	(15.2)	(3.6)	(8.9)
State B				
Math G4	Mean	28.8	10.0	18.3
(*n* = 50)	(*SD*)	(15.4)	(2.7)	(6.9)
Math G7	Mean	30.3	11.0	20.1
(*n* = 50)	(*SD*)	(12.5)	(3.6)	(7.2)
Math G8 Form 1	Mean	30.5	11.2	20.6
(*n* = 60)	(*SD*)	(17.2)	(5.4)	(11.4)
Math G8 Form 2	Mean	30.7	11.4	16.7
(*n* = 60)	(*SD*)	(17.1)	(3.5)	(12.1)
State A				
Science G5	Mean	30.8	12.4	22.0
(*n* = 44)	(*SD*)	(16.7)	(2.8)	(7.4)
Science G8	Mean	36.2	12.2	23.8
(*n* = 42)	(*SD*)	(18.2)	(3.2)	(8.7)

Academic vocabulary. Overall, a wide variety of general academic, context-specific, and technical academic vocabulary was found for both math and science tests. The science tests had more technical than general academic vocabulary, including words such as *antibody, bacteria, cell, ecosystem,* and *friction*.

As shown in Table 7, the average number of academic words per item ranges from 1.9 to 7.0 across the tests, and comprises 12.2% to 32.2% of the total unique words. The average amount of total academic vocabulary (combining general, context-specific, and technical) per item was significantly higher on the science than on math tests when compared within the same grade level for State A (d = 1.40 at Grade 8). In State A Grade 8, the typical science item had approximately twice as many academic words on average than the state's math test. Between the grade levels, science items had more academic vocabulary in Grade 8 than in

Grade 5 (*d* = 1.03). Between the States in Grade 8, State A math items had three academic words and State B math items had 4 academic words, on average (*d* = 0.55). In terms of the proportion, State A Grade 8 math items had 15.5% of academic words from the total number of unique words compared to State B Grade 8 math items, which had an average proportion of 24% academic words.

Table 7

The Mean Number, Standard Deviations (*SD*), and Percentages of Academic Vocabulary Words (general, context-specific, and technical) per Item for Each Content-Area Test, by State, Subject, and Grade

Test	Measure	General academic vocabulary	%	Context-specific vocabulary	%	Technical academic vocabulary	%	Sum of academic vocabulary	%
State A									
Math G5 (*n* = 46)	Mean (*SD*)	1.2 (1.1)	7.0	0.4 (0.7)	2.5	0.8 (1.0)	5.8	2.4 (1.6)	12.2
Math G8 (*n* = 51)	Mean (*SD*)	1.3 (1.2)	5.7	0.5 (0.8)	3.0	1.2 (1.2)	6.7	3.1 (1.9)	15.5
State B									
Math G4 (*n* = 50)	Mean (*SD*)	0.7 (1.0)	4.4	0.5 (0.9)	3.2	0.7 (1.0)	4.7	1.9 (1.7)	12.4
Math G7 (*n* = 50)	Mean (*SD*)	1.5 (1.8)	7.0	0.8 (1.0)	4.4	1.1 (1.3)	6.5	3.4 (2.7)	17.9
Math G8 Form 1 (*n* = 60)	Mean (*SD*)	1.9 (1.8)	9.7	1.1 (1.2)	6.8	1.1 (1.3)	7.5	4.1 (2.3)	24.0
Math G8 Form 2 (*n* = 60)	Mean (*SD*)	1.9 (2.0)	7.4	1.2 (1.0)	5.7	1.3 (1.2)	6.0	4.4 (2.3)	19.0
State A									
Science G5 (*n* = 44)	Mean (*SD*)	1.9 (1.6)	8.7	1.1 (1.3)	5.5	1.1 (1.2)	5.8	4.1 (2.1)	20.1
Science G8 (*n* = 42)	Mean (*SD*)	2.6 (2.1)	12.1	0.9 (1.4)	3.9	3.5 (3.2)	16.2	7.0 (3.5)	32.2

Note. Percentage here refers to the proportion of unique academic words out of the total unique words per item.

Grammatical features. Of the grammatical features selected for analysis, only a few were found per item for both math and science. The mean numbers of grammatical features per item in both states' math tests indicated that there was only one of the five grammatical features present in an item, on average. Science items had a slightly higher number of grammatical features (2.8 for Grade 5, and 2.0 for Grade 8). A closer look at the grammatical features showed that nominalizations contributed to a higher mean for grammatical features in science. The means of the nominalizations for the math tests ranged from 0.2 to 0.8, while those for the science tests ranged from 1.0 to 1.7 ($d = 0.89$ at Grade 5).

Cohesion. For both math and science tests, cohesive devices such as reference, substitution, adversative, causal, and temporal words were rarely used in the test items. Instead, lexical cohesion was the most prevalent cohesive device, often forming word chains within the text. An example of an item containing lexical cohesion is illustrated in the following text (To maintain confidentiality of the test items in this study, the following example is a simulated item that is similar to an actual test item.). The italicized words show three word chains created through lexical cohesion.

> Jane draws *a rectangle*. *The rectangle* measures 2 inches by 3 inches. What is the area of *the rectangle*? [Italics indicate lexical cohesion.]

Sentence type. For both math and science tests, simple sentences were the most frequent type of sentence identified. Table 8 shows the percentages of sentence types found for each content-area test. Simple sentence structures comprised 68% to 84% of all sentence types found per test, on average. Complex sentence and compound sentence types were also observed, but they were not as frequent as simple sentence types. Almost no compound-complex sentences were found on any tests. In general, a higher grade test tended to have more complex sentence structure than a lower grade test, regardless of the content area. In particular, State B math tests at Grade 8 had the most complex sentence structures compared to all the other tests; 30% of the sentences in these tests were complex.

Table 8

Percentage of Sentence Types Found in Content-Area Tests by State, Subject, and Grade

Test	Simple sentences	Complex sentences	Compound sentences	Compound-complex sentences
State A				
Math G5	80.2%	18.8%	1.0%	0.0%
Math G8	77.5%	20.7%	1.8%	0.0%
State B				
Math G4	76.7%	20.5%	4.1%	0.0%
Math G7	83.5%	14.6%	2.4%	0.0%
Math G8 Form 1	67.5%	30.1%	0.7%	0.7%
Math G8 Form 2	67.5%	30.1%	1.5%	0.0%
State B				
Science G5	79.1%	20.9%	0.0%	0.0%
Science G8	72.0%	26.9%	1.1%	0.0%

Non-linguistic features. With respect to the Form category (the proportion of language versus non-language on each test), State A math tests contained a large portion of visuals (50% of the test items for Grade 5 and 45% for Grade 8). The proportion of visuals on State B math tests ranged from 28% (for Grade 8) to 40% (for Grade 4) of the items. Science tests were also found to contain a large amount of visuals (41% for Grade 5 and 43% for Grade 8). The mean scores for the Visual category (the amount of language in visuals) ranged 1.4 to 1.9 for both states' math tests, indicating that there was not much language in the visuals. Likewise, the mean visual scores for the science tests ranged from 1.5 to 1.8, indicating that there was more language in the science test visuals. Table 9 illustrates the mean scores of visuals in all tests.

Table 9

Mean Scores for Visuals (amount of language in visuals) by State, Subject, and Grade

	Scores for visual		
Test	*N*	Mean	*SD*
State A			
Math G5	23	1.8	0.7
Math G8	23	1.4	0.6
State B			
Math G4	20	1.8	0.7
Math G7	15	1.9	0.6
Math G8 Form 1	21	1.5	0.7
Math G8 Form 2	17	1.9	0.8
State A			
Science G5	18	1.5	0.5
Science G8	18	1.8	0.5

For the Reliance category (the amount of language needed to correctly process the answer to an item), both math and science tests in State A ranged from 2.3 to 2.8, indicating that processing one or two key vocabulary words was more crucial than processing the entire sentence in solving an item. The mean scores for State B math test was slightly higher, ranging from 2.8 to 3.2. This score suggests that processing an entire sentence was necessary to solve an item correctly. Table 10 displays the comparison of mean Reliance scores across the tests.

Table 10

Mean Scores for Reliance (amount of language to process to answer item correctly) by State, Subject, and Grade

	Scores for Reliance		
Test	*N*	Mean	*SD*
State A			
Math G5	46	2.3	1.1
Math G8	51	2.6	1.3
State B			
Math G4	50	2.8	1.3
Math G7	50	3.2	1.1
Math G8 Form 1	60	2.9	1.1
Math G8 Form 2	60	2.9	1.0
State A			
Science G5	44	2.8	0.8
Science G8	42	2.4	0.8

Language Characteristics in an ELP test

The content analysis protocol was applied separately to items and then to either the passages or prompts in each of the modalities[9]: in the reading and listening passages, and in the speaking and writing prompts. The results of the content analysis ratings are summarized at the item level first for comparison to the items in content-area tests. Subsequently, the results for passages and prompts are reported for each modality of the ELP test (reading, listening, writing, and speaking).

Language Characteristics at the Item Level in the ELP Test

Note that the speaking section was entirely composed of constructed-response items and prompts, which contained very little language to be analyzed. Thus, the item level comparisons that follow were made only among multiple-choice items in the reading, listening, and writing sections.

Length. A similar pattern is noted across the two grade bands (3–5, 6–8) in terms of the number of total words per item. The number of the total words was higher on the reading and

[9] The researchers use the word modality to refer to reading, writing, listening, and speaking sections of an ELP test because that is the term used predominantly in the literature and in NCLB legislation. However, the researchers recognize these four modalities can be referred to as domains by some states and test publishers.

listening sections than it was on the multiple-choice items in the writing section. The reading and listening section items had 22 to 28 words per item on average. The writing section items had 11 words per item on average. The number of words per sentence was slightly higher on the 6–8 Grade band (mean of 8.5) than on the 3–5 Grade band (mean of 7.6) for all modalities (*d* = 0.66 for reading, for example).

Academic vocabulary. Both general academic and technical academic words were identified, particularly for the reading and listening sections. Most technical academic vocabulary was related to science, including words such as *digestive, microorganisms,* and *reproductive*. Overall, more academic words were found in the 6–8 Grade band (*d* = 0.86 for reading). Table 11 shows the means, standard deviations, and percentages of academic words out of total unique words per item across each subcategory (general, context-specific, and technical academic vocabulary). The average number of academic words ranges from 0.9 to 2.9 across the tests, and comprises 7.9% to 12.6% of the total unique words.

Table 11

The Mean Number, Standard Deviations (*SD*), and Percentages of Academic Vocabulary Words (general, context-specific, and technical) per Item in ELP Test Modalities

Test	Measure	General academic vocabulary	%	Context-specific vocabulary	%	Technical academic vocabulary	%	Sum of academic vocabulary	%
Reading									
Grades 3–5	Mean (*SD*)	1.2 (1.2)	6.8	0.0 (0.2)	0.2	0.2 (0.5)	0.9	1.4 (1.3)	7.9
Grades 6–8	Mean (*SD*)	2.4 (1.8)	11.9	0.0 (0.0)	0.0	0.5 (1.1)	2.6	2.9 (2.1)	14.5
Listening									
Grades 3–5	Mean (*SD*)	0.7 (1.0)	4.5	0.3 (0.5)	1.4	0.8 (1.2)	3.2	1.8 (1.6)	9.1
Grades 6–8	Mean (*SD*)	1.8 (1.4)	8.1	0.3 (0.8)	0.7	0.6 (0.9)	2.6	2.7 (2.1)	11.3
Writing									
Grades 3–5	Mean (*SD*)	0.6 (0.8)	5.5	0.6 (1.1)	3.7	0.4 (0.6)	3.3	1.6 (1.5)	12.6
Grades 6–8	Mean (*SD*)	0.4 (0.9)	6.4	0.1 (0.3)	1.1	0.3 (0.7)	2.9	0.9 (1.4)	10.5

Note. Percentage here refers to the proportion of unique academic words out of the total unique words per item.

Grammatical features. Almost no grammatical features from the protocol were observed at the item level. The average occurrences of grammatical features ranged from 0.2 to 1.2 per item across all tests.

Cohesion. Compared to other categories on the protocol, cohesive features were frequently observed, ranging from 2.2 to 4.5 cohesive words per item on reading and listening sections. Various cohesive forms were detected including reference, substitution, adversative, causal, temporal, and lexical cohesion across all sections. Overall, both reference and lexical cohesion were used more than other cohesion features.

Sentence type. Sentence type was primarily limited to simple sentence at the item level. Table 12 shows the percentages of simple and complex sentence types found in the items of the ELP test for each modality. Simple sentence structures comprised 74% to 89%

of all sentence types found per test, on average. Interestingly, the lower grade test sections had more complex sentence structures than the higher grade test.

Table 12

Percentages of Sentence Types at the Item Level in ELP Test Modalities

Test	Simple sentences	Complex sentences	Compound sentences	Compound-complex sentences
Reading				
Grades 3–5	86%	14%	0%	0%
Grades 6–8	89%	11%	0%	0%
Listening				
Grades 3–5	79%	13%	8%	0%
Grades 6–8	85%	8%	2%	5%
Writing				
Grades 3–5	74%	26%	0%	0%
Grades 6–8	79%	21%	0%	0%

Non-linguistic features. The rating for Form (the proportion of language versus non-language on each test) indicated that almost all the items on all sections across both grades ranges were composed entirely of language (mean score = 3), with only minimum presence of numbers or dates in some items. The Directness of items (the degree to which an item requires information directly from a passage) was almost uniformly high, meaning that all items could be answered correctly using the information given on the passage. The rating of the Scope (the degree to which the item asked for specific versus global types of information) was consistent across all sections for both grades, ranging from the mean score of 1.2 to 2.2. Both broad (global) and narrow (specific) scope items appeared, with narrow scope items being dominant. The Degree of Academic Specificity (the degree to which the item was subject-specific) revealed that the majority of items were moderately academic specific (77% in listening, 71% in reading, 50% in writing, and 75% in speaking). Between grades, there were more moderately academic specific items in Grades 6–8 (73%) than in Grades 3–5 (66%).

Type of comprehension. The majority of items for Type of Comprehension (the comprehension skills the item measured) on both grade levels asked literal questions, or questions that ask the student to identify and recall information that was directly stated. For the listening and reading sections of the Grades 3–5 and 6–8 tests, over 84% of all the items consisted of literal questions, except for Grades 6–8 listening in which literal questions comprised of 67% of the items. Interpretive and inference questions were observed, ranging from 3% to 33% of the test. The high percentages of literal questions support the findings for Scope, which indicated that most test items asked for narrow, specific information instead of broad, global information.

Language Characteristics in Passages and Prompts in the ELP Test

The reading and listening passages and the writing and speaking prompts were analyzed separately from the test items in both grade bands. Table 13 displays the means and standard deviations for number of words per sentence, academic vocabulary, grammatical features, cohesion, and form for each modality (Scope and Directness ratings did not apply to prompts and passages; therefore, those ratings do not appear in the analysis for this section.). Additional ratings including sentence type, academic specificity, and academic language functions are summarized in the following text.

Table 13

Means and Standard Deviations (*SD*) for Passages (listening, reading) and Prompts (speaking, writing) across Sections (*n* = number of passages or prompts)

		Number of words per sentence[a]		Academic vocabulary per passage		Grammatical features per passage		Cohesion per passage		Form per passage	
Modality	Measure	3–5	6–8	3–5	6–8	3–5	6–8	3–5	6–8	3–5	6–8
Listening (*n* = 5)	Mean (*SD*)	11.6 (2.5)	12.1 (1.4)	7.0 (5.6)	11.4 (7.7)	9.8 (1.3)	7.2 (3.1)	36.6 (10.3)	35.4 (15.5)	3.0 (0.0)	3.0 (0.0)
Reading (*n* = 3)	Mean (*SD*)	12.5 (2.8)	14.3 (2.2)	8.3 (7.1)	18.0 (9.5)	6.3 (2.9)	17.3 (7.0)	41.0 (7.0)	29.0 (6.1)	3.0 (0.0)	3.0 (0.0)
Speaking (*n* = 7)	Mean (*SD*)	8.7 (1.9)	9.4 (2.8)	0.6 (0.8)	0.9 (1.2)	0.1 (0.4)	0.4 (0.5)	6.7 (3.8)	6.6 (4.1)	2.9 (0.4)	2.9 (0.4)
Writing (*n* = 3)	Mean (*SD*)	7.8 (0.5)	9.3 (2.5)	2.7 (1.2)	6.0 (2.0)	0.0 (0.0)	2.7 (0.6)	8.7 (6.4)	7.3 (4.2)	3.0 (0.0)	2.7 (0.6)

[a] This is one of the subcategories to indicate overall length. See Table 4 for a list of the other subcategories under the length category.

The general trend for passages and prompts revealed that the tests for Grades 6–8 are more linguistically complex than those for Grades 3–5 in that the means tended to be higher in the categories for number of words per sentence, academic vocabulary per passage, and grammatical features per passage. Table 13 summarizes the results of the ratings for the passages and prompts. The only category in which the Grades 3–5 test had a higher mean across all the modalities than the Grades 6–8 test was in cohesion. The ratings for Form for both grade bands were the same (except in the writing modality in which the difference between the two means was small) and indicated that most of the passages and prompts were composed entirely of language in both bands.

Sentence type. Sentence type was limited primarily to simple sentences for both passages and prompts. Simple sentence structures comprised 51% to 58% of all sentence types found per test (Grades 6–8, 3–5, respectively) on average. The Grades 6–8 test had a slightly higher percentage (41%) of complex sentences than the Grades 3–5 test (36%). Both tests had equal percentages of compound sentences (4%).

Non-linguistic features. As shown in Table 13, the rating for Form was from 2.7 to 3.0, which means that the majority of the passages and prompts on all sections over both grades were composed entirely of language. The Degree of Academic Specificity rating showed that the majority of the passages and prompts were moderately academic specific (90% in listening, 75% in reading, and 100% in writing). Between grades, there were more moderately academic specific passages and prompts in the Grades 6–8 (74%) than in Grades 3–5 (63%).

Academic language functions. Exposition was found to be the predominant rhetorical mode for the listening and reading passages. In the listening modality for Grades 6–8, more organizational features (13 features) were identified than in Grades 3–8 (8 features). Common features were description, explanation, evaluation, comparison, contrast, exemplification, question, and generalization. In the reading modality, eight organizational features were found in the passages for Grades 3–5, and six organizational features were found for Grades 6–8.

Discussion

The present study investigated the academic language characteristics on content-area and ELP tests. Specific linguistic features that characterize academic English were examined, including vocabulary, grammar, cohesion, and sentence types. Academic language functions and types of comprehension were also examined for the reading and listening sections of the ELP test. In this section, the key findings for the characteristics of academic English on the tests are discussed, followed by a comparison of the language characteristics between the content-area and ELP tests.

Characteristics of Academic English in Content-Area Tests

Overall, few academic English grammatical features were found at the item level in terms of grammatical features across all tests. One possible explanation for this finding is that it reflects the inherent characteristics of the language of tests. As Bailey (2000) noted, test language has a conventional script with specific structures. For instance, a formulaic expression, ‘Which of the following is’ constrained an item to be a simple sentence structure with no cohesive forms. In fact, according to the recent technical report for State A assessment, State A, in its item development guidelines, explicitly directs item writers to avoid the conditional sentence structure and to use only simple and complex sentences, having consistent verb tenses, concise stems, and consistent forms between stems and choices sistently simple grammatical structure across math and science tests in State A may be related to this state’s intention of reducing test items’ linguistic complexity in order to better measure content knowledge. One interesting finding in grammatical features was that nominalizations were observed more than any other grammatical feature in the protocol. Science items had more nominalizations than math items, suggesting the different nature of each content area. That is, science language at these grade levels tends to utilize nominalizations more readily to express concepts and complex processes concisely.

The simplicity of language also was evident in cohesion and sentence structure for both math and science tests. Among the cohesive forms, lexical cohesion was the most prevalent. The use of lexical cohesion instead of other cohesive devices seems to be associated with the characteristics of test language. That is, lexical cohesion, by repeating words or phrases, seemed preferred for indicating textual relationships, rather than using substitution or reference. For example, a math problem would be written in the following way:

> *John* bought a *shirt* on sale for 50% off the original price. If the *shirt* originally cost $25, what was the final sale price that *John* paid for the *shirt*? [Italics indicate lexical cohesion.]

Notice in the example, the appropriate references such as “he” for John or “it” for shirt were not used. With respect to sentence structure, simple sentences were dominant, although some complex sentences were observed. It was interesting to note that more complex sentence structures were found in a State B math test than in any other test, showing variation between states as well as between content areas.

Based on the number of occurrences observed, the most prominent academic language feature was found to be vocabulary. Among all the categories, the Cohen’s *d* index detected a medium to a large effect size in academic vocabulary when compared between grades, content areas, and states. For the research team’s purpose, academic vocabulary was identified at three levels: general academic, context-specific, and technical vocabulary. This classification was not only intended to discern content and language knowledge, but also to pinpoint ELL students' difficulty for future studies. For example, one can hypothesize that technical words may be made more accessible to ELL students because those words are explicitly taught in classrooms. Context-specific words might have more variations in terms of learning because one can assume that students have different familiarity with those words from exposure in non-academic settings.

The prominence of academic vocabulary on the tests is more notable given that nominalizations and lexical cohesion were observed more than any other grammatical or discourse features. Both nominalization and lexical cohesion are realized in one or two vocabulary words. In particular, nominalization was often characterized as academic vocabulary (e.g., *scientific investigation, chemical reaction*). Interestingly, Reliance ratings results were also supportive of the importance of vocabulary knowledge on the tests. Both math and science items in one state obtained an average score between 2 and 3, suggesting that possessing vocabulary knowledge was critical to solving an item correctly. That is, on average in both math and science tests, solving an item correctly did not necessarily involve understanding an entire sentence. Although it is commonly assumed that science requires more language knowledge than math, the current analysis suggests that the linguistic complexity and language demands across tests may be more subject to specific test item-writing rules rather than to the difference in content areas.

Analysis of the language characteristics of the tests revealed variation between the two states. That is, the language demands of items on tests of the same content areas varied across the two states. Although both states' math tests covered a range of topics and standards, State B had a higher average number of academic vocabulary words per item than in State A, which indicated that there was more complex language to process in State B math tests. Reliance ratings in State B math tests were higher than they were in State A science tests, on average, indicating that students might need to process the entire sentence to solve an item correctly on the State B math tests; whereas, on the State A science tests, vocabulary knowledge might be sufficient for correctly answering an item.

Characteristics of Academic English in an ELP Test

As described earlier, the ELP test was analyzed for its items and passages/prompts separately. At the item level, almost no academic English features were present. It appears that test language characteristics greatly influenced the format of the item structure. The items in the reading section typically included simple WH-questions (e.g., 'What is the title/meaning/main idea?'). This structure seems related to the type of comprehension skills presented in items as well. The research team's ratings on the type of comprehension in reading and listening items indicated that the majority of items entailed extracting single, explicit information from given passages. In addition, the rating of Scope (the amount of information needed to answer an item correctly) also indicated that the most items required a narrow scope of information from a passage. Interestingly, listening items required broader information from a given passage than reading items, on average. It will be interesting to further examine how ELL students perform in each modality.

With respect to academic vocabulary, all three types of academic vocabulary were identified in the reading, listening, and writing sections. This result indicates that the given ELP test incorporated vocabulary words from different content areas. However, speaking items contained little, if any academic vocabulary, due to the nature of their prompts. Speaking prompts consisted of pictures in which students were asked to describe or label them.

In contrast to the academic language found at the item level, more diverse academic language features were exhibited in passages or prompts which had more extensive texts. All modalities included academic topics in passages and prompts. The identified areas were mostly math, science, and social studies, with only few specifically language arts topics. Various academic words were present, particularly in reading and listening passages. A wider variety of grammatical features were also observed, including passive voice, modals, conditional, and relative clauses. The fact that more cohesive devices were identified is most likely due to the presence of more text in the passages compared to the items. Reference, in particular, was most prevalent among the cohesive devices that were observed. Sentence structure was also diverse; compound and complex-compound sentences were observed in addition to simple and complex sentences. Additionally, complex sentences often included multiple embedded clauses. Organizational features ratings revealed that a variety of academic language functions were utilized in the passages. The features were consistent with ones that were found in science and social studies textbooks (Bailey et al., 2007; Butler et al., 2004). The passages in the test appeared similar in terms of linguistic characteristics to the texts that could be encountered in an academic context.

Conversely, speaking and writing prompts contained very few linguistic forms. Most prompts were presented in non-linguistic visual forms possibly to focus on the assessment of productive language. Although the visuals contained little language, they were related to a specific content area (e.g., math, science).

Comparison of Language Demands in Content-Area and ELP Tests

The pattern of language demand for the ELP test items was found to be similar to that observed in the content-area and ELP tests. In all instances, grammatical features, cohesion, and sentence structure were relatively simple at the item level. As discussed earlier, this finding seems to be partly due to the limited structure of test language, particularly at the item level. It also seems related to the test developer's intention of simplifying the language in test items in order to clearly convey the meaning of the items. A slight difference in grammatical features was that there were relatively fewer nominalizations on the ELP test than in the content-area tests. This is not surprising, considering that content-area tests conveyed more terms to describe content-specific concepts or processes than did the ELP test. Among cohesive devices, lexical cohesion was the most prevalent in content-area tests, whereas reference was the most common in the ELP test. This may be due to the content-area test developers' intention to avoid references or substitutions in order to denote references within the text. In contrast, the items in the ELP test frequently used references to refer to information in passages.

