An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document

CSE Report 611

Alison L. Bailey and Frances A. Butler
CRESST / University of California, Los Angeles

October 2003

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

Project 4.1 Developing Measures of Academic English Language Proficiency
Alison L. Bailey & Frances A. Butler, Project co-Directors, CRESST / UCLA

The work reported herein was supported under the Educational Research and Development Centers Program, PR / Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

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

Preface

An earlier version of this design document (July 2002) provided the basis for discussion for an invitational meeting of experts on K-12 language education hosted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA on July 22, 2002. The goal of the meeting was for the participants to react to the CRESST research plan for operationalizing academic language and provide feedback and guidance for refining the plan. The participants at the meeting brought a wide range of expertise to bear on the issue and were articulate in emphasizing the critical nature of the work given the current national educational climate under the *No Child Left Behind Act* (2001). Participant comments and suggestions have been incorporated in this final deliverable.

Colleagues from CRESST/UCLA also attended the meeting. Below is a complete list of the participants and their affiliations. We wish to express our sincere appreciation to all participants for a productive and highly engaging exchange of ideas. We also thank the participants in the session *Measuring and Supporting English Language Learning in Schools* at the 2002 Annual CRESST Conference for their comments and suggestions following the presentation based on this work.

Invited participants

Fred Davidson
University of Illinois

Lily Wong Fillmore
University of California, Berkeley

Margo Gottlieb
Illinois Resource Center

Jan Mayer
California Dept. of Education

Mary McGroarty
Northern Arizona University

Maria Seidner
Texas Education Agency

Catherine Snow
Harvard University

Marguerite Ann Snow
California State University, Los Angeles

CRESST participants

Eva Baker
Co-director CRESST/CSE

Joan Herman
Co-director CRESST/CSE

Frances Butler
Project Co-director

Alison Bailey
Project Co-director

Malka Borrego
Project Coordinator

Carol Lord
California State University, Long Beach

Jamal Abedi
Senior Research Associate

Zenaida Aguirre-Munoz
Senior Research Associate

Christy Kim-Boscardin
Senior Research Associate

Priya Abeywikrama
Graduate Student Researcher

Francisco Herrera
Graduate Student Researcher

AN EVIDENTIARY FRAMEWORK FOR OPERATIONALIZING ACADEMIC LANGUAGE FOR BROAD APPLICATION TO K-12 EDUCATION: A DESIGN DOCUMENT [1]

Alison L. Bailey & Frances A. Butler

CRESST/University of California, Los Angeles

Abstract

With the *No Child Left Behind Act* (2001), all states are required to assess English language development (ELD) of English language learners (ELLs) beginning in the 2002-2003 school year. Existing ELD assessments do not, however, capture the necessary prerequisite language proficiency for mainstream classroom participation and for taking content-area assessments in English, thus making their assessment of ELD incomplete. What is needed are English language assessments that go beyond the general, social language of existing ELD tests to capture academic language proficiency (ALP) as well, thereby covering the full spectrum of English language ability needed in a school setting. This crucial testing need has provided impetus for examining the construct of academic language (AL) in depth and considering its role in assessment, instruction, and teacher professional development. This document provides an approach for the development of an evidentiary framework for operationalizing ALP for broad K-12 educational applications in these three key areas. Following the National Research Council (2002) call for evidence-based educational research, we assembled a wide array of data from a variety of sources to inform our effort. We propose the integration of analyses of national content standards (National Science Education Standards of the National Research Council), state content standards (California, Florida, New York, and Texas), English as a Second Language (ESL) standards, the language demands of standardized achievement tests, teacher expectations of language comprehension and production across grades, and the language students actually encounter in school through input such as teacher oral language, textbooks, and other print materials. The initial product will be a framework for application of ALP to test specifications including prototype tasks that can be used by language test developers for their work in the K-12 arena. Long-range plans include the development of guidelines for curriculum development and teacher professional

[1] We wish to thank Malka Borrego for her insightful assistance in the preparation of this document, specifically the preliminary analyses of national, state and ESL standards, and Carol Lord for helpful comments and suggestions and her guidance for preliminary analysis with Priya Abeywickrama and Francisco Herrera of science texts. We wish as well to thank Laquita Stewart and Lawana Woods for their administrative assistance throughout the data analysis and document preparation process. We also acknowledge the expert assistance of Wade Contreras for final editing and formatting and Danna Schacter for final proofing.

development that will help assure that all students, English-only and ELLs alike, receive the necessary English language exposure and instruction to allow them to succeed in education in the United States.