Academic vocabulary was the most prominent feature of academic English characteristics that was observed in the test items. All three of the different types of academic vocabulary words were identified in both the content-area and ELP tests, although the items of the content-area tests were perceived to have more academic words than those of the ELP test. As expected, context-specific and technical vocabulary words were more prevalent in the content-area tests than in the ELP test. The presence of technical vocabulary in the ELP test reflects the current movement of aligning academic standards and ELP standards. Although the language in the ELP test items consisted mostly of non-academic language features (e.g., few academic vocabulary, few academic grammatical features), the language in passages and prompts did contain academic English features.

A summary of key comparison findings across the math, science, and ELP tests is qualitatively described in Table 14.

Table 14

Summary of Comparisons between Math, Science, and ELP Tests[a]

Category	Summary points
Length	The science test for Grade 8 had the highest means in all the subcategories for Length. Overall, the science tests had higher means than the math tests. The Grade 8 math test for State B and the Listening section of the ELP test for Grades 6–8 had similar means. Grade 8 tests had higher means than tests in Grades 4, 5, and 7, regardless of the content area.
Academic vocabulary	Overall, the science test for Grade 8 had the highest mean for academic vocabulary words. The second highest mean of the number of academic vocabulary was found in the Grade 8 math test for State B. The mean of academic vocabulary for State B math tests was higher than that of State A math tests for Grade 8. The ELP test had smaller means for academic vocabulary than content-area tests, across similar grade levels. Grade 8 tests had higher means than tests in Grades 4, 5, and 7, regardless of the content area.
Grammatical features	The science tests had higher means for overall grammatical features, due to its high mean of nominalizations. The Grade 8 science test had highest means of passive voice phrases. All other tests had a mean of around 1.0 for all subcategories, indicating a limited use of academic grammatical features.
Cohesion	The ELP test had higher means than any other tests. Lexical cohesion was the most prevalent in science tests. Reference was the most prevalent in the ELP test.
Sentence type	Most test items were in simple sentences across all tests. The State B math test for Grade 8 had the highest percentage of complex sentences, closely followed by the State A science test for Grade 8.
Form of presentation[b]	All the grade bands and sections of the ELP test had the highest means, indicating that the items were entirely composed of language. All math tests across states and grades had the low means, indicating that the items were not entirely composed of language, but contained visual forms (diagram, tables, pictures, etc.).

[a]For the ELP test, only item-level data were used in this table, not the data for passages and prompts. The purpose of this table is to compare common item types across content-area and ELP tests. Because content-area tests do not have passages and prompts that are found on the ELP test, their data were excluded from this table.
[b]Only the subcategory *Form of presentation* was compared among the non-linguistic features. This was the only subcategory that appeared across both the content-area and ELP tests.

One of the key findings in the researchers' analyses of the academic English in state content-area tests was the variation across the two states. As described in Table 14, State B math tests were found to have as many academic words as in the science tests. Although science tests were generally perceived to have more language demands than math tests in terms of length and academic vocabulary, the findings in this study suggest that a math test could be linguistically demanding. As expected, Grade 8 tests seemed to be more demanding in terms of length and academic vocabulary than lower grade tests, regardless of content areas. It should be noted that the findings of this study may be limited to the specific tests being analyzed. Given the variations that were found across the two states in the one content area (math) with tests from both states in this study, generalizations about the academic English characteristics identified on the tests is limited. Although the present study found that academic vocabulary was the dominant academic language feature, grammatical features or various sentence structures might be as dominant as vocabulary in other tests. The study is also limited in the range of academic English features it examined. This study focused on one set of features of academic English, linguistic forms, in order to examine tests in which items had limited text. Other academic language features, such as diverse discourse and cognitive demands used in academic contexts, were not investigated in this study.

Implications and Future Studies

Mastery of academic English is one of the most important determinants of success in academic content for students (e.g., Proctor, Carlo, August, & Snow, 2005). The results of the present study highlight the importance of academic vocabulary within the academic language construct. This finding carries an important implication for a future research that examines the validity and effects of accommodations. That is, it will be imperative to investigate the effects of vocabulary-related accommodations (e.g., glossary, customized dictionary), compared to those of other types of accommodations for ELL students. The current classification of academic vocabulary may provide a useful guide to provide such vocabulary-related accommodations in a principled way in order to assess student content knowledge. For example, context-specific and technical vocabulary should not be altered since those are part of content knowledge. In other words, accommodations that change *context-specific* and *technical* vocabulary are likely to change the construct that is tested, and thus raise questions about the validity of assessment-based interpretations. On the other hand, accommodations that change or gloss *general* academic vocabulary may preserve the construct that is tested, and thus support the validity of assessment-based interpretations.

The need for further studies is evident. It will be useful to examine the linguistic difficulty that ELL students encounter in taking content-area tests. The present study provides a base to conduct research in this respect. One of the research aims of the current project is to examine how ELL students perform on the items that contain more general academic or technical vocabulary, for instance. It will also be of interest to apply the research's content analysis protocol to other tests in different grades in order to further examine the nature of the language demands of content tests at grade levels and different content areas. Additionally, it will beneficial to refine the current content analysis protocol for broader applications and to better capture the language demands of tests imposed on ELL students that assess their academic and social language.

References

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

Bachman, L. F. & Carr, N. T. (1999). UCLA ESLPE Language Testing Project: Test Task Characteristics (TTC) Rating Instrument. Department of Applied Linguistics & TESL, University of California, Los Angeles.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An Investigation Into the Comparability of Two Tests of English as a Foreign Language*. Cambridge, England: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford, UK: Oxford University Press.

Bailey, A. L. (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* (CRESST Tech. Rep. No. 663). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2001/2004). *Towards the characterization of academic English in upper elementary science classrooms*. (Final Deliverable to OERI Contract No. R305B960002) (CSE Tech. Rep. No. 621). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey A. L., Butler, F. A., Stevens, R., & Lord, C. (2007). Further specifying the language demands of school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 103–156). New Haven, CT: Yale University Press.

Butler, F. A., Bailey, A. L., Stevens, R. A., Huang, B., & Lord, C. (2004). *Academic English in Fifth-Grade Mathematics, Science, and Social Studies Textbooks* (CRESST Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., & Castellon-Wellington, M. (2000/2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In J. Abedi, A. Bailey, F. A. Butler, M. Castellon-Wellington, S. Leon, and J. Mirocha, *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives* (CRESST Tech Rep. No. 663, pp. 47–78). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Chung, T. M. & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language, 15*(2), 103–116.

Cohen J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 43*(2), 213–238.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). Los Angeles: National Dissemination and Assessment Center.

Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire*. Clevedon, England: Multilingual Matters.

De Avila, E. A., & Duncan, S. E, (1990). *Language assessment scales, oral administration manual, English: Forms 2c and 2d*. Monterey, CA: CTB McGraw-Hill.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* New York, NY: Cambridge University Press.

Flowerdew, J. & Peacock, M. (2001). *Research Perspectives on English for Academic Purpose*. New York, NY: Cambridge University Press.

Herber, H. (1978). *Teaching reading in content areas,* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Jordan, R. R. (1997). *English for Academic Purposes. A Guide and Resource Book for Teachers.* New York, NY: Cambridge University Press.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Proctor, C. P., Carlo, M., August, D. & Snow (2005). Native Spanish-Speaking Children Reading in English: Toward a Model of Comprehension. *Journal of Educational Psychology*, *97* (2), 246–256.

Scarcella, R. (2003). *Academic English: A conceptual framework* (Tech. Rep. No. 2003–1). Santa Barbara: University of California, Linguistic Minority Research Institute.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, *12,* 431–459.

Short, D. J. (1993). Assessing Integrated Language and Content Instruction. TESOL Quarterly, 27(4), 627–656.

Short, D. J. (1994). Expanding Middle School Horizons: Integrating Language, Culture, and Social Studies. TESOL Quarterly, 28(3), 581–608.

Smith, R. J., & Barrett, T.C. (1974). *Teaching In the Middle Grades.* Reading, MA: Addision-Wesley.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: measuring the progress of ELLs* (Final Deliverable to OERI, Contract No. R305B60002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

U.S. Department of Education (2004). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. Office of Elementary and Secondary Education, Washington, DC.

APPENDIX CH1

Operational Definitions for Abridged Content Analysis Protocol

Categories and Subcategories of the Content Analysis Protocol for Language Demands

Categories	Subcategories
Length[a]	Total # of words (token), total # of words per sentence, total # of unique words (type), total # of unique content words by type.
Academic vocabulary[b]	General academic, context-specific, technical academic.
Grammatical features[c]	Passive voice phrases, modals, nominalizations, conditional clauses, relative clauses.
Cohesion	Reference, substitution, adversatives, causal, temporals, lexical.
Sentence type	Simple, complex, compound, compound-complex.
Non-linguistic features	Form of presentation, visual features, reliance on language, directness of information, scope of information, degree of academic specificity.
Academic language functions	Rhetorical mode, organizational features.
Thinking skills	Type of comprehension (literal, interpretive, inference, application).

[a] Additional analysis for length include total # of paragraphs, total # of sentences per paragraph, total # of words per paragraph, and total # of content words.
[b] Additional analysis in vocabulary include derived words, measurement and abbreviated words, and proper nouns.
[c] Additional analysis in grammatical features includes prepositional phrases.

Definition of Each Category

Length	
Total # of words per sentence	The total number of words (token), excluding words in visuals, divided by total number of sentences in an item.
Total # of words (type)	The total number of unique words in an item.
Total # of words (token)	The total number of words in an item.
Total # of unique words by type	The total number of unique content words in an item. Content words include nouns, verbs, adjectives, adverbs, and pronouns.
Vocabulary	
General academic vocabulary	Words used with the same or similar meaning across multiple content areas. "Multiple content areas" is defined as three or more of the following: science, mathematics, social studies, and language arts.
Content-specific academic vocabulary	Words which may be heard in daily life, but are also used in particular disciplines with specific meanings and are a part of academic contexts. Whether a word is classified as context-specific academic vocabulary or as nonacademic vocabulary will depend on the context in which the word appears.
Technical academic vocabulary	Words which are highly discipline-specific and seldom used outside of specific content area classes.
Grammatical features	
Passive voice phrases	A phrase in which the predicate of the phrase includes a form of be/become and past participle.
Modals	An auxiliary verb that express meanings such as necessity and possibility, usually followed by a verb. Modals include: will, would, shall, should, can, could, may, might, must, have to, ought to, had better, need, and dare.
Nominalization	A noun or noun phrase that may be derived from a verb or adjective (by means of a suffix such as –tion, –ate, etc.) and captures or encompasses a process/ activity/ event/ state/ property/ action/ state/ result.
Conditional sentence	A clause which expresses condition and result.
Relative clause	A clause that modifies a noun by identifying or classifying the noun.

Cohesion: The links that hold a text together across clause boundaries to give it meaning, and contains both grammatical and lexical components. Cohesive devices work within sentences and across clauses.	
Reference	A cohesive device in which pronouns (e.g., *he, she, it, we, you*) and deictic expressions (e.g., *here, there, now, then, this, that, the former, or the latter*) refer to elements in the text.
Substitution	A word which replaces an element with another that is not a personal pronoun. It occurs when a word or phrase is left out, and it is substituted for another, more general word.
Adversatives	Contradicting conjunctions (e.g., *however, but, although*) that are usually of a form or construction marking an antithesis. They extend meaning of one clause or sentence to a previous or subsequent one. They are conjunctions that create cohesion by relating sentences and paragraphs.
Causals	Adverbs that indicate or involve time. They extend the meaning of one clause or sentence to subsequent ones to show a causal relationship (e.g., *consequently, more over, additionally*).
Temporals	Conjunctions that refer to time and that extend the meaning of one sentence to a subsequent one (e.g., *before, after, now*). They are conjunctions that create cohesion by relating sentences and paragraphs.
Lexical cohesion	Cohesion that comes from the semantic (or meaning) relationships between words. Lexical cohesion "…is the result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units of text being 'about the same thing'…" Lexical chains can occur between noun words and over a succession of nearly related words spanning an entire text.
Sentence type	
Simple sentence	A sentence that contains one independent clause.
Complex sentence	A sentence that contains one independent clause and one or more dependent clauses.
Compound sentence	A sentence that contains two or more independent clauses, typically joined by a coordinating conjunction or with punctuation.
Compound-complex sentence	A sentence that contains one or more dependent clauses *and* two or more dependent clauses.

Non-linguistic features	
Form of presentation	The proportion of language to non-language (drawings, equations, etc.) in an item. The rating is a scale of 0 to 4 for the test item. (0) An item that is entirely composed of non-language. (1) An item contains primarily non-language and some language. (2) An item contains primarily language and some non-language. (3) An item that is entirely composed of language.
Visual features	A measure of the amount of language included in the item's visual features, if visual features are present. The rating is a scale of 0 to 3 for item stem and responses separately. (0) No visual features in stem or responses. (1) Visual feature(s) contains no language. (2) Visual feature(s) contains only the labeling (naming) of objects. (3) Visual feature(s) contains language or labels that describe or explain the visual feature(s).
Reliance on language	The extent to which the test taker needs to process the language in an item in order to answer it correctly. The rating of is a scale of 0 to 4 for the test item. (0) Knowledge of language is of no help in arriving at the correct answer (no reliance). (1) Knowledge of language is helpful but not necessary to arrive at the correct answer (no reliance) (2) In order to arrive at the correct answer, the test taker needs to process some language, but that language is only at the level of lexical items. (reliant) (3) In order to arrive at the correct answer, the test taker needs to process some language which is at the level of sentences. (reliant) (4) In order to arrive at the correct answer, the test taker needs to process some language which is at the level of textual relationships or connected discourse. (reliant)
Directness of information	A measure of the extent to which an item may be answered based solely on information directly provided in the corresponding passage of an ELP test. The rating of directness is a scale of 0 to 2 for passage-based items only. (0) Item requires the test taker to base response solely on topical knowledge and/or context. Test taker cannot answer item on the basis of the information provided in the prompt. (1) Item requires the test taker to base response equally on information provided in the prompt and topical knowledge and/or context. (2) Item does not require test taker to base response on topical knowledge and/or context. Test taker can answer item solely on the basis of information provided in the prompt

(*table continues*)

Non-linguistic features (continued)

Scope of information	A measure of the amount of information from a given passage needed to correctly answer a passage-based ELP test item. The rating of scope is a scale of 0 to 2 for passage-based items only. (0) Items can be answered without reference to the prompt. (1) Item requires an understanding of only a few words or an isolated sentence in the prompt. (2) Item requires the test taker to process multiple sentences in the prompt. (3) Item requires an understanding of the entire prompt.
Degree of academic specificity	Identifies the content areas addressed in the ELP test items and the degree to which those items are subject-specific. The rating of degree of academic specificity is a scale of 0 to 2 for the test item. (0) Item is not academically specific. (1) Item is somewhat academically specific. (2) Item is highly academic specific.

Academic Language Functions	
Level 1: Rhetorical	*Argument:* Text that brings something to the attention of the reader and makes a claim, comment, or argument. To support and prove the validity of an idea or point of view, sound reasoning, discussion, or a call for action are presented. Often there is controversy surrounding the topic.
	Description: Text that tends to focus on specifics in order to recreate, invent, or visually present people, places, events, or actions for the reader to visualize what is being described. Descriptive passages usually progress through a scene. Descriptive writing is often found within other rhetorical modes.
	Exposition: Text that provides, explains, or analyzes information, usually presenting it as non-controversial. Expository passages are dominant in textbooks, journals of information, or in-depth studies of particular topics.
	Narration: Text that presents a sequence of events that tells a story. Key to narration is the passing of time, or the successive stages in time.
Level 2: Organizational Features	*Analysis:* To identify the parts of a whole and their relationship to one another.
	Argument: To discuss a point of view with the purpose of creating agreement around a position or a conviction.
	Comparison/ Contrast: To examine or look for differences and/or similarities between two or more things.
	Definition: To say what the meaning of something, especially a word, is.
	Description: To say or write what someone or something is like, usually a physical description, or telling about one's thoughts.
	Evaluation: To judge and assign meaning or importance or quality to a particular experience or event; to review or analyze critically. Words or expressions that give an opinion or express emotions are usually evaluative, such as "good," "bad," "amazing," etc. Also, there must be information provided that supports an evaluative comment. In other words, there needs to be reasons associated with an evaluative comment.
	Exemplification: To provide evidence to support ideas; to list or name things separately, one by one.
	Explanation: to offer reasons for or a cause for ideas or processes; to provide instruction or guidance.
	Generalization: to infer a trend or a principle, or make a conclusion based on facts or statistics. Statements that take information about one thing and apply it to something else.
	Organization: to give structure to something, which includes information or data; to arrange or order things.
	Prediction: to say that an event or action will happen in the future, especially as a result of knowledge or experience; to form an idea or explanation for something that is based on known facts but has not yet been proved; to reason from circumstance.

(table continues)

Level 2: Organizational Features (continued)	*Question:* to seek information by forming questions, usually in the form of rhetorical questions
	Summary: to express the most important facts or ideas about something or someone in short and clear form
Thinking skills	
Type of comprehension	*Literal:* Identify and recall information directly stated. It is read at face value with little interpretation. Only a surface level understanding is necessary. All of the information needed to respond is found directly in the text.
	Interpretive: Interpret information contained in the text. Consolidation or reorganization of information may be necessary in order to respond to the item. Relevant text may not be explicitly stated, but additional information or inferences are not needed.
	Inference: Extrapolate information in order to show understanding. The reader may be required to respond to information implied but not directly stated. Inference items are typically prompt-based.
	Application: Ability to use acquired knowledge, facts, and techniques. Items may require synthesis of content and ideas from multiple sources. Connections with text, self, and world may also be necessary.

CHAPTER 2:

DETECTING TEST ITEMS DIFFERENTIALLY IMPACTING THE PERFORMANCE OF ELL STUDENTS

Jamal Abedi, Seth Leon, Mikyung Kim Wolf, and Tim Farnsworth.
CRESST/University of California, Los Angeles

To provide fair assessment and uphold standards on instruction for every child in this country, both federal (e.g., No Child Left Behind Act of 2001 [NCLB; 2002]) and state legislation now require the inclusion of all students, including English Language Learner (ELL) students, into large-scale assessments (Abedi & Gandara, 2006; Mazzeo, Carlson, Voelkl, & Lutkus, 2000). Such inclusion requirements have prompted new interest in modifying assessments to improve the level of ELL students' participation and to enhance the validity of inferences drawn from the assessments.

However, research on the assessment of ELL students and students with learning disabilities strongly suggests that language factors can threaten the validity and reliability of content-area assessments (Abedi, 2006; Abedi, Leon, & Mirocha, 2003; Abedi & Lord, 2001). Minor changes in the wording of content related test items can raise student performance (Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997; Cummins, Kintsch, Reusser, & Weimer, 1988; De Corte, Verschaffel, & DeWin, 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983). The results of recent studies have shown that reducing the unnecessary linguistic complexity of test items help to improve the performance of ELL students and students with learning disabilities without compromising the validity of assessment (see Abedi & Lord, 2001; Kiplinger, Haug, & Abedi, 2000; Maihoff, 2002).

The purpose of the DIF analysis conducted in Phase I of this project was two-fold: (a) to investigate test items that may function differentially for ELL students and (b) to examine the linguistic characteristics of those items. Furthermore, this study aimed to propose a sound method to systematically investigate potential item bias for ELL students, providing a source of validity evidence to support appropriate inferences from a given assessment.

Method

Description of DIF Technique

The DIF approach is often used to examine any systematic biases that a test item may have toward a specific group of students. DIF is said to be present when an item parameter (i.e., difficulty, discrimination, guessing) is different for two groups of students who are at the same ability level. More specifically, an item is said to exhibit uniform DIF when the probability of answering the item correctly is greater for one group than the other regardless of the ability level. Non-uniform DIF exists when the probability of answering the item correctly varies among students of different ability levels (Mellenbergh, 1982). In practice, DIF is typically conducted to detect biased items with respect to gender and ethnicity. Different methods have been suggested for examining the differential functioning of dichotomously and polytomously scored items (see Allen & Donoghue, 1996). Various techniques have been used to examine DIF, including the Quasi-chi-square (Scheuneman, 1975, 1979), log-linear (Alderman & Holland, 1981; Loyd, 1984; Mellenbergh, 1982), Mantel-Haenszel (MH; Holland & Thayer, 1988), standardization procedure (Dorans & Kulick, 1983, 1986), the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993), and the logistic regression approach (Spray & Carlson, 1988). National Assessment of Educational Progress (NAEP) has frequently used two different approaches, a graphical method and the MH procedure, while Educational Testing Service (ETS) also frequently uses the MH method. The Item Response Theory (IRT)-based DIF approaches have also been used in recent DIF literature (e.g., Thissen, 2001).

Regardless of which technique is used, in order to assess possible DIF between two groups, examinees from the focal group (i.e., the group to be studied) must be matched based on their abilities underlying the performance, or θ, to a reference group (the group which is used as a standard against which the performance of the focal group is compared). Therefore, the accuracy of DIF techniques depends to a great extent on the validity and comprehensiveness with which the underlying ability is measured. This underlying ability, θ, is estimated based on IRT, or sometimes it is estimated based on the total score of the test. In the MH approach, the subjects in the focal group and the reference group are matched based on the total score of the test (i.e., the total number of correct responses on the test). Traditionally, the MH approach is applied in testing programs where all examinees receive the same set of items. The IRT approach to DIF is based on information from the Item Response Curve (IRC). If a major difference is found between the IRC of the reference group and IRC of the focal group then that item may exhibit DIF. The magnitude of DIF depends on how the IRCs for the focal and reference groups vary. Figures 1 and 2 illustrate the IRCs of a DIF and a non-DIF item, respectively. As seen in Figure 1, at the low ability level the probabilities of these two groups responding correctly to the item is considerably different. On the other hand, the non-DIF item demonstrates identical IRCs for two groups, as shown in Figure 2.

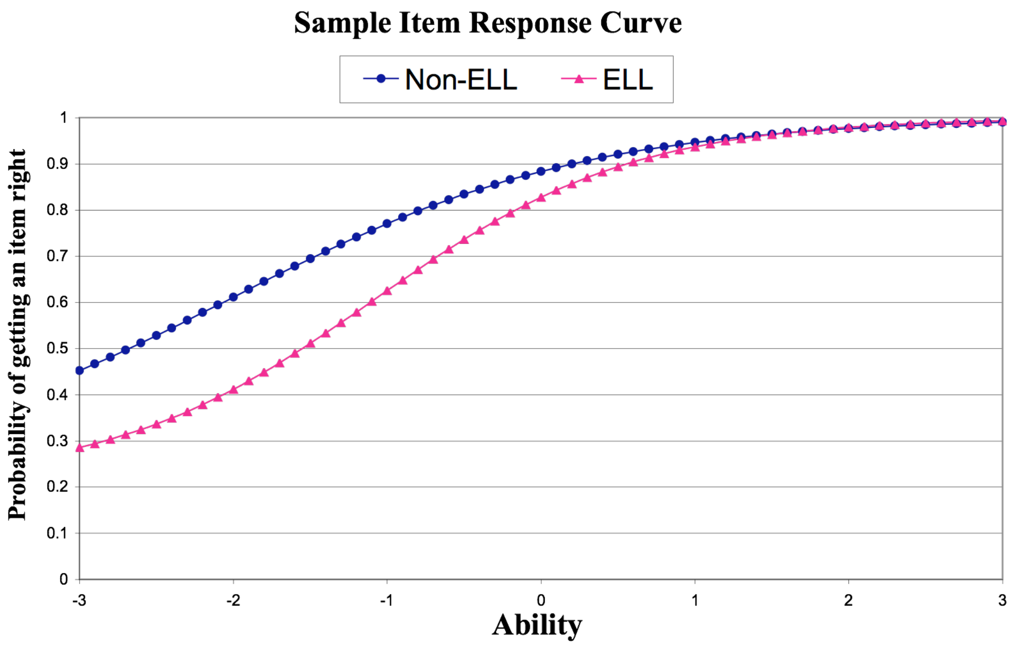

Figure 1. Example DIF Item Response Curve.

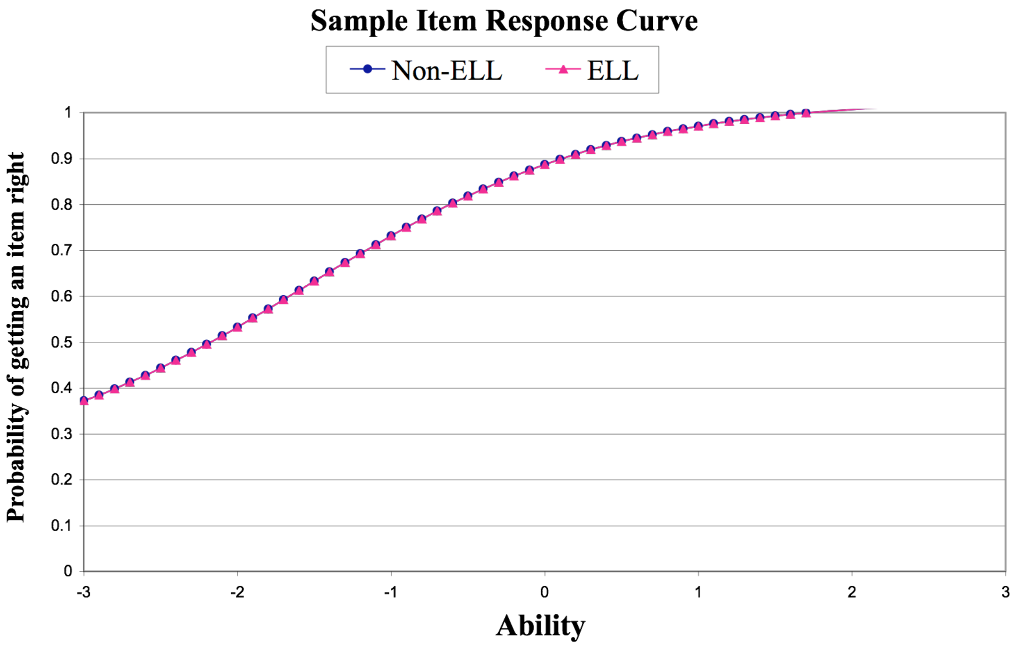

Figure 2. Example non-DIF Item Response Curve.

In this study, the researchers applied the IRT-based likelihood-ratio (IRT-LR) test developed by Thissen (2001), using IRTLRDIF software. The advantage of IRT-LR approach was that it "may detect DIF that arises from differential difficulty, differential relations with the construct being measured ('slopes'), or even differential guessing rates; alternative procedures vary in their effectiveness of DIF other than differential difficulty" (Thissen, 2001, p. 2). Thissen employs the 3-parameter model in item response theory in explaining how an item functions for a given group of students. The three parameters (a, b, and c) represent the function of an item's discrimination, difficulty, and guessing respectively. If there is a substantial difference in any of these 3 parameters between the groups of students being compared, then the item can be identified as exhibiting DIF. In other words, different IRCs for the same ability level groups can occur from an item's different functioning for the two groups in discrimination, difficulty, guessing, or combination of each. In relation to uniform and non-uniform DIF, uniform DIF typically is associated with either the difficulty or guessing parameters, while non-uniform DIF is associated with the discrimination parameter.

Data

Samples were drawn from three states in the analysis. As explained in Chapter 1, the three participating states provided students' scores on their standardized content-area tests including reading, math, and science.[10] Two of the states provided actual test items, which were analyzed for their linguistic structure as described in Chapter 1. The DIF analyses in this report were performed on samples where focal and reference groups were matched on ability (total assessment score). This design allowed for consistent comparisons of the magnitude of DIF effects across the various states, assessments, and grade levels. In general, there were many more students in the original reference groups then in the focal groups. This unbalanced focal to reference group ratio allowed for a matching procedure whereby each focal group student could be randomly matched to a reference group student on the assessment total score without a large loss of unmatched focal group students.

Students with disabilities were removed from the samples so that the results could be interpreted without that potentially confounding factor. Three samples were analyzed (where possible) for each assessment. ELL students served as a focal group in the first sample and were randomly matched with non-ELL students on the ability scale. In the other two samples, sub-groups of ELL students were analyzed. In the second sample, ELL students were

[10] State B did not have a standardized statewide science test at the time of this study. State C did not have a standardized statewide science test for Grade 4 at the time of this study.

grouped based on their reading proficiency level with students in the lower proficiency group serving as the focal group. In the third sample, ELL students were grouped based on their accommodation status whereby accommodated ELL students served as the focal group. In sum, ELL students, low reading proficient ELL students, and accommodated ELL students were used as focal groups in each analysis, assuming that these groups might be more sensitive to the way an item was presented due to their limited English language proficiency. Sample sizes for the analyses varied, according to available data, for each DIF type and test. The sample sizes for each analysis are summarized in Table 1.

Table 1

Summary of the Tests and Sample Sizes

State	Test	Total items	ELL/Non-ELL sample	High/low reading proficient ELL sample	Accommodated/ non-accommodated ELL sample
A	Math Grade 5	46	3,878	843	1,484
A	Math Grade 8	51	2,962	656	614
A	Science Grade 5	44	3,698	766	1,385
A	Science Grade 8	42	2,837	640	591
B	Math Grade 4 Form 1	50	990	NA	316
B	Math Grade 4 Form 2	50	842	NA	276
B	Math Grade 7 Form 1	50	568	NA	193
B	Math Grade 7 Form 2	50	544	NA	217
B	Math Grade 8 Form 1	60	795	210	303
B	Math Grade 8 Form 2	60	783	226	290
C	Math Grade 4	54	4,310	1,068	1,536
C	Math Grade 8	45	2,122	660	662
C	Science Grade 8	60	2,122	528	655

Note. NA = Analysis not performed due to insufficient sample size available.

Procedure

As mentioned earlier, IRT-LR was the primary method used to detect DIF in this report. Logistic regression and MH techniques were also employed to supplement the results obtained from IRT-LR. When sample sizes in each matched group (reference and focal) were greater than or equal to 500, then the IRT-LR method was used alone to determine whether a given item would be identified as DIF. If the matched group sample size was less than 500, a combination of multiple methods reaching DIF thresholds was required for DIF identification. Matched samples with fewer than 180 students in each group were considered too small to produce reliable results and therefore were not analyzed.

For each of the three methods, DIF criteria included both a test of significance and a measure of effect size to ensure that the differential functioning was of substantial magnitude. All three methods employ the chi-square test to measure significance. In both the IRT-LR method and logistic regression, the chi-square test is based on a likelihood ratio, and the PHI r measure is used to measure magnitude of effect. When reference and focal groups are equal in size, as they were in this report, the PHI r statistic will provide a consistent measure of effect size. Guidelines for detecting small to moderate, or moderate to large, DIF were set at PHI r levels of 0.10 and 0.15 respectively.

Identified DIF items were further analyzed for their linguistic complexity in order to examine possible causes of DIF. The linguistic complexity was measured using the content analysis protocol as described in Chapter 1. Selected linguistic features, which the linguistic content analysis found to be salient were examined for both DIF and non-DIF items. To measure the length of the items, the number of total words per item, the number of unique content (i.e., nouns, verbs, etc.) words, and the number of sentences were selected. The three types of academic vocabulary (i.e., general, context-specific, and technical academic vocabulary) were also included. Grammar and cohesion were calculated as the sum of all grammatical and cohesive devices, respectively, due to the relatively small amount of unique grammatical and cohesive features present in items. Finally, DIF and non-DIF items were compared on their *Form* (the proportion of language to non-language) and *Reliance* (the extent to which the language knowledge was needed to get an item correctly) ratings (see Chapter 1 for the description of these categories).

Results

In this section, the researchers present the DIF items identified using the IRT-LR method. Subsequently, the researchers present the characteristics of linguistic complexity of the DIF items compared to the non-DIF items.

DIF Items Across States' Tests

A summary of items detected as DIF across the three states for each of the three matched samples is displayed in Table 2. Detailed DIF statistic results are presented in Appendix CH2, including item parameter values contributing to DIF, as well as the alternative MH and logistic regression method results.

Table 2

Differential Item Functioning Detection across States Tests and Samples

			Number of items detected as DIF (percentage of DIF items)		
State	Test	Total items	ELL/non-ELL sample	High/low reading proficient ELL sample	Accommodated/non-accommodated ELL sample
A	Math Grade 5	46	1 (2%)	4 (9%)	0 (0%)
A	Math Grade 8	51	2 (4%)	8 (16%)	5 (10%)
A	Science Grade 5	44	5 (11%)	8 (18%)	1 (2%)
A	Science Grade 8	42	3 (7%)	8 (19%)	4 (9%)
B	Math Grade 4 Form 1	50	1 (2%)	NA	1 (2%)
B	Math Grade 4 Form 2	50	5 (10%)	NA	2 (4%)
B	Math Grade 7 Form 1	50	4 (8%)	NA	6 (12%)
B	Math Grade 7 Form 2	50	7 (14%)	NA	17 (34%)
B	Math Grade 8 Form 1	60	4 (7%)	7 (12%)	7 (12%)
B	Math Grade 8 Form 2	60	4 (7%)	10 (17%)	5 (8%)
C	Math Grade 4	54	1 (2%)	1 (2%)	1 (2%)
C	Math Grade 8	45	1 (2%)	1 (2%)	1 (2%)
C	Science Grade 8	60	1 (2%)	3 (5%)	2 (3%)

Note. NA = When the sample size was not sufficient, the analysis was not conducted.

As shown in Table 2, most of the variation in number of DIF items occurred across states, rather than between grade levels or subject matter. In comparison to the other states, State C had almost no DIF items, while State B had as many as 34% of items flagged for DIF in Grade 7 on a particular form. Within State A, science had more DIF items than math. There was little difference between the number of DIF items from lower to higher grades.

The majority of DIF items were found in the grouping analyzed by reading proficiency level (50 DIF items) and the grouping analyzed by accommodation provision (52 DIF items).

Perhaps surprisingly, the ELL/non-ELL comparison produced only 39 of the DIF items identified. For the ELL/non-ELL comparison groups, the average number of DIF items per test was about the same in both the math and science content tests (3 items per test). Similarly, across grades in ELL and non-ELL groups, there was an average of about 3 DIF items per test in Grades 7 and 8 (26 items total) and 3 DIF items per test in Grades 4 and 5 (12 items total). In State B, the number of DIF items varied greatly between forms (e.g., Grade 4 math Form 1 had 1 DIF item and Form 2 had 5 DIF items).

On average, there were fewer DIF items per test on math content tests (6 DIF items per test), than science (about 8 DIF items per test) for groups separated by reading proficiency level. Similar to the ELL and non-ELL group analysis, the number of DIF items varied greatly between forms (e.g., Grade 8 math Form 1 had 7 DIF items and Form 2 had 10 DIF items) for reading proficiency grouping analysis.

For groups separated by accommodation, similar patterns emerged. In State A, again Grade 8 contained more DIF items (5 items) than Grade 5 (0 items). Additionally, the greatest variation in the number of DIF items was between forms (e.g., Grade 7 math Form 1 had 6 DIF items and Form 2 had 17 DIF items).

Linguistic Features of DIF Items

Due to the small number of DIF items present, results were aggregated across all subjects, grades, and states in order to present a simplified and possibly more generalizable comparison. Table 3 compares non-DIF items versus DIF items against the focal group (ELLs), DIF items against the reference group (non-ELLs), and non-uniform DIF items, respectively. A smaller number of DIF items are listed in Tables 3, 4, and 5 than in Table 2 earlier because linguistic analysis described in Chapter 1 was conducted on only one form of each Grades 4 and 7 State B math tests. The current DIF analysis was conducted on two forms for each grade.

Table 3

ELL vs. Non-ELL DIF Items by Linguistic Features: Mean (*SD*)

Mean linguistic features	Non-DIF items (*N* = 379)	DIF against ELL items (*N* = 14)	DIF against non-ELL items (*N* = 7)	Non-uniform DIF items (*N* = 3)
# of total words	29.64 (15.33)	40.21 (25.30)	28.14 (14.37)	17.67 (8.39)
# of unique content words	13.80 (7.20)	19.00 (13.03)	13.00 (3.37)	7.67 (4.62)
# of sentences	2.52 (3.31)	3.14 (2.41)	1.86 (0.69)	1.67 (1.16)
# general academic vocabulary	1.57 (1.66)	3.00 (2.42)	2.14 (1.07)	1.33 (1.53)
# content vocabulary	0.83 (1.07)	1.07 (1.73)	0.57 (0.79)	0.00 (0.00)
# technical vocabulary	1.29 (1.70)	1.21 (1.48)	2.29 (2.56)	0.67 (0.58)
Total academic vocabulary words	3.69 (2.66)	5.29 (2.53)	5.00 (3.06)	2.00 (2.00)
Total academic grammar features	1.56 (1.85)	1.93 (1.73)	3.29 (2.63)	1.00 (0.00)
Cohesion	2.25 (2.54)	3.64 (3.48)	2.29 (3.50)	1.00 (1.73)
Non-simple sentences	0.55 (0.72)	0.79 (1.05)	0.43 (0.53)	0.00 (0.00)
Form	1.87 (0.64)	2.50 (0.65)	1.71 (0.95)	1.67 (0.58)
Reliance	2.79 (1.10)	2.93 (0.62)	1.43 (0.98)	2.33 (2.08)

Note. N = Number of items.

Generally speaking, the DIF items against ELL students contained substantially more features of academic English than either the non-DIF items or the DIF items against non-ELL students. This can be seen in the much higher number of total words, unique content words, sentences, and in the larger number of cohesive devices, as well as the overall form. Interestingly, there are substantially more general academic words and academic content words in these items, but not more technical words. The DIF items against non-ELL students were similar overall to the non-DIF items in most respects, except for the markedly higher number of technical academic vocabulary and grammar features.

The researchers provide an example that illustrates this finding. Figure 3 shows the IRC of a prototypical DIF item. This item was from State A Grade 5 science, and was identified as DIF against ELL students. The graph shows the differences in IRC curves between ELL students and non-ELL students when the guessing and difficulty parameters (a and c respectively) are constrained equally; ELL students of equal overall ability have a lower probability of answering this item correctly.

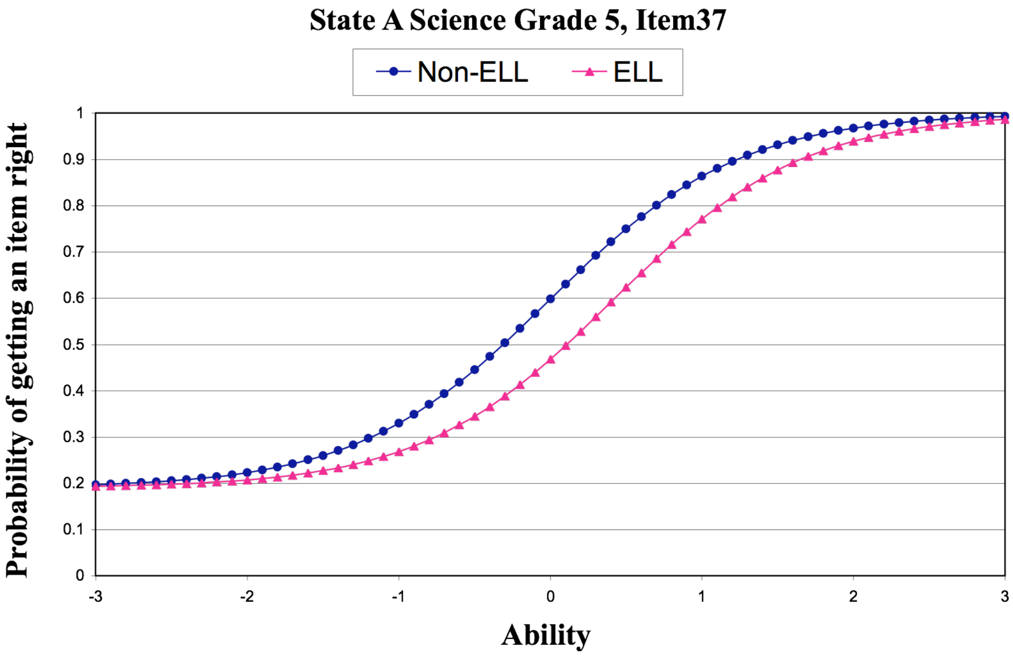

Figure 3. Example of the item characteristic curve of a DIF item against ELL students.

The content analysis of this item indicated that this item contained only language; in other words no numbers or visual features. The item contained a total of eight academic vocabulary words (1 general, 6 content, and 1 technical) out of only 17 unique content words. In other words, almost half the words on this item were rated as academic. The item was rated a 3 on *Reliance* on language, indicating that language processing at the sentence level was necessary to understand the item. It did not contain academic grammar features or an inordinate number of cohesive devices. It may be that a heavy academic vocabulary load contributed substantially to DIF on this item.

Table 4 summarizes results for high-reading versus low-reading ability ELL students.

Table 4

High ELLs v. Low ELLs Items by Linguistic Features: Mean (*SD*)

Mean linguistic features	Non-DIF items (*N* = 361)	DIF against low ELL (*N* = 14)	DIF against high ELL (*N* = 21)	Non-uniform DIF items (*N* = 7)
# of total words	30.03 (16.20)	30.79 (10.03)	22.71 (8.14)	42.43 (14.07)
# of unique content words	13.94 (7.50)	12.43 (8.10)	12.48 (6.05)	20.14 (5.87)
# of sentences	2.55 (3.33)	2.43 (1.40)	2.14 (3.21)	2.57 (0.54)
# general academic vocabulary	1.62 (1.70)	1.86 (1.23)	1.43 (1.08)	2.14 (3.53)
# content vocabulary	0.83 (1.08)	0.86 (1.35)	0.52 (0.68)	1.71 (1.60)
# technical vocabulary	1.27 (1.71)	1.14 (1.61)	1.86 (1.65)	1.71 (1.89)
Total academic vocabulary words	3.72 (2.69)	3.86 (2.48)	3.81 (2.23)	5.57 (3.41)
Total academic grammar features	1.58 (1.88)	1.79 (2.04)	1.67 (1.91)	1.57 (0.79)
Cohesion	2.33 (2.68)	2.50 (1.95)	0.86 (0.85)	3.71 (1.60)
Non-simple sentences	0.57 (0.74)	0.50 (0.65)	0.38 (0.59)	0.71 (0.95)
Form	1.86 (0.63)	2.36 (0.63)	1.95 (0.92)	2.00 (0.00)
Reliance	2.76 (1.12)	3.00 (0.68)	2.52 (0.98)	3.71 (0.76)

Note. *N* = number of items.

The patterns for this type of DIF for these two groups were less clear. Generally, the DIF items against low-reading ELL students had a very similar academic English profile to the non-DIF items, except for increased form and an increase in total academic grammar features. The DIF items against high-reading ELL students included slightly fewer features of academic English than the non-DIF items; in particular the lower number of total words and much lower number of cohesive devices. However, the use of technical vocabulary was higher for these items, which is similar to the findings for the DIF items against non-ELL students discussed earlier. In this case, the non-uniform DIF items, of which there were only six, had much more language than any other type of item, in contrast to the non-uniform DIF items in the ELL vs. non-ELL grouping discussed earlier.

Finally, Table 5 shows results of the DIF analyses for accommodated versus non-accommodated ELL students.

Table 5

Accommodated ELLs vs. Non-Accommodated ELLs Items by Linguistic Features: Mean (*SD*)

Mean linguistic features	Non-DIF items (*N* = 375)	DIF against accommodated items (*N* = 12)	DIF against non-accommodated items (*N* = 14)	Non-uniform DIF items (*N* = 2)
# of total words	30.07 (16.10)	26.92 (13.92)	28.14 (10.79)	26.00 (5.66)
# of unique content words	13. 91 (7.58)	13.75 (5.90)	13.86 (6.06)	16.00 (4.24)
# of sentences	2.56 (3.35)	1.75 (0.87)	2.29 (0.99)	1.50 (0.71)
# general academic vocabulary	1.64 (1.73)	1.92 (1.24)	1.07 (1.21)	1.50 (2.12)
# content vocabulary	0.83 (1.08)	1.00 (1.21)	0.71 (1.38)	1.00 (0.00)
# technical vocabulary	1.27 (1.65)	1.17 (0.84)	2.21 (3.22)	1.50 (0.71)
Total academic vocabulary words	3.74 (2.65)	4.08 (1.88)	4.00 (3.90)	4.00 (2.83)
Total academic grammar features	1.56 (1.74)	3.25 (4.07)	1.21 (1.72)	1.50 (2.12)
Cohesion	2.34 (2.62)	1.33 (1.97)	1.93 (2.53)	1.00 (0.00)
Non-simple sentences	0.56 (0.73)	0.50 (0.80)	0.50 (0.65)	0.00 (0.00)
Form	1.89 (0.65)	1.83 (0.58)	1.86 (0.66)	2.50 (0.71)
Reliance	2.77 (1.10)	2.92 (1.00)	2.71 (1.38)	2.00 (0.00)

Note. *N* = number of items.

For these two groups, fewer linguistic differences in academic English were found among the DIF items. The DIF items against accommodated ELL students had many more grammatical features, yet less cohesion than did the non-DIF items. The most notable differences occurred in the area of technical vocabulary, where the DIF items against non-accommodated ELL students had more technical words than did the non-DIF items. The non-uniform DIF items were quite similar to the non-DIF items in terms of features of academic English. Interpreting these findings is quite problematic because information regarding the types of accommodations and amount they were used was not available.

Discussion

The focus of the present DIF analyses was to determine if and to what extent test items might exhibit DIF for groups of ELL students. DIF analyses were performed in three states on math and science tests for multiple grade levels. The research team compared ELL students to non-ELL students, and within the ELL population the team also compared students grouped by their reading proficiency as well as by their usage of accommodations. In this study the researchers were also able to classify DIF items into types based on the three IRT model parameters (discrimination, difficulty, and guessing). The number of items identified as DIF varied depending on the state, test form where available, content area, sample group, and grade level analyzed.

Patterns of DIF Items

A notable variation in the number of DIF items was found across states, rather than between content areas or grade levels. There was less DIF exhibited in State C than in the other two states on both the math and science tests in Grade 8. In general there were slightly more DIF items identified on the science test than on the math test when both tests were analyzed in the same state and grade. This may be due to the somewhat more complex nature of the language on the science tests, and the somewhat greater overall amount of language on the science tests. In math content areas, the level of linguistic complexity was lower than the level of complexity in science. The exception to this general finding was in State A Grade 8 where the number of DIF items identified was similar on the science and math tests, even though there was substantially more language, both academic and overall, on the science tests.

In terms of grade-level, no clear pattern was found for the ELL and non-ELL grouping. However, in general, more items were flagged as DIF in higher grades than in the lower grades for the high/low reading proficient ELL grouping and the accommodated/non-accommodated ELL grouping. For example, in the State A Grade 5 accommodation sample, the researchers identified just one DIF item on the science test and no (zero) DIF items on the math test. By comparison, in the State A Grade 8 sample, the researchers identified four DIF items on the science test, and five DIF items on the math test. This trend was also true for the sample that compared accommodated ELL students to non-accommodated ELL students in State B. In the State B Grade 4 accommodation sample, the researchers identified just one DIF item on the math test for Form 1, and two DIF items on the math test for Form 2. In comparison, in the State B Grade 7 sample, the researchers identified six DIF items on the math test for Form 1, and seventeen DIF items on the math test for Form 2. As was found in the content analysis reported in Chapter 1, the language demands of the tests were somewhat greater in the higher grades, and this may have been at least partially responsible for the increased number of DIF items.

There were also more items in the samples of ELL students grouped by reading proficiency identified as DIF than in the samples grouped by ELL students vs. non-ELL students. This finding suggests that substantial variation is present within the ELL population and that ELL students in the lowest reading proficiency groups may be more vulnerable to nuisance factors unrelated to the test construct, in particular questions of opportunity to learn (OTL) the content. This may be true even though the DIF items against low-reading ELLs did not appear much different linguistically than the non-DIF items.

As indicated earlier, in this study the researchers were also able to classify DIF items into types based on the three model parameters (discrimination, difficulty, and guessing). Access to information on the type of DIF is important to determine why an item may function differently for two groups of students. While DIF items were detected in each parameter type, there were a substantial number of items detected as DIF primarily due to differences in the two groups of students' ability to guess correctly when the answer was very unlikely to be known. For example, in State A on the Grade 5 science test in the reading proficiency sample there were 8 items identified as DIF, and 6 of the DIF items were identified due to differences in the guessing parameter. Further inspection of these items may provide valuable information to help students with test taking strategies for specific types of items.

When an item exhibits DIF, it may be an indication that item bias or nuisance factors unrelated to the test construct could be affecting differential group performance. In this study the research team show that there was variation in the number of DIF items identified based on the state, content area, sample group, and grade level analyzed. It should be noted that DIF identification alone is not proof of bias. Items exhibiting DIF were further studied by type of DIF and the linguistic content of the items in order to aid in explaining the results.

DIF Items and Linguistic Analyses

The linguistic analyses do offer some intriguing possibilities for explaining the DIF findings and patterns. The most straightforward finding is that the items which exhibited DIF against ELL students had more academic vocabulary across grades and subjects, thus suggesting that it is often, or primarily, increased linguistic complexity which causes DIF for ELL students. However, if some of this academic vocabulary, that is, context-specific and technical vocabulary, is part of the knowledge students are expected to learn in their math and science classes, then it may be that the DIF items with context-specific or technical vocabulary may actually be measuring students' knowledge of math and science. Thus, rather than inadequate English proficiency being a source of bias, ELL's low performance on these items may be more directly related to inadequate OTL. That is, some ELL students' low English proficiency may be an obstacle not so much to performing well on the content assessment, as to acquiring the necessary academic vocabulary in the first place. Low-reading versus high-reading proficient ELL students did not show such a clear pattern. Future DIF studies may thus wish to compare low-reading proficient ELL students against non-ELL students as the reference group and examine how linguistic complexity varies across the DIF items.

The differences between the types of vocabulary and their frequency across DIF categories were especially interesting. The DIF items identified as being biased against ELL students had almost double the number of academic general vocabulary items (such as *furthermore* or *substantial*) and substantially higher academic content vocabulary as well. This suggests that the lack of academic vocabulary knowledge may be a key explanatory variable in causing DIF, more so than lack of grammatical or other types of linguistic knowledge. In contrast, DIF items identified as against both non-ELL and high-reading proficient ELLs had much more of the technical (highly specialized) academic vocabulary such as *square root* and *geothermal*. Students are generally exposed to words such as these only in classroom settings. This, along with the lower total number of words in these items, may indicate that the highly technical vocabulary is an aspect of content knowledge with which these students, who are hypothesized to have less trouble with academic vocabulary, are unfamiliar. In other words, other factors such as OTL the academic content itself may be a better explanation for DIF than language knowledge for these students.