1. Introduction

The number of students in U.S. schools for whom English is a second language has grown steadily during the past two decades. During this time, educators have struggled to implement approaches that help to ensure both quality instruction and ensure valid and reliable assessments for all students. Unfortunately, English language learners (ELLs) often enter the school system without the requisite English language skills to benefit from the mainstream curriculum, and thus, in the past, were often excluded from accountability systems. Now, however, with the *No Child Left Behind* (NCLB) *Act* (2001a; 2001b), focus on adequate yearly progress in math and reading for all students and special emphasis on ensuring that ELLs make steady progress in acquiring English virtually guarantees the inclusion of ELLs in state accountability systems.

Including ELLs in accountability systems is not without challenges, however. For example, the language demands of content-area assessments may be so great for ELL students as to invalidate the assessment of their content knowledge. Ultimately, we have not known if the performance of ELL students primarily reflects their language abilities or their content knowledge. Thus, including ELLs in the testing process, knowing that the interpretation of test scores may be invalid, is problematic. However, to exclude ELL students, in our view, is unacceptable, because if ELL students are not tested, information on their achievement is, in effect, absent from any decision-making that impacts their school careers.

A major issue at this juncture is how to determine if the requisite English language skills for demonstrating content knowledge on assessments have been acquired by ELL students. Most existing English language proficiency tests do not assess the type of language students must understand and be able to use in the mainstream classroom and on standardized content tests. Existing language tests tend to assess more social everyday language, rather than the more formal academic language (AL) of the classroom and content tests.[2] Thus a student could perform well on such a language test and not have the necessary language skills for academic tasks. There is, then, an important assessment gap between the type of English an ELL student may know and be able to use—that tested on current English language

[2]This is not to minimize the importance of social language for successful school and personal outcomes. Indeed the level of linguistic sophistication by which we navigate everyday informal situations would suggest we also need to foster student growth in the area of social language development (Bailey, forthcoming).

development (ELD) tests—and the language critical to school success (Stevens, Butler, & Castellon-Wellington, 2000).

States have perhaps demonstrated their recognition of this mismatch or shortfall in the usefulness of existing ELD assessments by making performance on a language arts subsection of content-area assessments part of their redesignation criteria. This practice, however, hinges on circular logic—some states and districts are apparently using the reading subtest of a content-area test to help determine if ELL students are indeed ready to take the same content-area test *that includes the subtest*. A test expressly developed to test knowledge of the kind of English language skills (i.e., AL) needed for school success would seem a more logical approach in the fair assessment of ELL students (Bailey & Butler, 2002).[3] Tests that tap academic language proficiency (ALP) could serve as an intermediate step between English language tests that primarily focus on social use of language and the content-area assessments that contain the language of academic settings (see Figure 1.) Such tests could fill the assessment gap mentioned previously.

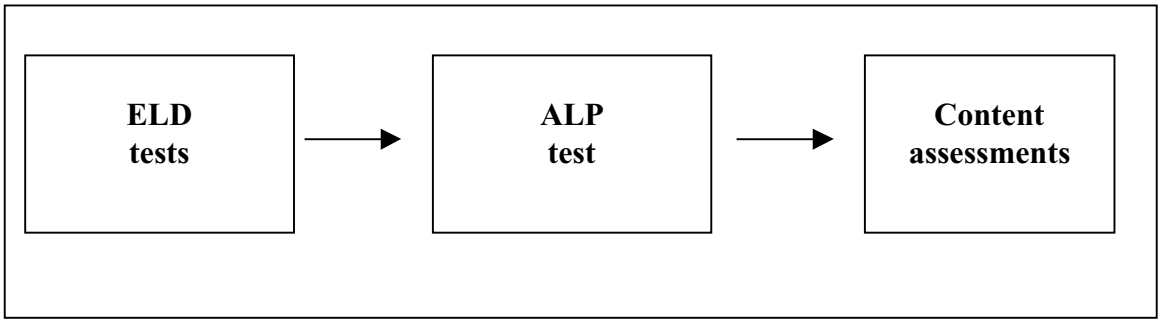

Figure 1. The role of ALP assessment in a socially responsive accountability system.