Item DIF and Accommodations

Results for accommodated versus non-accommodated students may have been complicated by the fact that accommodations were not randomly assigned and that the precise accommodation(s) given were unknown. Still, the higher amount of technical vocabulary for the items, which went against the non-accommodated ELL students, suggest that accommodations that change the language of the test input in some way may help reduce language-related test score variance for accommodated students. In other words, for example, the items on which students without dictionaries (or other accommodations) did badly may have had more technical vocabulary. In general, if the DIF favors the accommodated group, this could be a possible indication that the accommodation may have changed the construct being measured. For example, on an item asking a student for the meaning of a particular science vocabulary item such as *geothermal*, students using a dictionary may only have to look up the word to get it correct, thus changing the construct substantially. An alternative explanation may be that the accommodation was effective in removing nuisance factors from the accommodated group without affecting the construct being measured. For example, a glossary may allow accommodated ELL students access to test content, thus decreasing the relative difficulty of the item for that group. However, without knowing which specific accommodations were given to which ELLs, the relationship between language demand, DIF, and type of accommodations cannot be determined. Nevertheless, a careful review of the DIF items' content, including an analysis of the academic language and language demand of these items, do provide a basis for formulating hypotheses about how the language demand of the items and the type of accommodation provided may interact to produce DIF.

The items which exhibited non-uniform DIF did not exhibit a clear pattern of linguistic differences from the non-DIF items, and as there were relatively few, it may be best not to read too much into these results. Similarly, items with combined DIF did not exhibit a clear pattern with respect to linguistic features. However, as there were relatively small numbers of such items, it is difficult to reach any generalizable interpretations.

To summarize, although these DIF findings should be cautiously interpreted due to some limitations of the study, results do suggest that language factors may play a substantial role in causing DIF for ELL students, and thus may constitute a source of item bias. Furthermore, low-performing ELL students may be especially sensitive to item bias. This is because these students are more likely to be recent arrivals to the U.S., and may therefore be less attuned to necessary academic culture and/or classroom factors, which could cause item bias in addition to, or in combination with, linguistic content factors. A future study should examine this low-performing ELL group and compare performance against non-ELL students. The presence of substantial DIF both for and against accommodated ELL students, when compared to non-accommodated students, is potentially problematic and deserves greater attention. However, the lack of information as to accommodation type given to individual ELL students weaken any interpretations the researchers may make as the result of this study.

Limitations

There were some limitations in this study, of which the small sample is one. In these analyses, sample size was a major consideration in conducting analyses by subgroups within the ELL population. When focal group sample sizes are smaller than 500 subjects, the research team addressed this limitation to some extent by requiring DIF findings to be confirmed with multiple methods. It is also important to acknowledge that if systemic bias against ELL students or ELL subgroups was present across items in general, this type of bias would not be identified by the DIF techniques used in this study. For example, if nearly all items contain content which causes bias, DIF analyses will not be able to separate out DIF items properly. Since nearly all the items on all the tests which were analyzed contained some degree of academic language, and depended on some degree of language processing, this may be a major limitation for DIF analyses for linguistic bias. Therefore, the results of these analyses should be interpreted with caution. This also illustrates a potentially important methodological issue for future DIF research looking at linguistic sources of bias as opposed to gender or ethnic/cultural bias analyses.

References

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J. & Gandara, P. (2006, December). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practices, 26*(5), 36–46.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CRESST Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

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

Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups on the Test of English as a Foreign Language (TOEFL Research Rep. No. 9; ETS Research Rep. No. 81–16). Princeton, NJ: Educational Testing Service.

Allen, N. L., & Donoghue, J. R. (1996). *Detecting differential item functioning: Current methods and continuing problems*. Princeton, NJ: Educational Testing Service.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405–438.

De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77*(4), 460–470.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (Rep. No. ETS-RR-83-9). Princeton, NJ: Educational Testing Service, Research Publications.

Dorans, N. J., & Kulick, E. (1986, Winter). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355–368.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54*, 84–90.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math— not reading—on a math assessment: A language accommodations study of English language learners and other special populations*. Presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Loyd, B. (1984, April). *Evaluation of log linear models for detection of item bias: A comparison across samples*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES Publication No. 2000-473). Washington, DC: National Center for Education Statistics.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105–118.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.

Scheuneman, J. D. (1975, April). *A new method of assessing bias in test items*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED106359)

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143–152.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometricka, 58*, 159–194.

Spray, J. A., & Carlson, J. E. (1988). *Comparison of loglinear and logistic regression model for detecting changes in proportions* (Research Rep. No. 88-3). Iowa City, IA: American College Testing.

Thissen, D. (2001). IRTLRDIF v.2.02b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer Software]. Chapel Hill, NC: LL Thurstone Psychometric Laboratory.

Appendix CH2

Tables A1–A2
The Results of DIF Analysis: State A

Table A1

STATE A Math Grade 8 Differential Item Functioning Detection Method Comparison

Item	Matched P-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH-DIF
ELL (focal) and non-ELL (reference) students								
	Non-ELL *N* = 2,962	ELL *N* = 2,962	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
6	0.839	0.643	0.020	0.094	0.231	0.246	.041	-2.92
45	0.420	0.297	0.068	0.126	0.032	.140	.014	-1.52
Low ELL (focal) and high ELL (reference) reading proficient students								
	ELL Level (4, 5, +) *N* = 656	ELL Level (1, 2, 3) *N* = 656	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
6	0.646	0.399	0.029	0.086	0.240	0.266	0.011	-2.69
7	0.648	0.502	0.009	0.030	0.152	0.150	0.025	-1.49
45	0.250	0.166	0.080c	0.052c	0.069c	0.111	0.032	-1.27
52	0.282	0.424	0.000	0.089	0.122	0.144	0.038	1.55
53	0.250	0.351	0.082c	0.072c	0.030c	0.105	0.016	1.16
54	0.402	0.520	0.028	0.122	0.000	0.112	0.051	1.18
65	0.357	0.401	0.073	0.101	0.000	0.039	0.100	0.46
66	0.349	0.270	0.061c	0.054c	0.073c	0.091c	0.041c	-0.89
Accommodated (focal) and non-accommodated (reference) ELL students								
	Non-accom *N* = 614	Accom *N* = 614	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
6	0.674	0.526	0.041	0.058	0.154	0.162	0.048	-1.67
26	0.277	0.199	0.000	0.000	0.100	0.098	0.006	-1.06
40	0.671	0.774	0.065c	0.092c	0.000c	0.123	0.041	1.33
52	0.311	0.409	0.000	0.065	0.101	0.108	0.025	1.06
53	0.283	0.358	0.044c	0.096c	0.020c	0.084	0.042	0.85

Table A2

STATE A Science Grade 8 Differential Item Functioning Detection Method Comparison

Item	Matched *P*-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH-DIF
ELL (focal) and non-ELL (reference) students								
	Non-ELL *N* = 2,837	ELL *N* = 2,837	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non-uniform	
3	0.649	0.514	0.014	0.046	0.122	0.144	0.002	-1.44
11	0.441	0.531	0.021c	0.064c	0.092c	0.090	0.010	0.90
31	0.529	0.615	0.035c	0.035c	0.094c	0.086	0.019	0.86
Low ELL (focal) and high ELL (reference) reading proficient students								
	ELL Level (4, 5, +) *N* = 640	ELL Level (1, 2, 3) *N* = 640	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non-uniform	
2	0.809	0.653	0.000	0.063	0.102	0.186	0.012	-2.17
8	0.459	0.295	0.071	0.031	0.151	0.173	0.028	-1.76
11	0.370	0.450	0.020	0.144	0.025	0.075c	0.069c	0.72
31	0.481	0.597	0.109	0.081	0.000	0.114	0.021	1.04
34	0.348	0.411	0.101	0.009	0.022	0.061	0.024	0.54
36	0.291	0.442	0.000	0.000	0.170	0.157	0.004	1.45
39	0.313	0.323	0.064	0.115	0.054	0.007	0.077	0.03
41	0.214	0.256	0.123	0.023	0.000	0.047	0.080	0.40
Accommodated (focal) and non-accommodated (reference) ELL students								
	Non- accom *N* = 591	Accom *N* = 591	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non-uniform	
31	0.497	0.594	0.097c	0.000c	0.053c	0.100	0.041	0.96
39	0.348	0.411	0.047c	0.082c	0.044c	0.050	0.053	0.46
48	0.306	0.391	0.009c	0.047c	0.090c	0.090c	0.048c	0.90
57	0.193	0.289	0.041	0.045	0.104	0.114	0.025	1.28

Appendix CH2

Tables B1–B2

The Results of DIF Analysis: State B

Table B1

State B Math Grade 8 Form 1 Differential Item Functioning Detection Method Comparison

Item	Matched *P*-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH-DIF
ELL (focal) and non-ELL (reference) students								
	Non-ELL *N* = 795	ELL *N* = 795	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
P2-2	0.392	0.289	0.066c	0.055c	0.087c	0.113	0.036	-1.10
P2-8	0.702	0.610	0.016c	0.026c	0.095c	0.112	0.011	-1.21
P2-10	0.470	0.365	0.000	0.046	0.112	0.112	0.054	-1.08
P2-30	0.600	0.502	0.008	0.039	0.103	0.104	0.042	-1.00
Low ELL (focal) and high ELL (reference) reading proficient students								
	ELL Level (4, 5, +) *N* = 210	ELL Level (1, 2, 3) *N* = 210	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
P2-14	0.686	0.538	0.015	0.089	0.137	0.145	0.095	-1.38
P2-23	0.324	0.267	0.157	0.000	0.022	0.054c	0.088c	-0.71
P2-35	0.429	0.519	0.086c	0.027c	0.044c	0.105	0.028	1.08
P2-67	0.152	0.276	0.022	0.046	0.153	0.156	0.016	2.08
P2-73	0.324	0.433	0.123	0.076	0.000	0.123	0.034	0.99
P2-75	0.443	0.438	0.115	0.038	0.044	0.009	0.135	0.03
Accommodated (focal) and non-accommodated (reference) ELL students								
	Non-accom *N* = 303	Accom *N* = 303	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
P2-8	0.680	0.587	0.036	0.100	0.000	0.110	0.004	-1.19
P2-18	0.380	0.294	0.048c	0.091c	0.026c	0.095c	0.033c	-1.08
P2-25	0.373	0.502	0.110	0.115	0.000	0.133	0.038	1.39
P2-31	0.465	0.363	0.013	0.113	0.000	0.116	0.042	-1.19
P2-65	0.623	0.712	0.013	0.116	0.060	0.118	0.021	1.12
P2-73	0.399	0.413	0.063c	0.091c	0.079c	0.011	0.125	0.18
P2-79	0.317	0.446	0.091	0.130	0.000	0.140	0.015	1.50

Table B2

State B Math Grade 8 Form 2 Differential Item Functioning Detection Method Comparison

Item	Matched *P*-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH- DIF
ELL (focal) and non-ELL (reference) students								
	Non-ELL *N* = 783	ELL *N* = 783	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
P2-2	0.467	0.590	0.040	0.153	0.000	0.130	0.076	1.31
P2-30	0.640	0.553	0.028	0.042	0.107	0.090c	0.068c	-.91
P2-43	0.501	0.516	0.047c	0.024c	0.091c	0.015	0.108	0.11
P2-51	0.415	0.326	0.000	0.103	0.018	0.096c	0.033c	-0.92
Low ELL (focal) and high ELL (reference) reading proficient students								
	ELL Level (4, 5, +) *N* = 226	ELL Level (1, 2, 3) *N* = 226	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
P2-5	0.168	0.270	0.000	0.000	0.104	0.125	0.032	1.19
P2-7	0.451	0.496	0.092c	0.087c	0.036	0.056c	0.092c	0.36
P2-12	0.305	0.358	0.110	0.000	0.000	0.066c	0.090c	-0.62
P2-17	0.460	0.540	0.039	0.116	0.000	0.094c	0.072c	1.02
P2-25	0.332	0.469	0.056	0.155	0.042	0.150	0.051	1.43
P2-31	0.389	0.350	0.100	0.054	0.026	0.026	0.101	-0.39
P2-34	0.708	0.522	0.049	0.085	0.134	0.185	0.002	-2.31
P2-37	0.381	0.473	0.093c	0.054c	0.021c	0.107	0.033	1.04
P2-52	0.310	0.412	0.076	0.101	0.042	0.120	0.027	1.00
P2-56	0.447	0.319	0.026	0.015	0.134	0.122	0.082	1.51
Accommodated (focal) and non-accommodated (reference) ELL students								
	Non-accom *N* = 290	Accom *N* = 290	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
3	0.417	0.293	0.000	0.127	0.066	0.134	0.046	-1.27
7	0.445	0.534	0.000	0.103	0.019	0.094c	0.049c	0.87
31	0.466	0.369	0.000	0.116	0.000	0.101	0.065	-1.13
32	0.403	0.297	0.035	0.127	0.000	0.115	0.033	-1.33
39	0.241	0.317	0.037	0.104	0.073	0.089c	0.046c	0.81

Appendix CH2

Tables C1–C2

The Results of DIF Analysis: State C

Table C1

STATE C Math Grade 8 Differential Item Functioning Detection Method Comparison

Item	Matched *P*-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH- DIF
ELL (Focal) and Non-ELL (Reference) Students								
	Non-ELL *N* = 2,122	ELL *N* = 2,122	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
44	0.445	0.339	0.005c	0.024c	0.096c	0.104	0.006	-1.04
Low ELL (Focal) and High ELL (Reference) Reading Proficient Students								
	ELL Level (4, 5, +) *N* = 660	ELL Level (1, 2, 3) *N* = 660	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
44	0.379	0.292	0.033	0.000	0.111	0.104	0.010	-1.15
Accommodated (Focal) and Non-Accommodated (Reference) ELL Students								
	Non- accom *N* = 662	Accom *N* = 662	a- Discrimination	b- Difficulty	c- Guess	Uniform	Non- uniform	
13	0.465	0.554	0.009	0.104	0.017	0.103	0.019	1.01

Table C2

STATE C Science Grade 8 Differential Item Functioning Detection Method Comparison

Item	Matched *P*-values		IRT-LR parameter effects Phi (1-1) r			Logistic regression Phi (1-1) r		MH-DIF
ELL (focal) and non-ELL (reference) students								
No DIF items								
Low ELL (focal) and high ELL (reference) reading proficient students								
	ELL Level (4, 5, +) *N* = 528	ELL Level (1, 2, 3) *N* = 528	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
29	0.104	0.176	0.014	0.032	0.102	0.103	0.002	1.30
34	0.186	0.294	0.039	0.042	0.114	0.126	0.003	1.33
38	0.644	0.527	0.000	0.044	0.102	0.129	0.017	-1.20
Accommodated (focal) and non-accommodated (reference) ELL students								
	Non-accom *N* = 655	Accom *N* = 655	a-Discrimination	b-Difficulty	c-Guess	Uniform	Non-uniform	
1	0.299	0.376	0.025c	0.015c	0.096c	0.078	0.055	0.77
8	0.498	0.391	0.107	0.034	0.000	0.111	0.022	-1.10

CHAPTER 3:

INVESTIGATING ELL ASSESSMENT AND ACCOMMODATION PRACTICES USING STATE DATA

Jinok Kim and Joan L. Herman
CRESST/University of California, Los Angeles

Under the No Child Left Behind Act of 2001 (NCLB, 2002), schools are held accountable for the academic performance and progress of all students, including student subgroups such as English Language Learners (ELL) students. The focus on accountability has brought increased attention to the assessment of ELL students. Consequently, many challenging issues have arisen due in part to the uniqueness of the ELL population. While the previous two chapters analyzed test item level data, this chapter focuses on analyzing student level data (i.e., students' scores on the state tests) and aims to provide empirical findings that are concerned with the challenging issues around the assessment of ELL students. Particularly, the research team examines:

1. Achievement gaps among ELL, redesignated ELL, and non-ELL students,
2. ELL students' performance in content-area and English language proficiency (ELP) tests,
3. Characteristics of ELL students who exit the ELL status or who remain as ELL students for an extended period of time, and
4. States' accommodation practices.

States report that generally ELL students do not perform as well as non-ELL counterparts on measures such as academic content-area proficiency tests and exit examinations. However, little in-depth investigation of this achievement gap has been conducted using rigorous statistical methods. For example, the dichotomous division of ELL students and non-ELL students carries only partial information of ELL achievement, since it is important to monitor how redesignated ELL students (students who have exited from ELL status) perform. Many summaries and comparisons focus on the average achievement of ELL students as a group, which may be misleading without paying particular attention to students who improve and transition out of ELL programs or status. More importantly, the average comparisons as a group may overlook the diversity of students within the ELL population. Some ELL students may rapidly improve their ELP and exit from ELL status, while other ELL students may have severe difficulty in improving their ELP and continue as ELL students for relatively longer periods of time. Issues around such students who remain as

ELL students for an extended period are typically not addressed in many group comparisons. However, these long-term ELL students may have the most educational needs within the ELL population, as language barriers that exist for an extended period can severely hinder learning.

In light of these concerns, the researchers examine the expected achievement gaps among different groups (i.e., current ELL students, redesignated ELL students, and non-ELL students), employing rigorous statistical analyses. Furthermore, the researchers examine the factors that are correlated with exiting ELL status compared to ELL students who remain for a relatively longer period of time in our obtained data sets.

Prior research on the academic achievement of redesignated ELL students has yielded mixed results when their academic performance is compared against non-ELL students (Abedi, Leon, & Mirocha, 2003; Bibian, 2006; Stack, 2002). The underlying sources of the conflicting results have been difficult to ascertain. The research team approached their analyses with the hypothesis that one of the reasons for the mixed results might be related to the stringency or leniency of states' redesignation criteria. This hypothesis was examined in the context of the relationship between ELP and content-area assessments. A primary criterion for identification and redesignation of ELL students in this study's three participating states are the scores from assessments measuring students' ELP in the four modalities of speaking, listening, reading and writing. The relationships between the ELP and state content-area assessments may help us examine the stringency of redesignation criteria by estimating how well ELL students who just meet the ELP redesignation criteria tend to do relative to achieving proficiency on content-area tests. Additionally, studying these relationships may provide a way of examining the validity of the ELP test. Positive and strong relationships between ELP and performance on content-area assessments, under certain circumstances, may provide evidence that supports the concurrent validity of interpretations based on the ELP test. While some literature investigates this relationship (e.g., Stevens, Butler, & Castellon-Wellington, 2000), there is little current research, given substantial changes that have occurred to ELP assessment since the NCLB Act (2002).

Lastly, regarding the assessment of ELL students, great attention has been paid to the use of accommodations. To improve the accuracy of content-area assessment outcomes for ELL students, without giving them an unfair advantage relative to non-ELL students, research has attempted to identify accommodations that narrow the assessment performance gap without mitigating the validity of score-based interpretations (Abedi, Courtney, & Leon, 2003; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi, Hofstetter, & Lord, 2004; see Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006 for meta-analysis results). Towards this end, the research team examines use of accommodation strategies specifically for the ELL students in three participating states.

This chapter addresses the following questions:

1. How well do ELL students perform in various subjects of state content-area tests relative to their redesignated ELL and non-ELL peers? Are the patterns of performance among these three groups similar across states? If states show differences, what are the underlying sources of the differences in the patterns of performance?
2. How do differences in scores on state content-area tests relate to differences in scores in ELP tests? Do the relationships vary across schools? Do the relationships vary across content areas, grade levels, or states? Also, based on the relationships between ELP and content-area tests, how well do ELL students who just meet the ELP redesignation criteria tend to perform relative to achieving proficiency on state content-area tests?
3. What are the correlates of redesignation? What are the characteristics of students who are redesignated compared to students who continue as ELL students for an extended period of time (i.e., more than three years)?
4. What kind of accommodations and how frequently do states use these when assessing ELL students? What criteria are used to identify the types of accommodations to give ELL students?

Methods

In this section, the research team describes the data and provides an overview of primary methods employed for analyses. Because different types of models were constructed for each research question, depending on the available data for each state, additional details about the analyses are also presented in the results section following each research question.

Data

The data included students' scores in content-area tests for the 2005–06 academic year. Scores for 2 to 3 grade levels were obtained from each state, one from lower grades (4 or 5) and the other(s) from upper (7 or 8) grade levels. Table 1 shows the distribution of ELL students in the target grades for the three participating states for which data were collected.

Table 1

Percentage of ELL Students by Grade

State	Student status	Grade 4		Grade 5		Grade 7		Grade 8	
		N	Percent	*N*	Percent	*N*	Percent	*N*	Percent
A	Non-ELL			24,380	73.34			25,180	75.58
	RFEP[a]			3,854	11.59			4,304	12.92
	ELL			5,008	15.07			3,833	11.50
	Total			33,242	100.00			33,317	100.00
B	Non-ELL	70,511	91.41			75,404	94.10	83,956	94.77
	RFEP	2,539	3.29			2,160	2.70	2,079	2.35
	ELL	4,086	5.30			2,565	3.20	2,554	2.88
	Total	77,136	100.00			80,129	100.00	88,589	100.00
C	Non-ELL	47,467	85.19					52,299	89.92
	RFEP	2,020	3.63					2,355	4.05
	ELL	6,230	11.18					3,510	6.03
	Total	55,717	100.00					58,164	100.00

[a] Redesignated Fluent English Proficient.

As shown in Table 1, all three states showed substantial percentages of both current and redesignated ELL students, with lower grades having more current ELL students than upper grades. In States A and B about half of all the Grade 7 and 8 ELL students were current ELL students with the other half being redesignated ELL students, while in Grades 4 and 5 there were more current ELL students than redesignated ELL students. In State C, more of the ELL students were current than redesignated.

Consistently across the states, a much larger percentage of ELL students received Free or Reduced Lunch (FRL) as compared to non-ELL students. In State A, approximately 80% of ELL students received FRL, while about 40% of non-ELL students received it. The other states were similar, with somewhat smaller percentages among non-ELL students (about 35% in State B, and about 25% in State C). Given that FRL status is often used as an indicator for the socio-economic status of students' families, these results suggest that ELL students are more than twice as likely to come from socioeconomically disadvantaged families, which itself is often related to academic performance. Thus, in addition to linguistic and cultural differences, many ELL students have the same difficulties that non-ELL students from lower socioeconomic status families have.

Overview of Analysis Methods

Primarily hierarchical model (HM) techniques were employed to address the first two research questions. First, in order to estimate the average achievement levels of and gaps between ELL students, redesignated ELL students, and non-ELL students in state content-area assessments while controlling for student eligibility for FRL, the researchers used 2-level HMs, in which students were nested within schools. Separate analyses were conducted for each state, content area, and grade level. Second, the researchers investigated the relationships between the scores (or levels) of ELP and content-area assessments in reading, math, and science in an effort to provide empirical evidence of the validity of the ELP assessment and/or the validity of the redesignation criteria based on the ELP assessment. To estimate these relationships, the researchers used 2-level HMs, in which students were nested within schools. The use of HMs helped yield more accurate inferences on the parameter of interest (i.e., the average achievement levels/gaps and the relationship between ELP and content-area assessments), by taking into account the intra-class correlations among students within a school and at same time by controlling for some key covariates such as FRL status or interaction between the FRL and the ELP scores. Third, to study continuing or long-term ELL students (i.e., students who remain as ELL students for an extended period), the research team examined the proportion of the long-term ELL population and the correlates of redesignation among ELL students, in contrast to students who may continue as ELL students for more than three years. For the latter, the research team used logistic regression to predict the binary indicator of redesignation (i.e., redesignation in contrast to remaining as ELL students for more than three years).

Results

In this section, additional details on the analysis method and the results are described by each research question listed previously. The researchers focus on summarizing key findings and patterns across states here. Detailed results from all sets of analyses are presented in Tables A–D in Appendix CH3 (pp. 107–138).

Examining Achievement Levels and Gaps

The researchers first examine the expected achievement levels of and gaps among ELL students, redesignated ELL students, and non-ELL students in content-area assessments. To examine states' policies for monitoring redesignated ELL students for 2 years, the research team further divided redesignated ELL students into two groups: students who were redesignated in recent years (i.e., within one to two previous grades) and students who were redesignated earlier (i.e., prior to two previous grades).

Analysis method. The research team used 2-level HMs, in which students were nested within schools, to estimate the average achievement levels and gaps in various content areas of state tests, controlling for student eligibility for FRL. It was important to consider student FRL status, since ELL students and non-ELL students are substantially different in terms of the percentage of students receiving FRL, as previously discussed. The estimated gaps that do not account for FRL status are seriously confounded, which means that it is uncertain whether the gaps are the results of ELL students' average socioeconomic disadvantages or limited English proficiency.