1.1 Overview of the ALP Framework

We propose here a framework for the development of ALP tests, related curricula, and professional development materials. The framework is evidence-based in that we are documenting the type of evidence needed to operationalize ALP. The most immediate goal of this work is to characterize ALP for test development purposes so that test specifications and prototype tasks can be created to reflect language usage in academic settings.

[3] By *school success* we mean access to curriculum materials, understanding teacher talk, participation in class with teacher and peers, and ability to handle content assessments—standardized and teacher diagnostic assessments and measures of progress, etc.

The initial evidence of the impact of language demands in academic settings was present in studies of differential student performance where native speaking or English-only (EO) students tended as a group to outperform ELLs (Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; August & Hakuta, 1997; Butler & Castellon-Wellington, 2000; MacSwan & Rolstad, in press). Researchers have inferred from this finding that the English language is a barrier to student demonstration of content-area knowledge. Our plan is to go beyond this inference and follow new lines of inquiry that will allow us to more directly specify the ALP demands critical to school performance and that must underlie the development of specifications and prototype tasks.

The strength of an evidence-based approach to ALP is that it provides a mechanism for capturing not just the linguistic features of language—vocabulary, syntax, and discourse and the features of language use within and across content areas—but also the linguistic demands created and/or assumed by a broader array of stakeholders. That is, the broader educational community provides evidence of academic language in national content standards (e.g., National Science Education Standards of the National Research Council, National Council for the Social Studies), state content standards (e.g., California Department of Education, Texas Education Agency), and standardized achievement tests (e.g., the Stanford Achievement Test, the Iowa Tests of Basic Skills). To facilitate this effort, classroom language demands must be systematically identified and prioritized according to specific criteria. In determining the basis of evidence for the language demands, we will continue to concentrate on what would count as evidence of ALP in the classroom. We are interested in capturing the language students actually encounter in school through input such as teacher oral language, textbooks, and other print materials.

In sum, our current bases of evidence for specifying language demands (explicated further in Section 3) are the following:

1. Empirical studies of ELL/EO student performance and language demands of content and ELD assessments
2. The language prerequisites assumed in national content standards[4]
3. The language prerequisites assumed in state content standards
4. The language prerequisites assumed in English as a Second Language (ESL) standards
5. Teacher expectations for language comprehension and production

[4] The standards have assumptions about language expectations that we have to infer from the text.

6. Classroom exposure to AL, including teacher talk and textbooks

1.2 Content Areas and Grade Levels

Achieving operational descriptions of ALP will require research in a range of academic contexts and will necessitate attention to at least two dimensions of potential variation: content-specific subject matter and grade level (Butler & Stevens, 2001). The first dimension of this work will cut across content areas (science, social studies, math, and language arts) to investigate a common core of AL present in all subjects as well as to identify ALP uniquely required in individual content areas. Figure 2 shows the hypothesized relationships between common core AL and content area-specific AL.

The second dimension will cut across grade levels (elementary through high school) to determine what ALP is developmentally appropriate for the specific grade levels and grade-level clusters. Figure 3 shows the hypothesized relationships

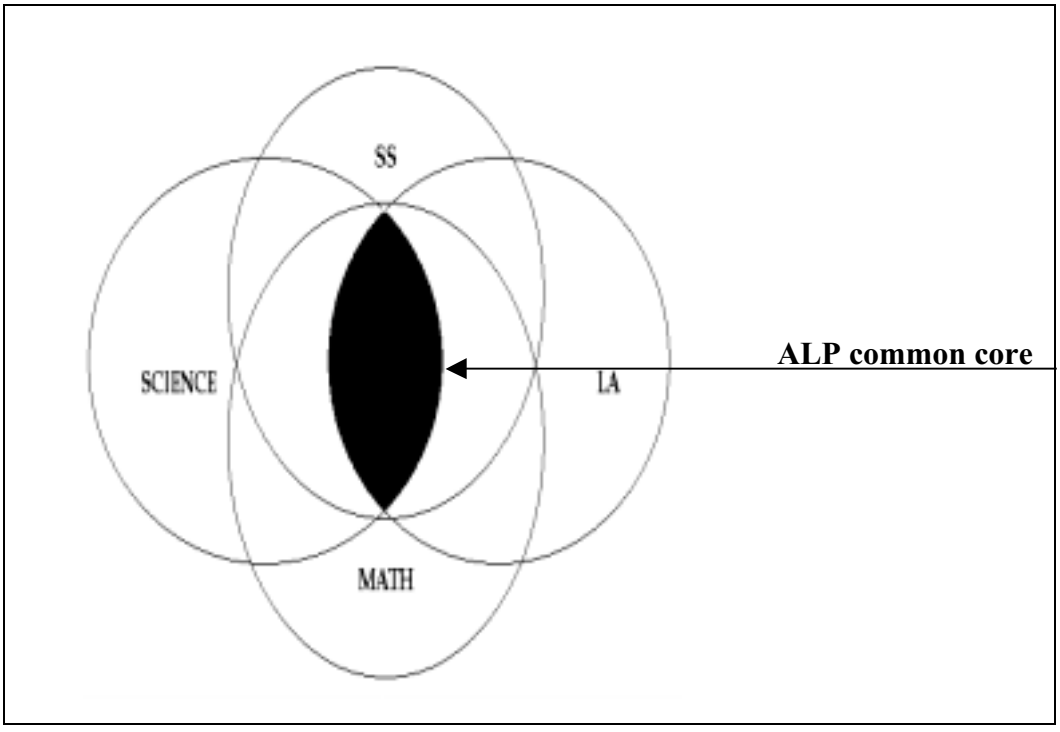

Figure 2. Hypothesized relationships between common core AL and content area-specific AL in the domain of math, science, social studies (SS), and language arts (LA).

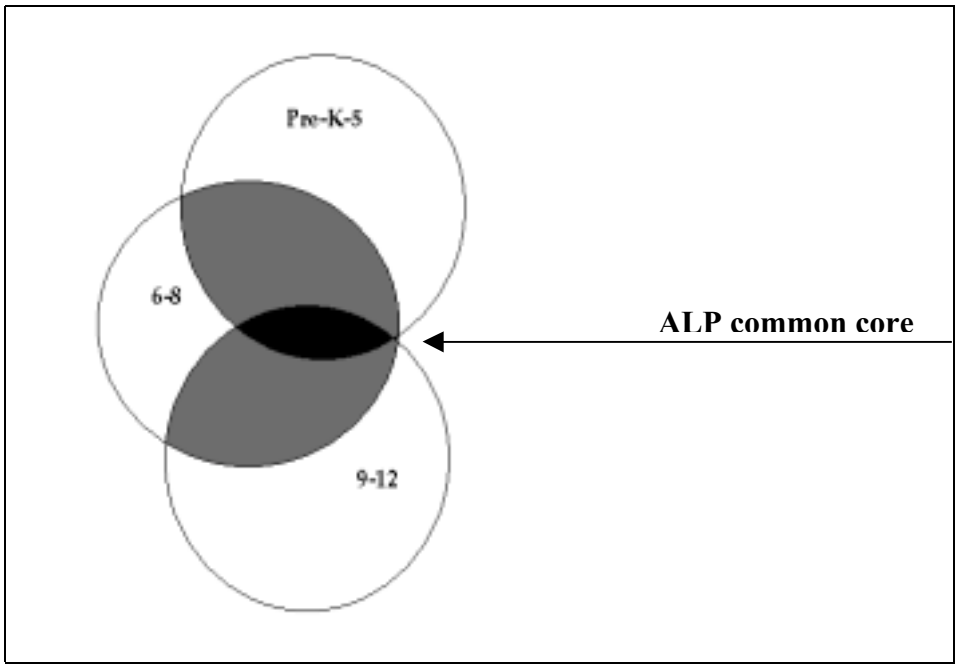

Figure 3. Hypothesized relationships between common core AL across grade clusters and cluster-specific AL.

between common core AL across grade clusters and cluster-specific AL. Both types of information are necessary to provide an evidentiary basis for the AL demands that must be assessed to determine ELL readiness for participation in the mainstream curriculum and for taking content tests in English.

The types of evidence laid out previously will allow a comprehensive characterization of ALP across content areas and grade levels and thereby provide a strong evidential base for the development of specifications for ALP tasks and curricula. Throughout this design document we concretize design decisions and chosen approaches with fifth-grade science content, classroom language, texts, standards, etc., where possible, for illustrative purposes.