Separate HM analyses were conducted for each state, grade level, and content area. HM yields relatively accurate inferences concerning key parameters by producing accurate standard errors. This contrasts with other techniques that do not take the nested structure of the data into account. HMs also partition the variability in outcomes into two levels, student and school levels, which enables us to examine to what extent the key within-school parameters (i.e., average levels or gaps) vary across schools. The HMs employed are as follows:

Student-level (Level-1)

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{ELL})_{ij} + \beta_{2j}(\mathrm{ExitMonitor})_{ij} + \beta_{3j}(\mathrm{Exit})_{ij} + \beta_{4j}(\mathrm{FRL})_{ij}, \qquad r_{ij} \sim N(0, \sigma^2)$$

School-level (Level-2)

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \qquad u_{1j} \sim N(0, \tau_{11})$$

$$\beta_{2j} = \gamma_{20} + u_{2j} \qquad u_{2j} \sim N(0, \tau_{22})$$

$$\beta_{3j} = \gamma_{30} + u_{3j} \qquad u_{3j} \sim N(0, \tau_{33})$$

$$\beta_{4j} = \gamma_{40} + u_{4j} \qquad u_{4j} \sim N(0, \tau_{44})$$

Equation 1. Two-level HM for estimating expected achievement levels of and gaps among ELL students, redesignated ELL students, and non-ELL students in content-area test achievement.

The outcome Y_{ij} was the achievement score in reading, math, or science in state exams of student i in school j. All predictors were binary indicators: *ELL* was coded as 1 when a student was an ELL and as 0 otherwise; *ExitMonitor* was coded as 1 when a student was recently redesignated and still monitored; *Exit* was coded as 1 when a student was redesignated more than 2 years ago and no longer monitored; and *FRL* was coded as 1 if a student was eligible for or receives FRL.

The parameters at level 1 represented levels or differences in the outcome within school *j*. The intercept β_{0j} captured the expected achievement of non-ELL students who did not receive FRL in school *j*; β_{1j} captured the expected difference or gap between ELL students and Non-ELL students in the outcome in school *j* controlling for whether or not the student was receiving FRL; β_{2j} captured the expected difference between recently redesignated students and non-ELL students in school *j*; β_{3j} captured the expected difference between redesignated students and Non-ELL students in school *j*; and β_{4j} estimated the expected decrement in achievement associated with students who were receiving FRL in school *j*.

At level 2, these within-school parameters were posed to vary across schools. The extent to which each parameter varied across schools was captured by the associated variance components, τ_{00} to $\tau_{44.}$

Summary of results. Separate HMs shown in Equation 1 were fitted for math, reading, and science outcomes for each grade and state. The discussion here focuses on trends that emerged across all three of the states regarding the average achievement levels of and the achievement gaps among ELL students, former ELL students, and non-ELL students. Immediate results from all analyses are presented in Tables A1–A7 in Appendix CH3 (pp. 107–116).

Results for all three states are summarized by comparing the patterns in terms of grades and various subjects and the estimates in terms of standard deviations (*SD*s) of outcomes are discussed. *SD*s provide a direct sense of the magnitudes of the estimated gaps or expected differences. For example, in studies of treatment effects (e.g., Cohen, 1988), researchers often use rough approximations to gauge the magnitude of treatment effects. For example, under certain circumstances, 0.2 *SD*s is considered "small," 0.5 *SD*s is considered "medium," and over 0.8 *SD*s considered "large." Although the results in this chapter are not effect sizes, researchers can use these approximations as a reference to understand the magnitudes of achievement gaps. More importantly, since the scales of the outcome measures were different by subject and grade, discussions in terms of *SD*s of outcomes facilitate comparisons among subjects and grades.

A statistically significant and large achievement gap was shown between current ELL students and their non-ELL peers in all three states, with a consistent pattern seen across grades and content areas: the gap is greater in upper grades than in lower grades, and greater for reading and science than for math. The magnitudes of average achievement gaps ranged from small to medium in math, whereas, in reading or science, they ranged from medium to large. Also, in the upper grades, the magnitudes tended to be larger by about 0.2 *SD*s than in the lower grades. One of the factors underlying these differences across content areas or across subjects may be the extent of linguistic difficulty ELLs encounter in various subjects and grades. In addition, the magnitudes of gaps varied across states as well. The between-state variability can be due to many important factors, including the differences in the characteristics of ELLs and the differences in the stringency of state redesignation criteria.

Looking specifically at recently redesignated students, the researchers found some mixed results within and between states. In State A, recently redesignated ELL students tend to perform lower in both grades overall. In State C, in Grade 4, recently redesignated students perform significantly better on average than non-ELL students, while in Grade 8, recently redesignated students perform significantly lower on average than non-ELL students. In State B, students who are redesignated tend to perform higher in all grades. However, given these mixed findings (i.e., recently redesignated students on average performed better than non-ELLs under some settings but performed worse in other settings), patterns across content areas and grades were in general consistent to the patterns found with current ELLs. Comparisons between recently redesignated students and non-ELLs showed that recently redesignated students also performed better in math than in reading and science and when they are in the lower grades than in the upper grades.

In all states, former ELL students who were redesignated earlier (i.e., at least two grades previously) were found to perform significantly better, on average, than non-ELL students when controlling for student eligibility for FRL, after 2 years of additional monitoring.

As demonstrated, significant gaps exist between current ELL students and Non-ELL students, while redesignated students who are no longer monitored generally perform as well as or better than non-ELL students. These trends are evident across states, different content areas and different grades. One hypothesis that could be drawn from these results is that the achievement gap between ELL and Non-ELL students tend to be narrowed or closed after a couple of years following redesignation. However, because the present results are based on comparing groups at a single time point, the 2005–06 school year, and with the ELL subgroup composition constantly changing over time, the researchers cannot strongly support this hypothesis.

Specifically, the achievement gap between ELL students and non-ELL students may seem large, while the achievement gap between redesignated ELL students and non-ELL students may seem narrow or nonexistent, as not all ELL students get redesignated or exit from ELL status. Relatively high-performing ELL students may exit ELL status over time, while relatively low-performing ELL students remain in that status. There may be incoming students who are newly arrived in the country and have limited English proficiency. In such settings, the differences in achievement between redesignated ELL students and ELL students may be in part due to these preexisting differences between students who exit and students who stay in the ELL status, rather than changes in ELL group achievement overall.

From these findings, the researchers may conclude that the achievement gap separating ELL students and non-ELL peers tend to close over time only for the students who can exit from the ELL status. While drawing this conclusion, the researchers must be clear that the current analysis and results do not provide such evidence for students who have not exited and may not explain why or how the gap changes for those who do exit. There may be a substantial portion of ELL students who never exit the ELL status and never catch up with their non-ELL peers in terms of academic achievement. A later section examining correlates of redesignation will address this issue.

Investigating Relationships between ELP and Content-area Assessments and Redesignation Criteria

ELP assessment aims to measure student ELP in various domains including speaking, listening, reading and writing. For all the participating states in this study, the scores in ELP assessment were the primary, if not only, criterion for identifying and redesignating ELL students. For example, in State A, an ELL student exits ELL status when he or she scores at level 4 or 5 on the 1–5 proficiency level scale in all five modalities (speaking, listening, reading, writing, and comprehension) and meets the proficiency standard for the state content-area test. In State B, an ELL student may exit the ELL status when he or she scores at level 6 on the 1–6 proficiency level scale in all four modalities (speaking, listening, reading and writing). In State C, although there are no state-wide criteria, the ELP assessment is one of the primary criteria used by districts and schools.

Given the critical roles of ELP assessment in identifying students as ELL students and in redesignating them out of this status, the research team investigates the relationships between the scores (or levels) of ELP and content-area assessments in reading, math, and science. The purpose of this is two-fold: first, to provide empirical evidence of the validity of the ELP assessment use; and second, to gauge the stringency of the redesignation criteria based on the ELP assessment.

Analysis method. ELP assessments can be viewed as valid if they assess ELP such that advancement in proficiency is associated with a decrease in language-related difficulties in school settings, such as regular instruction in English and assessments in English. Based on this, higher English proficiency assumed from the higher scores or levels on a valid ELP assessment should relate to better academic performance measured in state exams. Thus, positive and strong relationships between ELP and content-area assessment scores may provide concurrent validity of the ELP assessment.

To estimate relationships, the researchers used 2-level HMs, in which students were nested within schools. The content-area test scores in various subjects and grades were the outcomes, while ELP scores or levels were predictors of primary interest. Use of HMs takes into account the intra-class correlations among students within a school and thereby yields accurate inferences on the parameters of interest (i.e., the relationships between ELP and content-area assessments). Specifically, the following HMs were fitted to each subject outcome in different grades and states:

Student Level (Level 1)

$$Y_{ij} = \beta_{0j} + \beta_{1j}(ELP)_{ij} + \beta_{2j}(ELP)^2_{ij} + \beta_{3j}\,(FRL)ij +$$

$$\beta_{4j}(FRL)_{ij}(ELP)_{ij} + \beta_{5j}(FRL)ij(ELP)^{2}_{ij} +$$

$$[\beta_{6j}(Cohort4)_{ij} + \beta_{7j}(Cohort5)_{ij} + \beta_{8j}(Cohort6)_{ij}] + r_{ij}\,,\ r_{ij} \sim N(0, \sigma^2)\,,$$

School Level (Level 2)

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})\,,$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}\,,$$

Equation 2. Two-level HM for estimating the relationships between ELP and content-area assessments.

where Y_{ij} was a score in content-area assessments in math, science, or reading of student *i* in school *j*; ELP_{ij} was a score or level in ELP assessment for student *i* in school *j*, ELP^2_{ij} was a quadratic term of ELP_{ij}; and indicates whether student *i* in school *j* received FRL. Cohort variables, were included only in State A due to the availability of the variables in the existing data. $Cohort4_{ij,}$ $Cohort5_{ij,}$ and $Cohort6_{ij},$ indicate students who were identified as ELLs in the 2003–04, 2004–05, and 2005–06 academic years, respectively, with students who were identified in the 2002–03 year as being the base category.

Given the specification of the model, the key parameters of interest were β_{1j} and β_{2j}, which represented the relationship between the ELP levels or scores and the state assessment scores. The quadratic term of ELP captured the extent of curvature in the relationships. In estimating the ELP-content-area assessment relationships, the researchers not only controlled for the FRL status, but also estimated interactions between FRL status and ELP scores, with both linear and quadratic terms, of which the coefficients were β_{4j} and β_{5j}. The interaction terms were posed based on the hypothesis that the relationships between ELP and content-area assessment may depend on student FRL status. It must be noted that the ELP scores are centered around their grand means. By virtue of the grand-mean centering, the intercept β_{0j} represented the expected scores in state test scores for a student who had a mean value of ELP scores and did not receive FRL. In a case where the researchers used ELP levels instead of ELP scale scores, they centered the ELP variable around the medium level (i.e., 3 in a 5-point scale).

As for the redesignation criteria, one piece of evidence was found in the earlier section, in which the researchers compared the academic achievement of ELL students, redesignated ELL students, and non-ELL students. Regardless of states, subjects, and grades, students who were redesignated more than 2 years ago and no longer monitored tended to perform as well as or better than non-ELL students in content-area assessments, holding constant a proxy measure of socioeconomic status. This may provide a rough indication of the validity of redesignation criteria, in a sense that students who used to be ELL students perform as well as an average student who speaks English as his or her native language.

This section investigated the issue of redesignation in terms of the relationships between ELP and content-area assessments. Specifically, the research team estimated the expected scores on content-area assessments based on the ELP cut scores at which students were redesignated. They then compared the scores to the content-area test cut scores at which students are considered meeting or above proficiency. This may provide more details about the validity of redesignation criteria. More specifically, those details may show the extent of stringency of the ELP-related criteria, by estimating how well an ELL student would do when they exit the ELL status as compared to state-designated proficiency levels.

Summary of results. Detailed results for each state, content area and grade level are presented in Tables B1–B7 in Appendix CH3 (pp. 117–124). In looking at trends across all three states, the relationships between ELP and content-area assessments show extremely consistent results. For all grades and subjects examined, ELP scores or levels are strongly and positively associated with the performance in content-area assessments.

In addition to the strong, positive, and highly significant relationships, two other results are of particular note. First, after ELP levels or scores were controlled, student FRL status was neither a significant predictor of achievement nor, despite statistical significance, did it contribute to any substantively meaningful difference in achievement. Second, the relationships between ELP and content-area assessments did not vary across schools. In other words, a positive and strong relationship remained fairly consistent under different settings. Figures 1 and 2 show the results under two selected settings, providing examples of these estimated relationships.

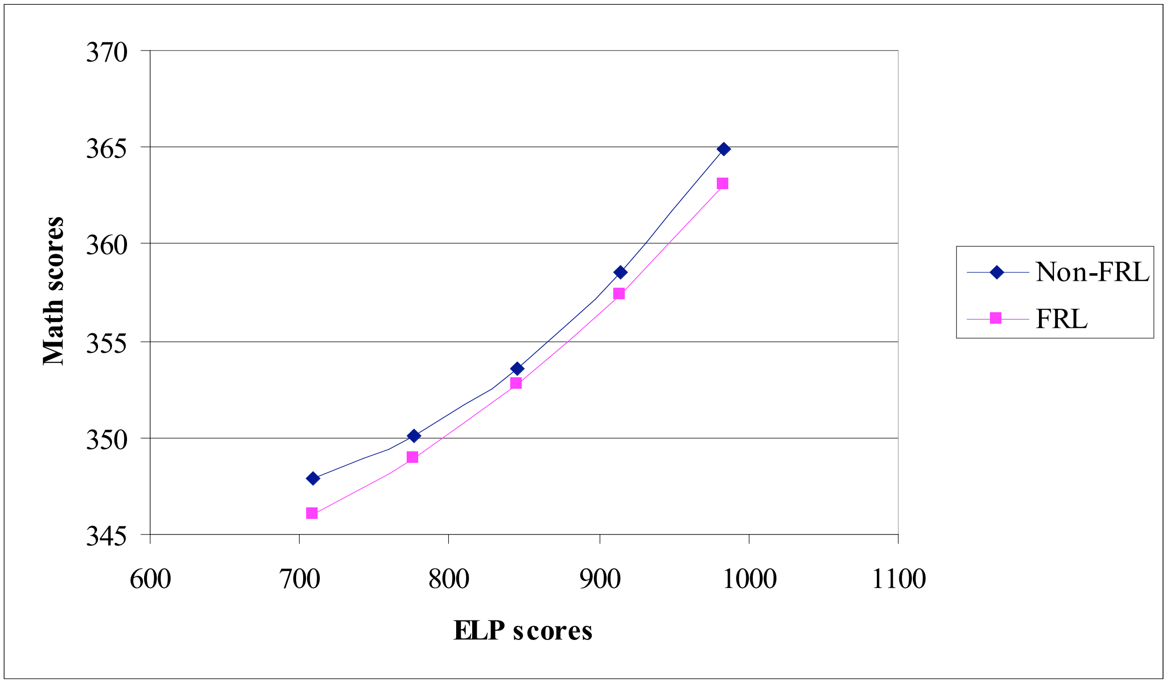

Figure 1. Expected Relationship between ELP and Math Assessment Scores for Grade 8 Students in State B.

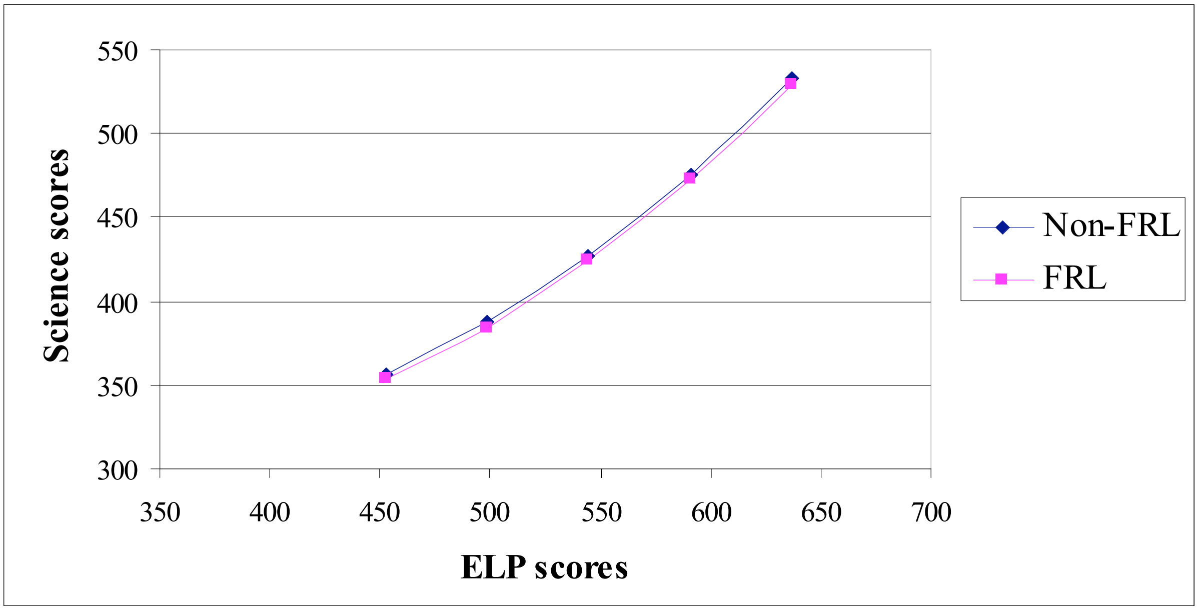

Figure 2. Expected Relationship between ELP and Science Assessment Scores for Grade 8 students in State C.

Findings may imply the concurrent validity of ELP test, in that higher levels of ELP scores on the ELP test are associated with better academic achievement. At the same time, the research findings suggests that, for ELL students, their ELP levels measured by the ELP assessment was a dominant factor in their academic achievement, and that improvement in ELP is almost always associated with a boost in academic achievement.

Another finding of note is that in most cases the relationships were not linear but quadratic. The quadratic terms were positive with a single exception (i.e., State C Grade 4 reading with a small negative term). Positive quadratic terms captured curvatures in the relationships, in which the expected levels in academic achievement tended to increase rather slowly for lower ELP scores, while the levels tended to increase more rapidly for higher ELP scores. Although the extent of curvature appears more apparent in some grades, content areas, or states than in others, the presence of a positive curvature may suggest that ELL students must reach a certain level of English proficiency in order for their academic achievement to start benefiting more from the improvement in ELP.

To further examine the redesignation criteria, the research team compared the estimated scores in state exams given the ELP level at which students are redesignated, with the cut scores set by the state to categorize students as meeting the standard or being at or above proficient. Unlike the relationships between ELP and content-area assessments, the findings about redesignation criteria fluctuated across states.

States A and C showed similar results in terms of the stringency of redesignation criteria. The results showed that the expected content-area scores given the ELP level of 4 or even of 5, with 5 being the highest level, in general do not reach the minimum scores to be at or above proficient in the state content-area test. In State A, for ELL students to exit, they must meet level 4 on the ELP test, which is "advanced intermediate," and meet proficiency in the state content-area test. Given the relationships between ELP levels and state test scores, the real challenge for ELL students to exit ELL status would be to meet proficiency in the state test rather than meeting level 4 proficiency in the ELP assessment.

Conversely, in State B, the expected content-area scores given the highest ELP level are as high as or higher than the proficiency cut-scores in the state test. This means that recently redesignated students tended to perform well enough to meet proficiency in content-area tests when they are just redesignated.

The conclusions drawn about the redesignation criteria based on the estimated relationships between ELP and content-area tests may explain differences among achievement patterns of recently redesignated students in the previous section. The different achievement patterns seem to result from differences in the stringency of the redesignation criteria.

In the state with the more stringent exiting criteria, students who recently exited ELL status perform as well as or better than their non-ELL peers. As English proficiency level is strongly associated with academic performance in ELL students, students who exit with higher levels of proficiency already perform well enough to meet proficiency when they are just redesignated. In contrast, in a state with seemingly more lenient exiting criteria (State A), recently redesignated students tend to perform lower than non-ELL students. State C does not have state-wide criteria of redesignation, which may explain the diverging results about achievement patterns of recently redesignated students within the state.

Because data were available with regard to the relationships between ELP and content-area assessments, the researchers conducted additional analysis on one state concerning when ELL students are identified. The base cohort, a group of students who were designated as ELL students in the 2002–03 school years, comprised the majority of the ELL population. The research results show that later cohorts who were designated as ELL students in the 2003–06 academic school years performed better in the content-area assessments given the same level of ELP assessment than the base cohort who were designated as ELL students in the 2002–2003 academic year, controlling for FRL status. The researchers speculate on potential reasons for the differences between the base and later cohorts in the subsequent discussion section.

Examining Correlates of Redesignation

In an effort to study differences between redesignated students and continuing or long-term ELL students who maintain ELL classification for more than three academic years, the research team examined the correlates of redesignation among ELL students, in contrast to continuing or longer-term ELL students.

The researchers used logistic regressions to predict the binary indicator of redesignation, that is, redesignation in contrast to remaining in ELL status for over 3 years for all three states. The specific models for analysis depended on the available variables in the existing state data bases with the approach being exploratory in nature. The sets of variables can be different from state to state. Important correlates may turn out insignificant due to small frequencies with regard to such correlates, while there may be important variables that are not available from the current data sets. In what follows, the researchers present the fitted models for each of the three states and summarize the findings.

Analysis method (State A). ELL students who were identified as ELL students in the 2002–03 academic year, and remained as ELL students in the 2005–06 academic year, were defined as continuing or long-term ELL students. Students who were redesignated between 2002 and 2006 were defined as redesignated or former ELL students in the analysis. These definitions led to 3,854 redesignated students and 2,936 long-term ELL students in Grade 5, and 4,304 redesignated students and 1,890 long-term ELL students in Grade 8. More than half of the ELL students examined were students who had been ELL students for over three years. For the binary outcome of interest, i.e., indicating redesignation in contrast to remaining in ELL status for over 3 years, the researchers fitted the following logistic regression model:

$$\begin{aligned}\text{Log odds}(Y_i) = {} & b_0 + b_1(\text{FRL})_i + b_2(\text{IEP})_i + b_3(\text{F\_IEP})_i + b_4(\text{S504})_i + b_5(\text{Title I})_i + \\ & b_6(\text{Migrant})_i + b_7(\text{Immigrant})_i + b_8(\text{ID})_i + b_9(\text{IS})_i + b_{10}(\text{NIC})_I + \\ & b_{11}(\text{Asian})_i + b_{12}(\text{Black})_i + b_{13}(\text{Hispanic})_i + b_{14}(\text{Indian})_i + \\ & b_{15}(\text{Gifted or Talented})_i ,\end{aligned}$$

Equation 3. Logistic regression model for examining correlates of redesignation (State A).

where Y_i was the redesignation status for ELL student *i*, coded as 1 if he or she had been redesignated, and as 0 if he or she continued ELL status for over 3 years based on the previous definition. All predictors were also binary indicators from state data in the 2005–06 year. The predictor *FRLi* indicated whether ELL student *i* received FRL; *IEPi* indicated IEP, whether student *i* was one with disability receiving Special Education Services; *F_IEPi* indicated whether student *i* used to be IEP in previous years but exited from it; *S504i* indicated whether student *i* had Section 504 Accommodation Plan; *Title Ii* indicated whether student *i* received Title I Targeted assistance; *Migranti* indicated whether student *i* was identified as a migratory student; *Immigranti* indicated whether student *i* was from an immigrant family; *IDi* indicated whether student *i* was new in the district this year; *ISi* indicated whether student *i* was new in the school this year; *NICi* indicated whether student *i* was new in the U.S. this year; and *Gifted or Talentedi* indicated whether student *i* participated in Gifted or Talented program. *Asiani*, *Blacki*, *Hispanici*, and *Indiani* were binary indicators of Asian, Black, Hispanic, and American Indian, respectively, with White as a base category.

Additionally, the researchers fitted another logistic regression model including a new category: the level of ELP when students were first identified as ELL students. The state categorization of ELP is on a 5-point scale: 1 (*entry*); 2 (*emerging*); 3 (*intermediate*); 4 (*advanced intermediate*); and 5 (*proficient*). Thus, four binary variables *entry2i*, *entry3i*, *entry4i*, and *entry5i* were added to the Equation 3, indicating the entry levels of 2, 3, 4 and 5, respectively, with the entry level of 1 being a base category.

The researchers fitted a separate logistic regression model when the entry level category was included because the inclusion of the category changes the study sample. Since the category was one of the variables recorded for ELL students, it was supposed to be missing for students who were redesignated and not monitored any more. Thus, the sample excluded a considerable proportion of redesignated ELL students, and rather, predicted recent redesignation in contrast to long-term ELL students. The resulting sample sizes for the second model was as follows: in Grade 5 there were 1,153 redesignated students and 2,934 continuing ELL students, and in Grade 8, there were 1,293 redesignated students and 1,890 continuing ELL students.

Analysis method (State B). ELL students who were identified as ELL for over 3 years in the 2005–2006 academic year were defined as “continuing or long-term ELL students.” Students who were redesignated during 2002–2006 were defined as “redesignated or former ELL students” in the analysis. These definitions led to the following sample sizes. In Grade 4, there were 2,539 redesignated students, while 3,303 students were continuing ELL students. In Grade 7, there were 2160 redesignated students, and 1,963 continuing ELL students. In Grade 8, there were 2079 redesignated students, and 1,903 continuing ELL students. For all grades, more than 75% of the ELL students examined were long-term ELL students.