2. The Academic Language Construct

2.1 A Working Definition of Academic Language

There is a growing literature on AL beginning with Cummins' (1980) notions of Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). Butler and Bailey (2002) point out, "The literature increasingly defines AL in more detail by looking at vocabulary, syntax, discourse, and language functions as they cut across different contexts of use and cognitive and textual

demands, most recently with emphasis on socio-cultural / psychological dimensions" (p. 3).

Chamot and O'Malley (1994) define AL in general terms as "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). A student who is academically proficient in a language (first or second) can use global and domain-specific vocabulary, language functions, and discourse (rhetorical) structures in one or more fields of study to acquire new knowledge and skills, interact about a topic, or impart information to others.

Solomon and Rhodes (1995) take a sociolinguistic view that defines AL in terms of register. Johns (1997) identifies a register of English used in professional books and characterized by the specific linguistic features associated with academic disciplines. Short (1994) has documented the range of language functions found in social studies classes, including explanation, description, and justification functions. Gibbons (1998) focuses on the *intertextual* nature of classroom language and emphasizes the importance of language use across the skill areas—the need to integrate oral and print language skills in classroom activities. The integration is important because oral and print language have different characteristics and consequently different demands (Schleppegrell, 2001).

Cummins (1980), mentioned previously, provides a multidimensional view of language proficiency by including cognitive demands alongside the formal / informal distinction in his characterization of oral language, a perspective reemphasized by Cummins (2000). The distinction contrasts Basic Interpersonal Communication Skills (BICS), acquired and used in everyday interactions, and Cognitive Academic Language Proficiency (CALP), acquired and used in the context of the classroom. Scarcella (in press) takes a broad view of AL building on Kern's (2000) model of academic literacy, which includes linguistic, cognitive, and sociocultural / psychological dimensions.

These definitions of academic language have been integrated with definitions used in earlier CRESST work (Bailey, 2000a; 2000b; Butler, Stevens & Castellon-Wellington, 1999; Stevens et al., 2000) that sought to characterize AL at the lexical (vocabulary), syntactic (forms of grammar), discourse (rhetorical), and language function levels. Most recently we have taken an empirical approach to language use

in the classroom (Bailey, Butler, LaFramenta, & Ong, 2001; Bailey, Butler, Borrego, LaFramenta, & Ong, 2002).

To achieve further specificity, academic language is defined in our work as language that stands in contrast to the everyday informal speech that students use outside the classroom environment. More specifically, at the lexical level, Stevens et al. (2000) identify three categories of words: (a) high-frequency general words, those words used regularly in everyday contexts; (b) nonspecialized academic words, those academic words that are used across content areas; and (c) specialized content area words, those academic words unique to specific content areas. Specialized content-specific language includes the conceptual terminology of, for instance, science (e.g., *osmosis, igneous, biodiversity*). The nonspecialized language that cuts across content areas is a form of AL that is not specific to any one content area, but is nevertheless a register or a precise way of using language that is often specific to educational settings. For example, formal vocabulary, such as *examine* and *cause* that students encounter at school, contrasts with everyday vocabulary such as *look at* and *make* that they encounter in less formal settings (e.g., Cunningham & Moore, 1993), and use of a simple preposition like *by* can take on the unfamiliar meaning of *according to* in a sentence like *sort by color* (Bailey, forthcoming).

Previous CRESST work cited above adds two additional features to the CALP definition of AL. First, AL implies the ability to express knowledge by using recognizable verbal and written academic formats. For example, students must learn acceptable, shared ways of presenting information to the teacher so that the teacher can successfully monitor learning. These formats or conventions may or may not be explicitly taught as part of a curriculum, but their use is expected of all students. Indeed, the opportunity to display knowledge is also an important feature of the classroom (Cazden, 2001) and may prove as crucial to students as the opportunity to learn. Second, AL use is often decontextualized whereby students do not receive aid from the immediate environment in order to construct meaning. There is little or no feedback on whether they are making sense to the listener or reader, so students must monitor their own performance (spoken or written) based on abstract representations of others' knowledge, perspectives, and informational needs (e.g., Menyuk, 1995; Snow, 1991). Indeed, the testing situation would appear to epitomize the decontextualized characteristic of AL—attempts to elicit feedback would constitute cheating.