For the binary outcome of interest, i.e., indicating redesignation in contrast to remaining in ELL status for relatively longer periods (i.e., over 3 years), the researchers fitted the following logistic regression model for each grade:

$$\text{Log odds}(Y_i) = b_0 + b_1 (\text{FRL})_i + \Sigma_{J=2,7}\, b_J (\text{PED}_J)_i + b_8 (\text{Asian})_i + b_9 (\text{Black})_i +$$

$$b_{10} (\text{Hispanic})_i + b_{11} (\text{Indian})_i + \Sigma_{K=1,6}\, b_{K+11} (\text{HW}_K)_i +$$

$$\Sigma_{L=2,5}\, b_{L+16} (\text{Read}_L)_i + \Sigma_{M=2,5}\, b_{M+19}(\text{Calcul}_M)_i + \Sigma_{N=2,6}\, b_{N+23}(\text{Comp}_N)_i +$$

$$\Sigma_{O=1,9} b_{O+29}(\text{Extraordinary}_O)_i\,,$$

Equation 4. Logistic regression model for examining correlates of redesignation (State B).

where Y_i is the redesignation status for ELL student *i*, coded as 1 if he or she had been redesignated, and as 0 if he or she continued ELL status for a long time, based on the previous definition. Predictor variables in the equation are shown in Table C3 in Appendix CH3 (pp. 127–128).

Analysis method (State C). ELL students who were identified as ELL for over 3 years in the 2005–2006 academic year were defined as "continuing ELL students," while students who were redesignated were defined as "redesignated or former ELL students" in the analysis. The definition and missing data led to the following sample sizes. In Grade 4, there were 1298 redesignated students, while 3630 students were continuing ELL students. In Grade 8, there were 1592 students, and 1920 continuing ELL students. More than half of the ELL students examined were continuing or long-term ELL students.

For the binary outcome of interest, i.e., indicating redesignation versus remaining in ELL status for relatively longer periods (i.e., over 3 years), the researchers fitted the following logistic regression model for each grade:

$$\text{Log odds } (Y_i) = b_0 + b_1(\text{FRL})_i + b_2(\text{IEP})_i + b_3(\text{Title I})_i + b_4(\text{Migrant})_i + b_5(\text{Immigrant})_i +$$

$$b_6(\text{Asian})_i + b_7(\text{Black})_i + b_8(\text{Hispanic})_i + b_9(\text{Indian})_i +$$

$$\Sigma\, b\, (\text{Extraordinary Conditions})_i + b\, (\text{Gifted or Talented})_i\,,$$

Equation 5. Logistic regression model for examining correlates of redesignation (State C).

where *Yi* is the redesignation status for ELL student *i*, coded as 1 if he or she had been redesignated, and as 0 if he or she had continued ELL status for more than 3 years, based on the previous definition. All predictors are also binary indicators from state data in the 2005–2006 academic year. *FRLi* indicates whether ELL student *i* received FRL. *IEPi* indicates IEP. *Title Ii* indicates whether student *i* received Title I funds. *Migranti* indicates whether student *i* was a migratory student; *Immigranti* indicates whether student *i* was from immigrant families; and *Gifted or Talentedi* indicates whether student *i* was Gifted or Talented. *Asiani*, *Blacki*, *Hispanici*, and *Indiani* are binary indicators of Asians, Blacks, Hispanics, and Indian Americans respectively, with White as a base category.

Summary of results. The focus of this set of analyses was ELL students who remained as ELL students for more than 3 years. In terms of English proficiency, the long-term ELL students the researchers examined in this section can be assumed to have the most educational needs and difficulties in improving English skills. Notably, the long-term or continuing ELL students composed a large percentage of the ELL population. In states A and C, more than 50% of students in the data set were of long-term ELL students; in State B more than 75% were long term ELL students.

In examining correlates of exiting ELL status versus continuing as ELL students for over 3 years across all three of the participating states, the researchers found that in all states, students who received FRL and students with disabilities were more likely to be long-term ELL students for all states. Also, controlling for all the other variables in the model, Asians and Hispanics were more likely to be long-term ELL students across states (although there were a few minor variations in the analysis).

Other predictor variables differed by state depending on data availability and configuration. Significant predictors found in at least one state included: immigrant status (students who were new in the country were more likely to be long-term ELL students) and parental education (students whose parents did not graduate high school were more likely to be long-term ELL students). Given slight variations of specific variable names across states, the findings were consistent in terms of the demographic trends of long-term ELL students in all three states. That is, they tended to be the socio-economically disadvantaged or those who were immigrants or new in the country.

Some variables beyond demographics were used in the analysis, depending on their availability in each of the participating states. In State A, the level of ELP at school entry measured by ELP assessment was used in the analysis. The entry level was a highly significant predictor, that is, students who entered at a lower ELP level were less likely to be redesignated. For example, the odds of being redesignated in less than 3 years when the entry level was 2 or 3 was more than two times the odds when the entry level was 1. When the entry level was 4, the expected odds were more than six times in Grade 5 students and more than three times in Grade 8 students. It is notable that after controlling for the entry level, neither FRL status nor ethnicity group was significant predictors of redesignation. Only immigrant status and disability conditions remained significant. The analysis could only be done in one state with the available data, but this finding may suggest that entry level may be one of the key variables underlying the relationships between demographic variables and redesignation.

In State B, variables that measure academic-relevant behaviors in and out of school were available. After controlling for all the demographic variables, some of these behavioral variables still turned out to be significant in the expected direction: the amount of homework, free reading hours, calculator use in math classes, and use of computers at home for school work. However, unlike the entry ELP level in State A, none of these variables fully explained the relationships of demographic variables.

Summary of States' Reported Uses of Accommodation for ELLs

This section summarizes the uses of accommodation for ELL students in three participating states. Detailed tables with states' reported accommodation strategies are presented by state in Tables D1–D14 in Appendix CH3 (pp. 133–138).

In looking at accommodations use across the states, State A data do not include information about the types of accommodation, but only provides information on whether or not a given student used an accommodation in their state exam for each subject matter, with percentages ranging from 6–25% depending on grade and subject area. Four accommodation types are used most frequently for ELL students in State B, which are extended time of test administration (16–32%); test administration in a separate room (10–30%); reading tests aloud (9–26%); and the use of a translation dictionary (4–16%). In State C, three accommodation types are used most frequently for ELL students: oral presentation of the entire test (15–21%); extended test time (11%); and teachers reading aloud the directions (4–5%).

In all three states, accommodation tends to be used more frequently in the primary-school grades than in the secondary-school grades. However, one specific type of accommodation, provision of translation dictionaries, tends to be administered more frequently as the grade gets higher. Translation dictionaries are administered to 3.5% of the Grade 4 ELL students, 7.1% of the Grade 7 ELL students, and 16% of the Grade 8 ELL students.

With regard to co-administration of different types of accommodation, State B and State C show contrasting results. In State B, various accommodations tend to be administered in combination with others. For example, almost every time when the translator dictionary accommodation is administered or when the test is read aloud, more test time is offered and also the test is given in a separate room. These co-occurrences of accommodations are found regardless of the ELL status of students. In contrast, in State C among ELL students and in the three types of accommodation that are examined, no single co-occurrence is found. At this point, it is unclear whether it is due to no co-administration, or due to reporting system, such as reporting only one main type of accommodation for individual students. As State A data do not have information about the types of accommodation available as mentioned earlier, such analysis was not possible.

Lastly, the researchers examined the percentages of ELL students that receive each type of accommodation by ELP level, as measured by the ELP assessment, in all states and grade levels. With breakdowns by English proficiency level, one can see that the percentages of accommodation uses are much greater than was seen in overall results. This is mainly due to the fact that ELL students in overall summaries include all former ELL students who are not monitored and not allowed to use accommodations any more. With regard to ELP levels that tend to use accommodations more frequently, states show mixed results.

In State A, math and science show a similar pattern. Among ELL students with lower English proficiency, from levels 1 to 3, about 50–55 % of Grade 5 students, and about 40–47% of Grade 8 students, use some kind of accommodation. Among ELL students at ELP level 4, a slightly lower percentage, 44% of Grade 5 students and around 30% of Grade 8 students use accommodations. ELL students with the highest ELP (i.e., level 5), i.e. the students who exceed the cutoff of redesignation in terms of ELP assessments, use accommodation significantly less, only 11–12% of them in the Grade 5 Students and 6–7% in the Grade 8 Students.

In State C, two types of accommodation, oral presentation of the entire test and teacher-read directions only, are used more frequently as ELL students have lower English proficiency. For example, among Grade 4 ELL students with the lowest ELP levels, levels 1 and 2, oral presentation of the entire test is used 41–48%, 26% among ELL students with level 3, and 10% among ELL students with level 4. Students with the highest ELP level, level 5, rarely used the accommodation (2%). Teacher-read directions are only used 7–8% among students with levels 1 or 2, 6% among students with level 3, 4% among students with level 4, and less than 1% among students with the highest ELP. However, extended time shows a different pattern, with the accommodation being used rather evenly across English proficiency levels. The Grade 8 ELL students show extremely similar patterns.

Unlike the two other states, in State B, accommodation is used most frequently among ELL students with higher English proficiency, levels 4 and 5; and less frequently as ELP levels become lower. For example, read aloud is administered 78–88% among ELL students of levels 4 and 5, 75% among ELL students of level 3, 62% among ELL students with level 2, and 48% among ELL students with level 1. Accommodation is rarely used in the highest level, level 6 or superior, which is expected. State B does not allow ELL students with a superior level of reading to receive any accommodation.

Discussion and Implications for Future Studies

In this section, the researchers discuss the key findings related to the research questions and suggest implications for possible future studies from the findings of this study.

It is well known that achievement gaps exist between the socioeconomically disadvantaged and their non-disadvantaged peers. This has an important implication for the ELL population because ELL students have a much higher percentage of students receiving FRL compared to their non-ELL peers. It has been established that ELL students are low-performing subpopulations due to limited proficiency in English language, but their socio-economically disadvantaged status may also be a contributing factor. Remembering this confound, the achievement comparisons among different groups (i.e., current ELL students, recently redesignated students, redesignated and no-longer monitored students, and non-ELL students) controlled for student FRL status. Consistent findings across the participating states included appreciable average gaps between the current ELL students and non-ELL students, and the research team found that the redesignated and no-longer monitored students tended to perform as well as non-ELL students.

Other consistent findings show that both current and redesignated ELL students performed better in primary school grades than in secondary school grades, and also performed better in math than in reading or science. One of the possible factors that may explain the differences across grades and subjects would be the extent of linguistic difficulty that students may encounter in various grades and subjects. In other words, students may encounter more linguistic difficulty in learning from instructions and in taking tests in upper grades than in lower grades, and may have more difficulty in language arts or science than in math.

Results concerning recently redesignated students fluctuated across states. Based on the relationships between ELP scores and content-area assessment scores, the researchers examined the extent of stringency on ELP-related redesignation criteria in terms of levels of English proficiency required for redesignation. As hypothesized, the results show that in states with more stringent criteria, recently redesignated students performed as well as or better than non-ELL students. In states with more lenient criteria, recently redesignated students performed worse than non-ELL students. States A and B may exit students earlier than State C, but at relatively lower English proficiency levels. However, redesignated students seem to achieve as well as non-ELL students after a couple of years, given that no-longer monitored students perform better than non-ELL students. Based on these factors, earlier exit may be comparable to later exit.

However, it is important to note that the average academic performance of redesignated ELL students is only one of the many important considerations in evaluating redesignation policy. There may be other important factors, such as tailored instruction for ELL students, policies around accommodations for ELL students, and the welfare of ELL students. For example, if ELL students benefit from tailored instruction even when they have a higher level of English proficiency, it would be more desirable to exit them later than earlier. However, if earlier exit prevents them from using accommodations that they may still need or forces them to immerse too early either culturally or in terms of the use of language, later exit would be more desirable. Thus, the determination of the optimal levels of English proficiency for the purpose of redesignation will depend on various factors at the student, school, and policy levels. More research is needed to help determine optimal levels of English proficiency for redesignation within the context of these other factors. For example, future studies should identify available tailored instructions that may benefit ELL students at various levels. The relationship between implementation of these instructions and ELL students' academic performance and progress should also be examined.

The relationships between ELP scores and content-area assessment scores were found to be strong, positive, and highly significant. The relationships are strong enough to account for other predictors, including student FRL status. Also, the relationship between ELP and content-area assessments did not vary across schools but was consistent, which implies that a boost in content-area performance is almost always associated with higher performance in ELP. This may imply the concurrent validity of ELP assessment as mentioned earlier. On the other hand, this may raise the question of whether the ELP assessment overlaps substantially with content-area assessment, rather than measuring English proficiency. This would be a question of construct validity, which again goes beyond the scope of this chapter and can be answered by corresponding analysis of the contents of both assessments.

One analysis conducted only in State A due to data availability shows that ELL students who were identified in the 2002–03 academic year (base cohort) performed better than those who were identified in later years (later cohort), consistently across multiple subjects and grades. It is unclear without further investigation why the differences arose between the base and later cohorts. One hypothesis is that students who were identified earlier are different in various ways from those who were identified later, which enables them to perform academically better given similar levels of ELP. Another hypothesis is with regard to the assessments. ELP assessment may have changed. The criteria may have become more stringent since the 2003–04 academic year, classifying higher ELP students into a lower level compared to the 2002–03 year. Another possibility is with regard to the change in the state assessment. State assessment may have been modified to decrease the language complexity from 2003–04. In addition, accommodations may have begun to be effectively used for ELL students to help decrease the challenges of ELP. One meaningful line of research may empirically examine the underlying sources of differences between base and later cohorts, which examine both state data and changes in practices or policies.

The research team conducted analyses focusing specifically on continuing or long-term ELL students. Continuing ELL students account for 50–75% of examined ELL students across the states. Correlates of long-term ELL students were identified, but the findings were limited because the variables available in the state data were mainly limited to demographics. Thus, the sets of analyses in this section are only a small step towards an effort to examine long-term ELL students. Further studies should collect direct measures of other student variables with regard to language acquisition. The entry level of English proficiency that was available in only one state was found to be promising. When the entry levels of ELP are taken into account, higher entry levels were significantly associated with redesignation, accounting for the relationships of demographics. Also, this may strongly suggest the need for longitudinal analysis of individual ELL students for an extended period of time, from their identification to redesignation status. Longitudinal studies can address questions such as where ELL students start in their English proficiency level (i.e., initial status, or the entry ELP level from the earlier discussion), and how rapidly ELL students improve in their English proficiency over time (i.e., growth rates). This approach enables us to examine (1) the relationship between where ELL students start and how rapidly they improve in their English proficiency; and (2) the correlates of both initial status and growth rates.

Further studies may also need to address possible intragroup differences within the ELL population. ELL students have diverse backgrounds, which may imply qualitatively different stages or issues in their language development, while some may have limited education both in their native language and English, others may have considerable education in their native language. Different types of ELL students may have distinct needs and patterns or acquisition, leading to different accommodation use or instructional strategies. While the first may learn English along with the content knowledge, the second may instead transfer what they know in their native language. Again, based on theories of second language acquisition, collection of such variables would be required to conduct such a focused study.

As for accommodations, state data allowed us to identify frequently used types of accommodation for ELL students in two states. It also allowed us to examine how accommodations are implemented depending on levels of English proficiency. However, the researchers did not pursue further analysis, e.g., how they relate to academic achievement of ELL students. One of the main reasons was the likelihood of inaccuracies in the accommodation data. For example, the frequencies of accommodation are much smaller than expected, which may imply under-reporting or differential under-reporting across schools or districts. State accommodation data are collected to serve much broader purposes than only focusing on ELL students, and in many ways this data is less than ideal for the current purposes. The credibility of the data should be ensured before conducting meaningful studies of how the data relate to academic achievement.

References

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CRESST Tech. Rep. No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodation for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CRESST Tech. Rep. No. 666). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CRESST Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bibian, G. C. (2006). *An ethnographic case study of a reconstituted urban middle school and the factors that contribute to improved ELL redesignation rates.* Unpublished doctoral dissertation, University of California, Los Angeles.

Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments.* Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved November 21, 2006, from http://www.centeroninstruction.org/files/ELL1-Interventions.pdf

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Stack, J. (2002, September). *California English Language Development Test (CELDT): Language proficiency and academic achievement.* Paper presented at the Annual Conference of the National Center for the Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles, CA.

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

APPENDIX CH3

Tables A1–A7

Investigating ELL students' performance in various subjects of state content-area tests relative to their non-ELL peers and redesignated students performance

Table A1

Estimating expected levels of and gaps among ELLs, redesignated ELLs, and non-ELLs in content-area test achievement (State A, Grade 5)

Math

Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	318.72	1.55	205.32	<.0001
ELL, γ_{10}	-39.96	1.34	-29.80	<.0001
ExitMonitor, γ_{20}	-3.96	1.92	-2.07	0.0389
Exit, γ_{30}	28.64	1.58	18.08	<.0001
FRL, γ_{40}	-30.24	1.21	-24.95	<.0001

Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	398.03	47.40	8.40	<.0001
ELL, τ_{11}	73.48	30.68	2.39	0.0083
Exit, τ_{33}	91.60	38.88	2.36	0.0092
FRL, τ_{44}	122.80	26.21	4.69	<.0001
Residual	3972.31	31.33	126.79	<.0001

Reading

Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	298.15	1.34	222.53	<.0001
ELL, γ_{10}	-56.59	1.32	-42.78	<.0001
ExitMonitor, γ_{20}	-8.94	1.86	-4.80	<.0001
Exit, γ_{30}	23.37	1.32	17.69	<.0001
FRL, γ_{40}	-32.53	1.14	-8.44	<.0001

(table continues)

Table A1 (*continued*)

Reading				
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	283.04	35.22	8.04	<.0001
ELL, τ_{11}	76.27	30.47	2.50	0.0062
Exit, τ_{33}				
FRL, τ_{44}	103.33	23.91	4.32	<.0001
Residual	3776.45	29.72	127.08	<.0001

Science				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	316.86	1.60	198.26	<.0001
ELL, γ_{10}	-57.75	1.53	-37.64	<.0001
ExitMonitor, γ_{20}	-19.86	2.04	-9.76	<.0001
Exit, γ_{30}	17.68	1.60	11.02	<.0001
FRL, γ_{40}	-36.77	1.20	-30.74	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	418.97	49.53	8.46	<.0001
ELL, τ_{11}	128.13	39.05	3.28	0.0005
Exit, τ_{33}	66.67	37.97	1.76	0.0396
FRL, τ_{44}	99.66	24.79	4.02	<.0001
Residual	4489.33	35.41	126.80	<.0001

Table A2

Estimating expected levels of and gaps among ELLs, redesignated ELLs, and non-ELLs in content-area test achievement (State A, Grade 8)

Math				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	296.31	5.15	57.52	<.0001
ELL, γ_{10}	-64.08	2.64	-24.29	<.0001
ExitMonitor, γ_{20}	-44.68	3.74	-11.93	<.0001
Exit, γ_{30}	20.28	2.58	7.86	<.0001
FRL, γ_{40}	-38.45	2.02	-19.06	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	2294.41	384.10	5.97	<.0001
ELL, τ_{11}	192.98	69.84	2.76	0.0029
ExitMonitor, τ_{22}	307.58	125.39	2.45	0.0071
Exit, τ_{33}	151.04	62.13	2.43	0.0075
FRL, τ_{44}	148.41	43.13	3.44	0.0003
Residual	7492.15	58.96	127.07	<.0001

Reading				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	297.69	4.77	62.36	<.0001
ELL, γ_{10}	-54.13	2.28	-23.75	<.0001
ExitMonitor, γ_{20}	-30.68	3.00	-10.22	<.0001
Exit, γ_{30}	13.68	1.79	7.66	<.0001
FRL, γ_{40}	-26.89	1.47	-18.25	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	2044.06	326.80	6.25	<.0001
ELL, τ_{11}	182.12	57.40	3.17	0.0008
ExitMonitor, τ_{22}	236.10	89.12	2.65	0.004
Exit, τ_{33}	64.79	31.73	2.04	0.0206
FRL, τ_{44}	79.25	23.49	3.37	0.0004
Residual	3951.95	31.11	127.05	<.0001

(table continues)

Table A2 (*continued*)

Science				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	309.16	4.91	62.90	<.0001
ELL, γ_{10}	-66.47	2.51	-26.48	<.0001
ExitMonitor, γ_{20}	-39.21	3.24	-12.12	<.0001
Exit, γ_{30}	11.05	2.10	5.25	<.0001
FRL, γ_{40}	-35.89	1.69	-21.30	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	2120.58	349.04	6.08	<.0001
ELL, τ_{11}	198.24	68.82	2.88	0.002
ExitMonitor, τ_{22}	222.79	98.47	2.26	0.0118
Exit, τ_{33}	85.72	42.23	2.03	0.0212
FRL, τ_{44}	97.65	31.97	3.05	0.0011
Residual	5665.85	44.72	126.69	<.0001

Table A3

Estimating Expected Levels of and gaps among ELLs, Redesignated ELLs, and non-ELLs in Content-Area Test Achievement (State B, Grade 4)

	Math				Reading			
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	351.14	0.09	3747.74	<.0001	255.26	0.07	3445.98	<.0001
ELL, γ_{10}	-2.05	0.15	-13.32	<.0001	-3.93	0.15	-26.54	<.0001
ExitMonitor, γ_{20}	3.80	0.25	15.22	<.0001	2.69	0.23	11.62	<.0001
Exit, γ_{30}	3.98	0.25	15.92	<.0001	2.69	0.23	11.71	<.0001
FRL, γ_{40}	-5.46	0.08	-68.52	<.0001	-5.00	0.07	-71.94	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	9.13	0.43	21.05	<.0001	5.21	0.27	19.63	<.0001
ELL, τ_{11}	1.97	0.74	2.66	0.004	2.43	0.68	3.58	0.0002
Exit, τ_{33}	2.30	1.58	1.45	0.0729	1.72	1.20	1.44	0.0756
FRL, τ_{44}	3.25	0.29	11.02	<.0001	2.09	0.22	9.58	<.0001
Residual	69.23	0.31	221.75	<.0001	59.34	0.27	221.13	<.0001

Table A4

Estimating Expected Levels of and Gaps among ELLs, Redesignated ELLs, and non-ELLs in Content-Area Test Achievement (State B, Grade 7)

	Math				Reading			
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	360.30	0.15	2456.29	<.0001	264.36	0.11	2388.50	<.0001
ELL, γ_{10}	-3.94	0.21	-19.20	<.0001	-6.34	0.21	-30.05	<.0001
ExitMonitor, γ_{20}	1.82	0.34	5.33	<.0001	0.91	0.31	2.95	0.0031
Exit, γ_{30}	3.18	0.27	11.60	<.0001	1.88	0.23	8.04	<.0001
FRL, γ_{40}	-5.73	0.12	-48.93	<.0001	-5.01	0.10	-50.85	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	10.68	0.75	14.19	<.0001	5.72	0.44	12.89	<.0001
ELL, τ_{11}	2.45	0.91	2.70	0.0034	4.64	1.07	4.32	<.0001
Exit, τ_{33}	4.93	1.50	3.30	0.0005	2.77	1.11	2.50	0.0061
FRL, τ_{44}	4.25	0.42	10.19	<.0001	2.79	0.30	9.23	<.0001
Residual	69.30	0.35	197.79	<.0001	55.56	0.28	197.46	<.0001

Table A5

Estimating Expected Levels of and Gaps among ELLs, Redesignated ELLs, and non-ELLs in Content-Area Test Achievement (State B, Grade 8)

	Math				Reading			
Fixed effect	Coefficient	*SE*	t ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	360.11	0.17	2080.84	<.0001	265.37	0.13	1974.43	<.0001
ELL, γ_{10}	-2.82	0.19	-14.75	<.0001	-6.21	0.22	-27.66	<.0001
Exit monitor, γ_{20}	2.89	0.53	5.46	<.0001	1.56	0.50	3.13	0.0017
Exit, γ_{30}	3.53	0.23	15.19	<.0001	1.89	0.22	8.73	<.0001
FRL, γ_{40}	-5.06	0.11	-46.89	<.0001	-5.00	0.09	-52.80	<.0001
Random Effect	Variance Component	*SE*	*z* ratio	*p* value	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	18.31	1.15	15.92	<.0001	10.58	0.71	14.88	<.0001
ELL, τ_{11}	2.08	0.80	2.59	0.0048	6.39	1.24	5.17	<.0001
ExitMonitor, τ_{22}	10.04	4.45	2.26	0.0121	7.97	4.17	1.91	0.028
Exit, τ_{33}	2.73	1.06	2.56	0.0052	2.00	0.93	2.15	0.0158
FRL, τ_{44}	4.47	0.38	11.79	<.0001	3.20	0.29	10.90	<.0001
Residual	62.89	0.27	229.25	<.0001	57.55	0.25	228.78	<.0001

Table A6

Estimating expected levels of and gaps among ELLs, redesignated ELLs, and non-ELLs in content-area test achievement (State C, Grade 4)

	Math				Reading			
Fixed effects	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	503.73	0.94	535.19	<.0001	603.03	0.67	899.25	<.0001
ELL, γ_{10}	-42.06	1.39	-30.22	<.0001	-49.31	1.37	-35.98	<.0001
ExitMonitor, γ_{20}	15.26	2.20	6.95	<.0001	8.86	1.68	5.29	<.0001
Exit, γ_{30}	11.57	3.06	3.79	0.0002	8.76	2.25	3.89	<.0001
FRL, γ_{40}	-31.14	0.78	-39.70	<.0001	-25.66	0.65	-39.69	<.0001
Random effects	Variance component	*SE*	Z ratio	*p* value	Variance component	*SE*	Z ratio	*p* value
Intercept, τ_{00}	660.03	36.61	18.03	<.0001	308.46	18.76	16.44	<.0001
ELL, τ_{11}	350.13	56.10	6.24	<.0001	504.90	64.28	7.86	<.0001
ExitMonitor, τ_{22}	311.33	101.66	3.06	0.0011	110.97	52.71	2.11	0.0176
Exit, τ_{33}	294.12	193.93	1.52	0.0647				
FRL, τ_{44}	74.49	21.22	3.51	0.0002	58.97	14.89	3.96	<.0001
Residual	4188.73	25.78	162.50	<.0001	2768.30	17.11	161.83	<.0001