To be "proficient," then, a student needs to be able to interact in situations in which there may be fewer opportunities to negotiate meaning or to use context than one might find in many social settings. Ultimately, students must learn to recognize and make sense of the various conventional ways academic material will be presented to them and expected of them in decontextualized settings. For example, ELL students who are reasonably proficient speakers of everyday (BICS) English, but who have not had as extensive an exposure to complex syntax, idioms, and depth of vocabulary (e.g., antonyms, synonyms, etc.) as native speakers of English of the same age, may find lessons more challenging because their language proficiency levels do not match the demands of the academic language of the classroom. ELL students may easily understand an interrogative sentence such as "Where do you think the fins on a whale are?" because it is the sort of language that they may encounter in everyday speech or in widely read materials, such as newspapers and magazines. However, the same request for information may be conveyed in very different language in a classroom setting. Bailey et al. (2001) document the following example: "What is your best estimate of where the fins are located on a whale?" This version of the question includes not only the unfamiliar use of *best* (to mean *most accurate*) and the formal *are located* for *are*, but also an embedded wh- question in the second clause: "…where the fins are located on a whale?"

Clearly, the acquisition of AL is critical for effectively negotiating the content and interaction of the classroom. Collectively the research cited here has documented the existence of an AL phenomenon. However, there have been few attempts to operationalize the concept sufficiently for utilization in assessment development.

3. Evidentiary Basis for Operationalizing Academic Language Proficiency

To move from descriptions of the AL construct to the development of tasks and test items that measure its proficiency, we need to adopt a framework that documents the sources of information that will feed into the operationalization of the construct. The recent National Research Council (NRC) (2002) call for evidentiary bases to educational research in general provides one of the main motivators (along with need to reflect the many different contexts within which AL arises) to assemble a wide array of data from a variety of sources to operationalize the ALP construct; thus we integrate within one model (Bailey & Butler, 2002) content assessment language demands, national standards of professional

organizations (e.g., NRC), state content standards (California, Florida, New York, and Texas), ESL standards, teacher expectations of language comprehension and production across grades, and analyses of classroom language and textbooks. As mentioned above, in each case where possible, we have focused on the science content area and the fifth grade (or clusters including fifth grade) in order to provide in-depth examples. However, the following framework we describe can be applied to all grade levels or grade clusters, to the four major content areas (i.e., language arts, math, science, and social science), as well as all four modalities of language (i.e., listening, speaking, reading, and writing). Figure 4 schematizes the different types of evidence we have been able to identify for the purpose of operationalizing proficiency in AL. We turn now to examine each of the sources.

3.1 ELL/EO Student Performance and Language Demands of Content and ELD Tests

Research on the assessment of ELL students to date has focused primarily on the validity of content assessments and the role of accommodations to provide a valid and reliable measure of ELL school performance. As mentioned earlier, numerous studies of differential student performance have found that native speaking or EO students tend as a group to outperform ELLs (Abedi & Lord, 2001; Abedi, Lord, et al., 2000; August & Hakuta, 1997; Butler & Castellon-Wellington, 2000; MacSwan & Rolstad, in press). Abedi, Leon and Mirocha (2000), for example, found that ELL student performance suffers in those content area subtests that are thought to have greater language complexity than others. These findings suggest that student language proficiency impacts performance on standardized content assessments according to the nature of the English language demands of the content area assessed. While this finding is not surprising and has already been widely assumed, Abedi, Leon, et al. (2000) provide statistical evidence, across multiple school districts and across multiple states, of weaker ELL performance in contrast with much higher EO performance in general and in content areas that have greater language demands (e.g., language arts).

Furthermore, Butler and Castellon-Wellington (2000) examined student performance on language proficiency assessments and concurrent performance on content assessments to provide baseline data for identifying a threshold of language ability needed to determine whether ELLs' content assessment performance is considered a valid measure of their content knowledge. They found evidence of