Table A7

Estimating expected levels of and gaps among ELLs, redesignated ELLs, and non-ELLs in content-area test achievement (State C, Grade 8)

Math

Fixed effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	571.92	1.55	369.94	<.0001
ELL, γ_{10}	-46.78	1.98	-23.60	<.0001
ExitMonitor, γ_{20}	-0.82	2.59	-0.32	0.7526
Exit, γ_{30}	13.68	2.39	5.72	<.0001
FRL, γ_{40}	-34.67	1.06	-32.75	<.0001

Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	950.77	74.95	12.69	<.0001
ELL, τ_{11}	374.99	80.91	4.63	<.0001
ExitMonitor, τ_{22}	182.46	80.81	2.26	0.012
Exit, τ_{33}	239.53	95.21	2.52	0.0059
FRL, τ_{44}	182.91	28.27	6.47	<.0001
Residual	4118.11	24.53	167.85	<.0001

Science

Fixed effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	510.07	1.26	406.41	<.0001
ELL, γ_{10}	-59.36	1.73	-34.40	<.0001
ExitMonitor, γ_{20}	-12.62	1.98	-6.37	<.0001
Exit, γ_{30}	1.94	1.97	0.99	0.3242
FRL, γ_{40}	-28.32	0.80	-35.46	<.0001

Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	633.03	50.07	12.64	<.0001
ELL, τ_{11}	347.03	64.90	5.35	<.0001
ExitMonitor, τ_{22}	100.56	46.14	2.18	0.0146
Exit, τ_{33}	204.23	66.79	3.06	0.0011
FRL, τ_{44}	97.45	15.51	6.28	<.0001
Residual	2465.17	14.72	167.50	<.0001

(table continues)

Table A7 *(continued)*

		Reading		
Fixed effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept, γ_{00}	6614.00	1.16	571.69	<.0001
ELL, γ_{10}	-70.85	2.06	-34.40	<.0001
ExitMonitor, γ_{20}	-9.86	1.94	-5.09	<.0001
Exit, γ_{30}	5.55	1.93	2.87	0.0041
FRL, γ_{40}	-29.13	0.85	-34.10	<.0001
Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept, τ_{00}	511.91	42.03	12.18	<.0001
ELL, τ_{11}	556.25	92.80	5.99	<.0001
ExitMonitor, τ_{22}	33.90	41.70	0.81	0.2082
Exit, τ_{33}	123.38	61.92	1.99	0.0232
FRL, τ_{44}	106.89	17.78	6.01	<.0001
Residual	2971.97	17.74	167.57	<.0001

Appendix CH3

Tables B1–B7

Investigating the relationship between ELL students' performance on the content-area and the ELP tests, the expected scores on content-area tests based on ELP-to-content-area test relationships and how these estimated scores compare to the content-area test cut scores for meeting proficiency

Table B1

Estimating the relationships between ELP and content-area tests (State A, Grade 5)

Math				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	219.12	2.86	76.60	<.0001
ELP	30.79	1.66	18.53	<.0001
ELP^2	4.82	0.50	9.69	<.0001
FRL	–3.12	2.63	–1.18	0.2367
ELP *FRL	–3.31	1.73	–1.91	0.0557
cohort4	17.68	2.78	6.36	<.0001
cohort5	26.06	2.53	10.32	<.0001
cohort6	20.71	2.58	8.04	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	166.02	32.27	5.15	<.0001
Residual	2751.97	52.39	52.53	<.0001

(table continues)

Table B1 (*continued*)

Science				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	197.01	2.97	66.23	<.0001
ELP	35.58	1.73	20.54	<.0001
ELP^2	3.22	0.52	6.22	<.0001
FRL	–7.43	2.74	–2.71	0.0067
ELP *FRL	–3.69	1.80	–2.05	0.0405
cohort4	13.66	2.89	4.73	<.0001
cohort5	19.75	2.63	7.51	<.0001
cohort6	14.26	2.69	5.31	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	173.29	35.06	4.94	<.0001
Residual	2976.35	56.76	52.44	<.0001

Reading				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	178.43	2.28	78.27	<.0001
ELP	32.55	0.80	40.49	<.0001
ELP^2	6.59	0.45	14.59	<.0001
FRL	–6.97	1.87	–3.74	0.0002
cohort4	15.92	2.52	6.31	<.0001
cohort5	22.86	2.30	9.96	<.0001
cohort6	18.67	2.34	7.98	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	156.15	27.26	5.73	<.0001
Residual	2272.27	43.17	52.63	<.0001

Table B2

Estimating the relationships between ELP and content-area tests
(State A, Grade 8)

Math				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	163.92	4.27	38.35	<.0001
ELP	26.61	1.38	19.35	<.0001
ELP ^{2}	7.98	0.77	10.40	<.0001
FRL	–6.63	2.58	–2.58	0.01
ELP ^{2}*FRL				
cohort4	35.04	4.32	8.10	<.0001
cohort5	47.10	3.85	12.24	<.0001
cohort6	30.82	4.20	7.33	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	439.91	127.61	3.45	0.0003
Residual	5011.02	108.69	46.11	<.0001

Science				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	169.72	3.92	43.25	<.0001
ELP	29.57	1.17	25.32	<.0001
ELP ^{2}	9.06	1.14	7.98	<.0001
FRL	7.48	3.75	1.99	0.0461
ELP ^{2}*FRL	–3.86	1.27	–3.04	0.0024
cohort4	30.55	3.67	8.32	<.0001
cohort5	38.51	3.27	11.78	<.0001
cohort6	25.97	3.58	7.25	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	142.58	51.49	2.77	0.0028
Residual	3584.68	78.07	45.91	<.0001

(table continues)

Table B2 (*continued*)

Reading				
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value
Intercept	184.19	3.25	56.60	<.0001
ELP	26.80	1.00	26.75	<.0001
ELP^2	5.58	0.56	9.99	<.0001
FRL				
ELP^2*FRL				
cohort4	25.77	3.14	8.20	<.0001
cohort5	29.77	2.80	10.63	<.0001
cohort6	19.89	3.07	6.49	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value
Intercept	386.78	113.73	3.40	0.0003
Residual	2649.06	57.60	45.99	<.0001

Table B3

Estimating the Relationships between ELP and Content-Area Tests (State B, Grade 4)

	Math				Reading			
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept	344.40	0.25	1369.38	<.0001	246.56	0.20	1232.04	<.0001
ELP	0.08	0.00	46.79	<.0001	0.09	0.00	68.82	<.0001
ELP^2	0.00	0.00	8.71	<.0001	0.00	0.00	10.42	<.0001
FRL	-1.18	0.26	-4.61	<.0001	-0.88	0.21	-4.25	<.0001
Random effect	Variance component	*SE*	*z* ratio	*p* value	Variance component	*SE*	*z* ratio	*p* value
Intercept	4.13	0.57	7.31	<.0001	1.78	0.31	5.77	<.0001
Residual	42.91	0.87	49.42	<.0001	27.46	0.56	48.76	<.0001

Table B4

Estimating the Relationships between ELP and Content-Area Tests (State B, Grade 7)

	Math				Reading			
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept	352.10	0.41	866.24	<.0001	253.46	0.30	835.49	<.0001
ELP	0.08	0.01	14.43	<.0001	0.09	0.00	44.91	<.0001
ELP^2	0.00	0.00	0.51	0.6068	0.00	0.00	7.35	<.0001
FRL	-1.31	0.42	-3.12	0.0018	-0.64	0.32	-2.01	0.0449
ELP*FRL	-0.02	0.01	-2.37	0.0179				
ELP^2*FRL	0.00	0.00	2.02	0.0433				
Random effect	Variance component	*SE*	*z* ratio	*p* Value	Variance component	*SE*	*z* ratio	*p* value
Intercept	4.68	0.92	5.08	<.0001	0.70	0.35	2.00	0.023
Residual	41.91	1.26	33.31	<.0001	31.82	0.94	33.83	<.0001

Table B5

Estimating the Relationships between ELP and Content-Area Tests (State B, Grade 8)

	Math				Reading			
Fixed effect	Coefficient	*SE*	*t* ratio	*p* value	Coefficient	*SE*	*t* ratio	*p* value
Intercept	353.57	0.30	1161.39	<.0001	254.14	0.28	912.21	<.0001
ELP	0.06	0.00	36.68	<.0001	0.07	0.00	22.62	<.0001
ELP^2	0.00	0.00	8.46	<.0001	0.00	0.00	12.55	<.0001
FRL	-0.87	0.31	-2.77	0.0057	-0.20	0.30	-0.66	0.5092
ELP*FRL					0.01	0.00	4.13	<.0001
ELP^2*FRL	0.00	0.00	-2.37	0.0179	0.00	0.00	-2.08	0.0374
Random effect	Variance component	*SE*	*z* ratio	*p* value	Variance component	*SE*	*z* ratio	*p* value
Intercept	3.99	0.68	5.89	<.0001	1.66	0.42	3.94	<.0001
Residual	36.71	0.95	38.59	<.0001	31.15	0.82	37.94	<.0001

Table B6

Estimating the relationships between ELP and content-area tests (State C, Grade 4)

	Math				Reading			
Fixed effects	Coefficient	*SE*	t ratio	*p* value	Coefficient	*SE*	t ratio	*p* value
Intercept	429.54	1.09	394.03	<.0001	533.56	1.47	363.49	<.0001
ELP	1.17	0.01	85.44	<.0001	1.35	0.03	46.05	<.0001
ELP^2	0.00	0.00	14.54	<.0001	0.00	0.00	-14.78	<.0001
FRL					1.50	1.48	1.01	0.3125
ELP*FRL					-0.11	0.03	-3.25	0.0011
Random effects	Variance component	*SE*	Z ratio	*p* value	Variance component	*SE*	Z ratio	*p* value
Intercept	317.07	34.78	9.12	<.0001	108.10	16.29	6.64	<.0001
Residual	2381.22	39.27	60.64	<.0001	1644.41	27.82	59.11	<.0001

Table B7

Estimating the relationships between ELP and content-area tests
(State C, Grade 8)

Math

Fixed Effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept	492.16	1.60	308.01	<.0001
ELP	0.97	0.02	44.16	<.0001
ELP^{2}	0.00	0.00	8.06	<.0001
FRL				
ELP^{2}*FRL				

Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept	241.45	45.05	5.36	<.0001
Residual	3044.36	67.22	45.29	<.0001

Science

Fixed effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept	427.24	1.68	253.57	<.0001
ELP	0.96	0.02	62.47	<.0001
ELP^{2}	0.00	0.00	10.98	<.0001
FRL	-3.34	1.61	-2.08	0.0377
ELP^{2}*FRL				

Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept	118.72	22.44	5.29	<.0001
Residual	1445.85	32.07	45.09	<.0001

(table continues)

Table B7 *(continued)*

Reading				
Fixed effects	Coefficient	*SE*	*t* ratio	*p* value
Intercept	570.86	2.25	253.77	<.0001
ELP	1.20	0.02	61.45	<.0001
ELP ^{2}	0.00	0.00	0.02	0.9857
FRL	-4.74	2.32	-2.04	0.0412
ELP ^{2}*FRL	0.00	0.00	2.17	0.0302
Random effects	Variance component	*SE*	*z* ratio	*p* value
Intercept	116.93	24.72	4.73	<.0001
Residual	2240.88	49.54	45.23	<.0001

Appendix CH3

Tables C1–C6

Investigating the characteristics of students who are redesignated versus those who continue as ELLs for an extended period of time and the correlates of redesignation

Table C1

Examining correlates of redesignation (State A, Grade 5)

	Analysis without entry level			Analysis with entry level		
Variable	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value
Intercept	1.08	0.23	<.0001	–1.89	0.38	<.0001
FRL	–0.73	0.07	<.0001	–0.16	0.11	0.15
IEP	–2.32	0.11	<.0001	–1.84	0.17	<.0001
F_IEP	–0.48	0.15	0.00	–0.68	0.25	0.01
S504	–0.45	0.54	0.40	–0.07	0.74	0.93
TITLE1	0.37	0.76	0.63	–10.28	233.70	0.96
MIGRANT	0.33	0.82	0.68	0.05	1.31	0.97
IMMIGRANT	–0.83	0.13	<.0001	–0.53	0.23	0.02
ID	–0.10	0.21	0.65	–0.10	0.33	0.76
IS	0.20	0.11	0.07	0.36	0.16	0.03
NIC	–2.21	1.07	0.04	–2.46	1.64	0.13
Gifted/Talented	1.44	0.21	<.0001	–0.94	0.45	0.04
Ethnicity						
Asian	0.49	0.19	0.01	0.59	0.30	0.05
Black	0.84	0.46	0.07	1.04	0.64	0.10
Hispanic	–0.13	0.16	0.40	0.25	0.25	0.31
Indian	0.06	0.61	0.92	0.29	0.92	0.75
Entry level						
Level 2	[a]	[a]	[a]	0.81	0.12	<.0001
Level 3	[a]	[a]	[a]	0.78	0.11	<.0001
Level 4	[a]	[a]	[a]	1.84	0.10	<.0001
Level 5	[a]	[a]	[a]	5.22	0.75	<.0001

[a] Does not apply.

Table C2

Examining correlates of redesignation (State A, Grade 8)

	Analysis without entry level			Analysis with entry level		
Variable	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value
Intercept	2.14	0.29	<.0001	–0.35	0.41	0.40
FRL	–0.37	0.07	<.0001	–0.14	0.10	0.17
IEP	–2.13	0.10	<.0001	–1.52	0.15	<.0001
F_IEP	–0.29	0.20	0.16	–0.36	0.31	0.24
S504	–0.14	0.76	0.86	–0.50	1.30	0.70
MIGRANT	–12.54	327.90	0.97	–11.72	458.20	0.98
IMMIGRANT	–1.99	0.15	<.0001	–0.81	0.25	0.00
ID	0.04	0.24	0.88	–0.68	0.32	0.04
IS	–0.10	0.13	0.42	0.19	0.20	0.34
NIC	–13.73	327.90	0.97	–12.85	458.20	0.98
Gifted/Talented	1.13	0.77	0.14	2.18	0.84	0.01
Ethnicity						
Asian	0.07	0.24	0.79	–0.49	0.36	0.17
Black	–0.87	0.48	0.07	–0.60	0.68	0.38
Hispanic	–0.75	0.21	0.00	–0.61	0.30	0.05
Indian	–0.12	0.72	0.87	–0.85	1.23	0.49
Entry level						
Level 2	[a]	[a]	[a]	1.04	0.17	<.0001
Level 3	[a]	[a]	[a]	0.94	0.17	<.0001
Level 4	[a]	[a]	[a]	1.20	0.15	<.0001
Level 5	[a]	[a]	[a]	6.62	0.60	<.0001

[a] Does not apply.

Table C3

Description of Variables Used in the Analyses (State B)

Variable	Description
FRL	biniary indicator of receiving FRL
Parental education	
ped1	binary indicator of "Did not finish high school" (base category)
ped2	biniary indicator of "High school graduate"
ped3	biniary indicator of "Some additional education after high school, but did not graduate"
ped4	biniary indicator of "Trade or business school graduate"
ped5	biniary indicator of "Community, technical or junior college graduate"
ped6	biniary indicator of "Four-year college graduate"
ped7	biniary indicator of "Graduate school degree"
Ethnicity	
Indian	biniary indicator of "American Indian"
Asian	biniary indicator of "Asian"
Black	biniary indicator of "Black"
Hispanic	biniary indicator of "Hispanic"
White	biniary indicator of "White" (base category)
Homework	
hw1	biniary indicator of "No homework is ever assigned by all their teachers"
hw2	biniary indicator of "Less than one hour each week"
hw3	biniary indicator of "Between 1 and 3 hours"
hw4	biniary indicator of "More than 3 but less than 5 hours"
hw5	biniary indicator of "Between 5 and 10 hours"
hw6	biniary indicator of "More than 10 hours"
hw7	biniary indicator of "Has homework, but does not do it" (base category)
Reading hours	
read1	biniary indicator of "None" (base category)
read2	biniary indicator of "About 30 minutes"
read3	biniary indicator of "About 1 hour"
read4	biniary indicator of "Between 1 and 2 hours"
read5	biniary indicator of "More than 2 hours"

(table continues)

Table C3 (*continued*)

Variable	Description
Use of calculators	
calcul1	biniary indicator of “Never use a calculator in math class” (base category)
calcul2	biniary indicator of “I hardly ever use a calculator in math class”
calcul3	biniary indicator of “Once or twice a month”
calcul4	biniary indicator of “Once or twice a week”
calcul5	biniary indicator of “Almost every day”
Use of computers	
comp1	biniary indicator of “I use a computer at home for school work almost every day” (base category)
comp2	biniary indicator of “Once or twice a week”
comp3	biniary indicator of “Once or twice a month”
comp4	biniary indicator of “Hardly ever”
comp5	biniary indicator of “Never, even though there is a computer at home”
comp6	biniary indicator of “There is no computer at home”
Extraordinary conditions	
extraordinary1	biniary indicator of “Educable Mentally Disabled”
extraordinary2	biniary indicator of “Other Health Impaired”
extraordinary3	biniary indicator of “Speech-Language Impaired”
extraordinary4	biniary indicator of “Specific Learning Disabled”
extraordinary5	biniary indicator of “Autistic”
extraordinary6	biniary indicator of “Severe/Profound Mentally Disabled”
extraordinary7	biniary indicator of “Trainable Mentally Disabled”
extraordinary8	biniary indicator of “Academically/Intellectually Gifted (AIG) ”
extraordinary9	biniary indicator of “Hearing Impaired”

Table C4

Examining Correlates of Redesignation (State B)

	Grade 4			Grade 7			Grade 8		
Variable	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value
Intercept	-1.09	0.46	0.02	-0.34	0.59	0.56	-0.44	0.53	0.41
FRL	-0.44	0.09	<.0001	-0.33	0.10	0.00	-0.35	0.09	0.00
Parental education									
ped2	0.28	0.04	<.0001	0.14	0.04	0.00	0.11	0.04	0.01
ped3	0.23	0.06	0.00	0.17	0.06	0.00	0.18	0.06	0.00
ped4	0.18	0.07	0.01	0.15	0.07	0.03	0.12	0.08	0.13
ped5	0.16	0.04	0.00	0.13	0.04	0.00	0.09	0.04	0.03
ped6	0.25	0.03	<.0001	0.18	0.03	<.0001	0.19	0.03	<.0001
ped7	0.20	0.05	<.0001	0.20	0.06	0.00	0.18	0.05	<.0001
Ethnicity									
Indian	0.45	1.52	0.77	11.05	319.60	0.97	-13.77	447.00	0.98
Asian	-0.56	0.17	0.00	-0.55	0.20	0.01	-0.41	0.18	0.02
Black	-0.32	0.27	0.25	-0.89	0.28	0.00	-0.83	0.28	0.00
Hispanic	-0.49	0.15	0.00	-0.46	0.18	0.01	-0.17	0.16	0.29
Homework									
hw1	-0.64	0.48	0.18	-0.62	0.42	0.14	-0.23	0.41	0.57
hw2	-0.31	0.37	0.40	-0.30	0.29	0.29	0.06	0.28	0.82
hw3	-0.15	0.36	0.68	-0.11	0.29	0.71	0.13	0.28	0.65
hw4	-0.08	0.37	0.83	-0.19	0.31	0.53	0.21	0.30	0.49
hw5	-0.26	0.37	0.49	0.11	0.32	0.74	0.30	0.31	0.34
hw6	-0.06	0.40	0.88	0.35	0.46	0.45	0.72	0.46	0.12
Reading hours									
read_5	0.15	0.19	0.45	-0.30	0.11	0.01	-0.15	0.11	0.17
read1	0.38	0.20	0.05	-0.11	0.13	0.39	0.05	0.13	0.68
read12	0.53	0.21	0.01	0.05	0.16	0.73	0.01	0.16	0.93
read2	0.58	0.23	0.01	0.98	0.24	<.0001	0.49	0.21	0.02

(table continues)

Table C4 (*continued*)

	Grade 4			Grade 7			Grade 8		
Variable	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value	Coefficient	*SE*	*p* value
Use of calculators									
calcul2	0.66	0.18	0.00	0.33	0.49	0.49	-0.08	0.50	0.87
calcul3	0.51	0.19	0.01	0.49	0.49	0.31	-0.36	0.51	0.47
calcul4	0.53	0.18	0.00	0.63	0.47	0.18	0.37	0.45	0.41
calcul5	0.20	0.19	0.30	0.67	0.47	0.15	0.39	0.44	0.37
Use of computers									
comp2	0.15	0.09	0.11	0.40	0.12	0.00	0.24	0.13	0.07
comp3	0.38	0.10	0.00	0.55	0.11	<.0001	0.53	0.11	<.0001
comp4	0.21	0.14	0.13	0.44	0.12	0.00	0.29	0.11	0.01
comp5	-0.18	0.13	0.16	0.33	0.12	0.01	0.13	0.11	0.24
comp6	-0.09	0.17	0.59	0.21	0.17	0.21	0.11	0.15	0.46
Extraordinary conditions									
emental	11.33	341.30	0.97	12.42	452.30	0.98	1.14	1.29	0.38
other	-0.02	0.80	0.98	-0.34	1.08	0.75	11.58	298.10	0.97
lang	-0.33	0.24	0.16	-1.01	0.67	0.13			
specific	-1.28	0.21	<.0001	-1.07	0.17	<.0001	-1.54	0.20	<.0001
autistic	0.18	0.35	0.60	-0.10	0.40	0.81	-0.36	0.40	0.38
smental	-10.25	306.90	0.97						
tmental	-0.44	0.74	0.55				0.30	1.24	0.81
gifted							2.71	1.03	0.01
hearing							-12.52	447.00	0.98

Table C5

Examining correlates of redesignation (State C, Grade 4)

Parameter	Coefficient	*SE*	p value
Intercept	0.38	0.16	0.02
FRL	-0.55	0.10	<.0001
IEP	-0.45	1.13	0.69
PL_504	2.51	0.82	0.00
Title I	-0.85	0.08	<.0001
Migrant status	0.59	0.15	<.0001
Immigrant status	0.28	0.20	0.16
Ethnicity			
Indian	0.75	0.38	0.05
Asian	-0.66	0.19	0.00
Hispanic	-0.46	0.16	0.01
Black	-0.40	0.35	0.26
Extraordinary status			
Dis1	-12.15	289.50	0.97
Dis2	-0.64	1.37	0.64
Dis3	-1.24	1.15	0.28
Dis4	-1.06	1.56	0.49
Dis5	-1.68	1.35	0.21
Dis6	-0.26	1.16	0.82
Dis7			
GIFTED	1.53	0.15	<.0001

Table C6

Examining correlates of redesignation (State C, Grade 8)

Parameter	Coefficient	*SE*	p value
Intercept	1.20	0.18	<.0001
FRL	-0.63	0.09	<.0001
IEP	-1.34	1.18	0.26
PL_504	0.22	0.62	0.73
Title I	-0.45	0.09	<.0001
Migrant status	-0.17	0.17	0.31
Immigrant status	-0.08	0.25	0.75
Ethnicity			
Indian	-1.02	0.40	0.01
Asian	-0.38	0.22	0.09
Hispanic	-0.70	0.18	0.00
Black	-0.79	0.34	0.02
Extraordinary status			
Dis1	-0.83	1.58	0.60
Dis2	0.75	1.27	0.56
Dis3	-0.83	1.20	0.49
Dis4	-0.87	1.60	0.59
Dis5	-0.05	1.32	0.97
Dis6	0.07	1.26	0.95
Dis7	-10.84	317.50	0.97
GIFTED			

Appendix CH3

Tables D1–D14

Investigating states' accommodation practices for ELLs

Table D1

The frequency and percentage of students receiving accommodations (ELLs), (State A)

	Grade 5 (*N* = 8862)			Grade 8 (*N* = 8137)		
Test condition	Math	Science	Reading	Math	Science	Reading
Accommodation	2250 (25.4)	2202 (24.9)	1353 (15.3)	1069 (13.1)	1063 (13.1)	1007 (12.4)
Modification	12 (0.1)	11 (0.1)	16 (0.3)	10 (0.1)	4 (0.1)	5 (0.1)
Regular	6600 (74.5)	6649 (75.0)	7493 (84.6)	7058 (86.7)	7070 (86.9)	7125 (87.6)

Note. The total numbers for each grade represent the combined number of current and redesignated ELLs.

Table D2

The frequency and percentage of students receiving accommodations (all students), (State A)

	Grade 5 (*N* = 33126)			Grade 8 (*N* = 33106)		
Test condition	Math	Science	Reading	Math	Science	Reading
Accommodation	3863 (11.7)	3798 (11.5)	2763(8.3)	2327 (7.0)	2283 (6.9)	2207 (6.7)
Modification	52 (0.2)	49 (0.2)	58 (0.2)	53 (0.2)	42 (0.1)	40 (0.1)
Regular	29211 (88.2)	29279 (88.4)	30305 (91.5)	30726 (92.8)	30781 (93.0)	30859 (93.2)

Table D3

The percentage of students receiving each test version by their English proficiency (State A, Grade 5)

Test	English proficiency level				
	1 (*N* = 299)	2 (*N* = 268)	3 (*N* = 756)	4 (*N* = 2584)	5 (*N* = 1086)
Mathematics					
Accommodation	50.50	51.87	55.69	44.58	11.68
Regular	49.50	48.13	43.52	55.26	88.32
Science					
Accommodation	49.83	51.12	54.63	43.46	11.46
Regular	50.17	48.88	44.58	56.39	88.54
Reading					
Accommodation	28.43	29.48	35.98	25.54	7.42
Regular	71.57	70.52	63.23	74.15	92.58

Note. The percentages may not add up to exact 100% because the "Modification" category was not shown in the table.

Table D4

The percentage of students receiving each test version by their English proficiency (State A, Grade 8)

Test	English proficiency level				
	1 (*N* = 235)	2 (*N* = 212)	3 (*N* = 466)	4 (*N* = 1608)	5 (*N* = 1877)
Mathematics					
Accommodation	46.81	40.09	34.55	26.87	6.50
Regular	53.19	59.91	65.24	72.82	93.39
Science					
Accommodation	46.81	40.09	34.55	26.62	6.45
Regular	53.19	59.91	65.24	73.26	93.50
Reading					
Accommodation	44.68	37.74	31.97	25.44	6.02
Regular	55.32	62.26	67.81	74.44	93.87

Note. Percentages may not exactly equal 100% because the "Modification" category is not included in this table.

Table D5

The Frequency and Percentage of Students Receiving Accommodations (State B)

	Grade 4		Grade 7		Grade 8	
	All students	ELLs	All students	ELLs	All students	ELLs
Accommodation	(*N* = 102385)	(*N* = 6625)	(*N* = 106414)	(*N* = 9450)	(*N* = 107695)	(*N* = 4633)
Extended time	13703 (13.4)	2128 (32.1)	13262 (12.5)	1500 (15.9)	13458 (12.5)	1491 (32.2)
Separate room	12556 (12.3)	1849 (27.9)	10342 (9.7)	931 (9.9)	10043 (9.3)	975 (21.0)
Read aloud	9420 (9.2)	1701 (25.7)	6930 (6.5)	803 (8.5)	6557 (6.1)	789 (17.0)
Translator dictionary	422 (0.4)	229 (3.5)	1043 (1.0)	667 (7.1)	1091 (1.0)	742 (16.0)

Note. The total *N* for ELLs represents the combined sample size of current and former ELLs.

Table D6

The Percentage of Students Receiving Each Accommodation by English Proficiency Level (State B, Grade 4)

	English proficiency level					
Accommodation	1 (*N* = 1085)	2 (*N* = 764)	3 (*N* = 449)	4 (*N* = 184)	5 (*N* = 16)	6 (*N* = 1530)
Extended time	64.70	75.92	83.30	86.96	93.75	12.55
Separate room	53.00	67.54	76.17	84.78	93.75	10.33
Read aloud	48.20	61.52	75.28	77.72	87.50	7.84
Translator dictionary	6.36	8.77	10.02	10.33	18.75	1.50

Table D7

The Percentage of Students Receiving Each Accommodation by English Proficiency Level (State B, Grade 7)

	English proficiency level					
Accommodation	1 (*N* = 710)	2 (*N* = 1166)	3 (*N* = 619)	4 (*N* = 518)	5 (*N* = 140)	6 (*N* = 447)
Extended time	43.66	56.86	67.53	74.13	72.86	11.86
Separate room	22.68	32.59	43.94	54.63	54.29	6.26
Read aloud	15.92	28.64	38.29	49.42	55.00	3.36
Translator dictionary	19.44	25.99	30.21	42.66	51.43	3.80

Table D8

The Percentage of Students Receiving Each Accommodation by English Proficiency Level (State B, Grade 8)

	English proficiency level					
Accommodation	1 (*N* = 662)	2 (*N* = 780)	3 (*N* = 836)	4 (*N* = 531)	5 (*N* = 286)	6 (*N* = 275)
Extended time	41.54	55.13	63.76	72.32	69.93	9.45
Separate room	20.69	33.85	40.19	51.79	58.04	4.00
Read aloud	10.57	27.18	34.33	46.14	53.50	3.27
Translator dictionary	16.62	26.41	35.89	47.08	48.25	3.27

Table D9

The co-administration of different accommodations (State B, Grade 4 ELLs)

Accommodation	Extended time (*N* = 2128)	Separate room (*N* = 1849)	Read aloud (*N* = 1701)
Extended time (*N* = 2128)			
Separate room (*N* = 1849)	1777		
Read aloud (*N* = 1701)	1623	1502	
Trans_dic (*N* = 229)	221	191	179

Table D10

The co-administration of different accommodations (State B, Grade 7 ELLs)

Accommodation	Extended time (*N* = 1500)	Separate room (*N* = 931)	Read aloud (*N* = 803)
Extended time (*N* = 1500)			
Separate room (*N* = 931)	886		
Read aloud (*N* = 803)	767	641	
Trans_dic (*N* = 667)	622	372	335

Table D11

The co-administration of different accommodations (State B, Grade 8 ELLs)

Accommodation	Extended time (*N* = 1491)	Separate room (*N* = 975)	Read aloud (*N* = 789)
Extended time (*N* = 1491)			
Separate room (*N* = 975)	919		
Read aloud (*N* = 789)	744	639	
Trans_dic (*N* = 742)	679	465	364

Table D12

The frequency and percentage of students receiving each accommodation (State C)

	Grade 4		Grade 8	
	All students	ELLs	All students	ELLs
Accommodation	(*N* = 55628)	(*N* = 8248)	(*N* = 57967)	(*N* = 5863)
Oral presentation of entire test	4366 (7.85)	1726 (20.93)	2091 (3.61)	876 (14.94)
Extended timing	3722 (6.69)	909 (11.02)	2622 (4.52)	643 (10.97)
Teacher-read directions only	1109 (1.99)	398 (4.83)	673 (1.16)	220 (3.75)

Note. The total *N* for ELLs represents the combined sample size of current and former ELLs.

Table D13

The percentage of students receiving each accommodation by their English proficiency (State C, Grade 4)

	English Proficiency Level				
Accommodation	1 (*N* = 560)	2 (*N* = 1092)	3 (*N* = 2311)	4 (*N* = 3328)	5 (*N* = 731)
Oral presentation of entire test	48.57	40.66	26.44	10.43	1.92
Extended timing	8.21	13.74	13.50	11.57	7.39
Teacher-read directions only	7.32	8.42	6.10	3.76	0.68

Table D14

The percentage of students receiving each accommodation by their English proficiency (State C, Grade 8)

	English Proficiency Level				
Accommodation	1 (*N* = 297)	2 (*N* = 355)	3 (*N* = 1043)	4 (*N* = 20548)	5 (*N* = 715)
Oral presentation of entire test	48.15	42.82	27.90	10.27	2.52
Extended timing	7.74	9.01	11.98	12.51	9.93
Teacher-read directions only	7.41	6.20	6.62	4.53	0.98

CHAPTER 4:

SUMMARY AND IMPLICATIONS

Mikyung Kim Wolf, Joan L. Herman, Noelle Griffin, and Jinok Kim
CRESST/University of California, Los Angeles

As mentioned earlier, this project entailed two main research phases. The first phase involved exploring existing state data and state practices in English Language Learner (ELL) assessment in order to address important validity issues in assessing ELL students. Informed by the findings from the first phase, the second phase will undertake an intensive research effort focusing on areas of identified need. The present report describes the Phase I research activities and discusses key findings. The assessments of interest in this phase included states' English Language Proficiency (ELP) and content-area tests of math and science. For the three participating states, ELP tests served multiple purposes. The tests were used primarily to determine No Child Left Behind (NCLB; 2002) Title III annual measurable achievement objectives (AMAOs), annually measuring the students' ELP and progress. The states also used the test results for the purpose of redesignating ELL students as Fluent English Proficient (FEP) students. The common purpose of the content-area tests was to measure annual student achievement of content knowledge and report information for Adequate Yearly Progress (AYP). Some states also used the content-area test results as another source to determine ELL students' redesignation status. Professional standards mandate that there be diverse sources of validity evidence to support each purpose for which a test may be used. Such evidence includes the analysis of test content, the item characteristics, internal structure of the test, and relationships between test results and other variables and outcomes. This collected evidence assists in determining the extent to which test results provide appropriate data for each intended use and is informative in refining and improving the use of the assessment.

In examining diverse sources of validity evidence, the researchers explored such areas as (a) the language demands exhibited on the content-area and ELP tests, (b) ELL students' performance on state content-area and ELP tests and the relationships between the two, (c) the criteria for redesignation and the characteristics of redesignated students, and (d) the practice of accommodation use with content-area tests. The first area, an analysis of language demands, provided a fundamental piece of evidence about what the tests were measuring. Additionally, the exploration of multiple biases indices employing differential item functioning (DIF) techniques, provided information on the amount and nature of items that may disadvantage ELL students. The second area, investigating students' performance on multiple measures, provided information on the extent to which the test results were validly used for their intended purposes. The third area, examining the redesignation criteria and the characteristics of redesignated students provided not only important validity evidence on the redesignation criteria based on the test results, but also evidence about ELL students' heterogeneity. The last area, examining the use of accommodations in testing content knowledge raised important questions about the validity and comparability of ELL students' test results. In this chapter, the researchers summarize key findings from each area across the participating states and discuss implications for ELL assessment.

Language Demands in Content-Area and ELP Tests

An examination of the linguistic characteristics of test item content revealed the nature of language demands that ELL students face in test-taking. It also provided a lens through which to view the constructs being measured by the tests. Considering that a new generation of ELP tests has been developed to assess academic English language, the language of schooling, rather than strictly social language, the analyses concentrated on the characteristics of academic language presented in the ELP tests. In addition, the researchers examined the correspondence between the characteristics of academic language in content-area tests compared to ELP tests. On the basis of previous theory and research, a content analysis tool was devised to systematically analyze the academic language features and language demands of the tests. This instrument encompassed a wide range of academic language features in lexis, grammar, and discourse, as well as the language demands from non-linguistic features such as visual presentation and type of comprehension skills needed to solve an item. While this content analysis instrument was employed to describe the language demands of test items qualitatively, a DIF technique also was utilized to quantitatively investigate the test items that potentially disadvantaged ELL students.

The results of the linguistic analysis suggested that academic vocabulary was the most prominent feature characterizing the language demands across both content-area and ELP tests. Academic vocabulary in this research was categorized into general academic, context-specific, and technical vocabulary. The last two categories included the vocabulary which was related to specific content areas, such as math and science. For both math and science tests, more content-related academic vocabulary was evident than general academic vocabulary. One interesting finding was that the math tests, which are typically assumed to present lower language demands than science, contained a wide variety of general academic vocabulary. This finding suggests that vocabulary may be a key source of difficulty for ELL students taking math and science tests in English. The linguistic analysis of items which functioned differentially for ELL students than for other students, those exhibiting DIF, also supported the importance of academic vocabulary. In general, the identified DIF items against ELL students contained a greater amount of language and more academic vocabulary compared to non-DIF items. Particularly, DIF items were found to have more general academic vocabulary, rather than technical vocabulary, suggesting that ELL students might have more difficulty with general academic vocabulary than with technical vocabulary. One possible explanation for this finding is that technical vocabulary is explicitly taught to all students as part of the content in instruction. This finding implies the importance of examining ELL students' opportunities to learn, for instance, uncovering the ways that ELL students are exposed to and instructed on both general and specific academic language.

Among the academic language features included in the analyses, there were very few occurrences of academic grammatical and discourse features in either of the content-area tests. This finding may be explained, in part, by the inherent nature of test language. The test items tended to contain formulaic expressions with simple WH-questions (e.g., which of the following is…, what is…), which limited the use of diverse grammatical and discourse features. In fact, one state explicitly mentioned in their technical manuals that the test developers attempted to reduce the linguistic complexity by using simple grammar and sentence structure.

Although a similar trend of the prominence of academic vocabulary was found across the content-area tests, a closer look at the results demonstrated that the science tests were more linguistically demanding than the math tests, particularly in terms of the amount of the language (i.e., length, the number of the words per item), the amount of academic vocabulary, and the use of nominalization. This difference may be reflected in the DIF analysis, in which more DIF items were detected for science than in math in State A where both math and science test data were available. The language demands were perceived to be higher for the Grade 8 tests than the Grade 4 tests for both content areas. The amount of language, of academic vocabulary, and the number of complex sentence structures per item increased for the higher grade level regardless of the content areas.

The linguistic analysis of one state's ELP test yielded a similar pattern to that of the content-area tests. While few academic grammatical and discourse features were observed, academic vocabulary was highly prevalent in the sections of listening, reading, and writing on the ELP test. Although the speaking section questions were presented primarily in social or daily language, the scoring guideline implied that students were expected to produce both social and academic language in their responses. Comparing the content-area and ELP tests, slightly higher academic language demands were found in the content-area tests. More academic vocabulary and technical vocabulary were observed in the content-area tests in general. This finding was not surprising in that the ELP test developers described the construct of their test as including both social and academic language. The observation of both general academic and technical vocabulary in the ELP test thus reflected the current reform efforts: that is, the newly-developed ELP tests attempt to measure the language proficiency that ELL students will need in an academic context.

In addition to the linguistic features, the analyses also addressed the type of comprehension that items elicited. The results indicated that the majority of the ELP reading items were limited to extracting or identifying specific literal information contained within the passages. The results suggest that while the passages in the reading section may be comparable to passages encountered in academic settings, the reading test items addressed only a limited range of tasks that usually occur in those same academic settings. The findings further suggest that despite the attempts of newly-developed ELP tests to measure academic language, these tests may not adequately sample the range of reading tasks that students encounter in academic settings. Developers of such tests should consider expanding their test items to include a wider and more representative sample of different classroom reading tasks. This would assure that inferences of ability based on the tests generalize to the academic tasks of which the students are learning. The finding also implies that providing more detailed score reporting for ELL teachers and decision makers (e.g., sub-scores based on the type of different comprehension skills for a reading section) instead of providing a total score for each domain will be more useful.

Another interesting finding is the variation of the language demands in the tests across states, even within the same content area. For example, the math test of one state contained more language, more academic vocabulary, and more complex sentence structures than that of the other state at the same grade. The DIF results also showed that the number of DIF items against ELL students varied more across states than across grade levels or content areas. The state which contained the least number of DIF items specifically mentioned that they applied ‘Universal Design’ and ‘plain English’ in their test development. Furthermore, the researchers found that the highest numbers of DIF items were encountered in the test identified as making the greatest language demands. This finding raises a critical issue about the effects of linguistic simplification, or the use of ‘plain’ English in state content-area tests. On the one hand, it can be argued that purposive test designs reduce the linguistic load and thus offer fewer obstacles for ELL students taking content-area tests. On the other hand, it can also be argued that such linguistic simplification or use of ‘plain’ English may result in the elimination of or significant reduction in the number of context-specific and technical vocabulary items in the test. Consequently, the test may not assess the vocabulary knowledge that is part of the construct. This would lead to construct under-representation, and thus raise questions about the validity of the score-based interpretations of math and science achievement from that test.

A major caveat of attempting to reduce language demand, either through universal design, or the use of ‘plain’ English in content-area tests, then, is that these not be altered so as to completely eliminate the language features, like context-specific and technical vocabulary, that are likely to be part of the construct. Considering the prominence of academic vocabulary that was found on the content-area tests of the participating states, it would appear that explicit item writing rules and specific principles for test construction did not result in an appreciable reduction in the context-specific and technical vocabulary. On the same note, the DIF analyses indicate the academic vocabulary was clearly associated with DIF items against ELLs, which suggests that the real issue with the test items may lie in inadequate opportunity to learn, rather than in language bias.

ELL Students’ Performance in Content-Area and ELP Tests

In comparing the state tests performance of current ELL students, non-ELL students and former ELL students who had been redesignated as English proficient, the research team was provided with opportunities to explore another validity issue in ELL assessment. That is, by investigating the relationship between students’ scores on ELP and content-area tests the researchers were able to provide evidence of concurrent validity in determining the level of English proficiency needed for success in mainstream classrooms and attainment of content proficiency.

The examination of the achievement scores on the math and science tests yielded consistent results across states. As expected, ELL students performed significantly lower than non-ELL students, particularly on science compared to math tests. The performance gap between ELL and non-ELL students was greater at the upper grades than at the lower grades. Although a number of factors may be involved in this performance gap, the language demands of the test items seemed to be one factor, considering the findings from the linguistic and DIF analyses of the items in these tests. As reported earlier, the items on science tests contained a greater amount of language than did the math tests within one state, and the items were more linguistically demanding at the upper grades.

What is particularly noteworthy is that despite the variation of the linguistic complexity of the tests across different states, the achievement gap trends were consistent across states: ELLs showed a wide performance gap, former ELL students beyond 2 years of redesignation performed comparably or as well as non-ELL students, and more recently redesignated students performed significantly better than current ELL students, under certain settings, performing comparably relative to non-ELL students. A number of plausible reasons may account for this finding, including the differences between states in ELL students' backgrounds, the effects of stringency of redesignation criteria, effective opportunities to learn, and the effects of language demands or the accommodation used in the content-area tests.

As mentioned earlier, a primary use of state ELP tests is to determine ELL students' readiness to exit from ELL status. Thus, it was anticipated that a significant, positive relationship between ELP and content test performance would provide a source of validity evidence about using the ELP test results for making such redesignation decisions. Relationships between ELL students' performance on ELP and content-area tests was found to be strongly positive across all states, regardless of the content area, grade, or school. Coupled with the results from the linguistic analysis, this finding provides evidence to uphold the claim that the construct of the ELP test included language ability necessary to deal with academic materials. Considering that the traditional ELP tests failed to predict ELL students' readiness for a mainstream classroom, this result is promising, suggesting that the new generation of ELP tests has begun to address previous limitations. In addition, these relationships support the use of ELP test scores as one source to determine the redesignation status.

Even though this was not a direct issue of study, one notable, recurring theme from the analyses was the heterogeneity of the ELL student population: results varied for students at different levels of ELP, for those at different levels of reading skill, and for redesignated students. For example, a closer look at the relationship between the ELP and content-area test scores showed a quadratic relationship, meaning that relationships differed for ELL students at different levels of ELP. That is, expected levels of academic achievement tended to increase rather slowly for students with lower ELP scores, while the levels tended to increase more rapidly for those achieving higher ELP scores. Furthermore, when the DIF analysis was conducted for high and low levels of ELL students based on their performance on the state reading test, more DIF items were detected across all states' tests, compared to the DIF analysis when grouping ELL and non-ELL students. As might be expected, this finding underscores the serious challenge to test validity for ELL students who lack rudimentary English reading skills.

In addition, the examination of redesignated students' performance further highlighted the heterogeneity of the ELL group. Some redesignated ELL students performed better in mainstream classrooms than their English-only peers; others failed to improve their ELP sufficiently and showed a wide performance gap relative to non-ELL students. This substantial variation was also present depending on the time of redesignation. In all three states, students who were redesignated over 2 years ago on average performed as well as or better than non-ELL students. However, more recently redesignated students showed mixed results across states as described above, possibly the result of differences in the stringency of states' redesignation criteria, which is discussed below. Nonetheless, these findings highlight both the presence of important differences within the ELL group and the importance of identifying and examining ELL student subgroups.

Redesignation of ELL Students

Redesignation is one of the most influential decisions for ELL students. Once a student is exited from ELL status, she or he no longer receives special instructional services and takes content-area tests under standard conditions without the use of accommodations. Furthermore, redesignation decision-making has substantial impact on accountability system results, considering that schools must meet annual goals for ELL student subgroups (e.g., AMAO). Hence, validating the criteria used for redesignation decisions is of paramount importance.

The mixed results about the performance of recently redesignated students noted earlier seemed to be associated with differences in states' redesignation criteria. That is, for the state that required students to meet a relatively higher level of ELP for redesignation, recently redesignated students on average performed comparably to their non-ELL counterparts. On the other hand, even though the tests are not directly comparable, for the two states that allowed ELL students to exit with lower ELP levels on their tests, the recently redesignated students tended to perform lower than non-ELL students and in general did not reach proficiency on content-area tests. As discussed in Chapter 3, further research is needed to evaluate states' redesignation criteria, including the interaction between the cut scores and redesignation criteria, the readiness of students for mainstream classrooms, and the consequences of redesignating students.

Performance patterns among redesignated students also were intriguing. Even though recently redesignated students still struggled with content-area tests under certain settings, those who were at least two years beyond redesignation appeared to overcome learning obstacles, in that no-longer-monitored redesignated students on average performed better than non-ELL students. These results also suggest the importance of longitudinal studies to investigate ELL students' progress and the effects of the ELL-related polices, including the influence of NCLB (2002), new assessments, and redesignation criteria.

The Use of Accommodations for ELL students in Content-Area Tests

The last area the analyses attempted to address was the use of accommodations for ELL students in content-area tests in the three states participating in the study. The research team was interested in conducting analyses to shed light on (a) the extent to which the content-area tests accurately measure ELL students' content knowledge and skills when accommodations are provided, (b) whether accommodations inadvertently alter the construct to be measured, and (c) how accommodations influence students' performance at different levels of language proficiency. A review of the documents regarding approved accommodations in the three states revealed that there was considerable variation across and within states. However, serious flaws in available state data limited the exploration of the use or effects of these accommodations. The use of accommodations was noted for only a small proportion of ELLs, raising questions about the completeness of data reported by local agencies to the state. Data on specific accommodations used was available for only two of the three states. Based on these data, the most common accommodations in State B were extended time of test administration; test administration in a separate room; reading tests aloud; and the use of a translated dictionary. In State C, the most frequently used accommodations were oral presentation of the entire test, extended test time, and teachers reading aloud the directions. It is noteworthy that there is virtually no research support for the validity of any of these commonly used accommodations, and in fact available research raises questions about such validity (see Abedi, Hofstetter & Lord, 2004; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006). Considering the potential effects of accommodations on the validity of ELL test scores, there is an immediate need for states to develop both specific research-based guidelines regarding the provision of accommodations for ELL students and, a uniformly structured database for local districts to report accommodation use for individual students. This, in turn, will allow practitioners and researchers to monitor accommodation use and further examine the validity of their scores from accommodated tests in order to improve the ELL assessment practices.

Figure 3 summarizes the findings of Phase I that have been discussed and provides preliminary recommendations for practitioners and researchers. Providing validity evidence to support intended uses of an assessment is an ongoing process. As seen in Figure 3, a number of areas need further investigation to improve the validity of ELL assessment practices. The areas that the research team has explored in the first phase will also need to be regularly revisited as part of the validation process. This process will eventually help improve the understanding of ELL students' performance and ELL assessment systems.

Findings from Phase I (Sources of Validity Evidence)	Implications
Content-area test items in math and science contained a wide variety of academic vocabulary.	Vocabulary knowledge may be a significant factor influencing ELL students’ performance.
DIF items against ELL students tended to include more general academic vocabulary, not technical academic vocabulary.	ELL students may have less access to opportunity to learn general vocabulary compared to non-ELL students.
More DIF items were detected between high and low reading proficient ELL students, compared to between ELL and non-ELL students.	Language demands may be a critical factor impacting the students’ performance. Furthermore, the finding signifies the importance of addressing different needs within an ELL group.
One specific ELP test was found to include both social and academic language characteristics in its construct to measure; however, the extent to which academic language features were included in the items varied across four modality sections (reading, listening, speaking, and writing). Furthermore, the item types were limited in terms of the types of comprehension skills required of the student.	The new ELP test reflects the current reform efforts of assessing academic language proficiency; however, it still needs to include a wider range of language tasks to encompass those that are encountered in academic contexts.
ELP test scores were strongly and positively associated with ELL students’ content-area scores.	The use of ELP tests scores as a redesignation criterion is supportive.
Recently redesignated ELL students’ performance was widely varied across states.	State’s redesignation criteria may play a role in this mixed result.
State variations were evident in the linguistic features of the test items, redesignation criteria, and accommodation types used.	Comparability across and even within states should be made with caution.

Recommendations for Research and Practice
(Targeted Areas to Improve)

(1) Test development: Provide a clear item-writing guideline to avoid unnecessary linguistic complexity.

(2) Comparability of test scores: Use caution in comparing each state’s AYP reports considering the different characteristics of the measures being used.

(3) Redesignation criteria: Examine the criteria to ensure it reflects the ELL students’ readiness for mainstream classroom.

(4) ELP test use: Examine the construct being measured; Examine cut scores for redesignating an ELL student; Consider reporting a detailed score profile for diagnosing and instructional purposes.

(5) Longitudinal research: Continue research on the progress of ELL students for further examining validity evidence for the use of tests and redesignation criteria.

(6) Accommodations: Request for an immediate need of systematic accommodation data reporting and management; Request for a principled provision of accommodations while attending to the different needs within an ELL group.

(7) OTL: A measure of OTL for ELL students is imperative and should be utilized as part of the state’s validation process.

Figure 3. Summary of the findings and recommendations.

References

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved November 21, 2006, from http://www.centeroninstruction.org/files/ELL1-Interventions.pdf

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).