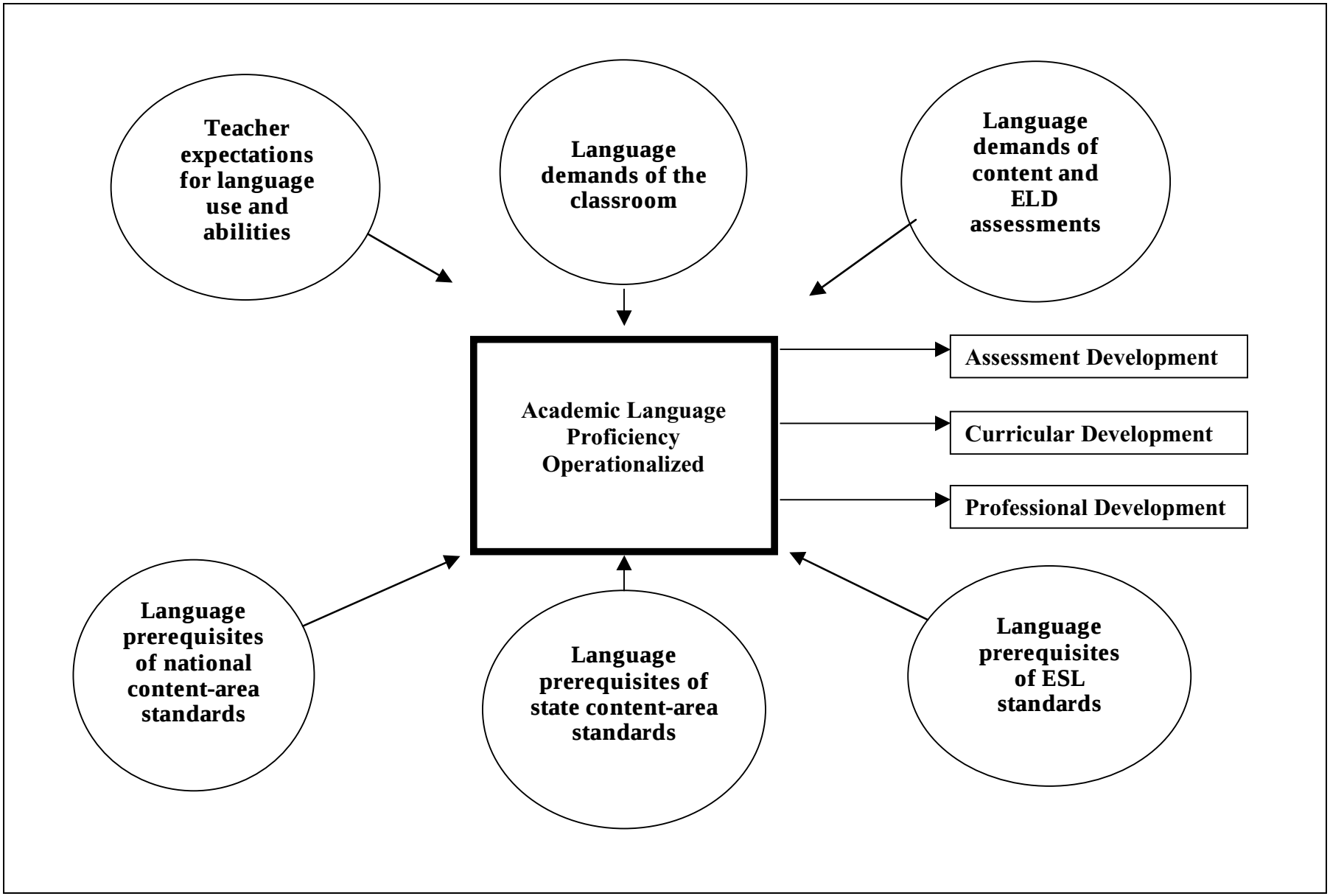

Figure 4. Evidentiary bases for the operationalization of Academic Language Proficiency (ALP) (based on Bailey & Butler, 2002).

some ELL students, those designated in this report as fluent English proficient (FEP) and redesignated fluent English proficient (RFEP), performing on a par with EO students (50 national curve equivalent [NCE] or above) on the content assessment subtests, suggesting that for those students, performance on the content test reflected their content knowledge. Other ELL students, however, those designated as limited English proficient (LEP), while scoring in the competent range on the language assessment, did not score on par with EO students on the content assessment subtests. These results suggest that further differentiation of language performance in the upper proficiency range on ELD assessments is imperative to determine if these particular students are struggling with language, content, or both.

The research to date shows clear relationships between language proficiency and performance on content tests for ELLs. However, these findings are strongly tempered by a number of major concerns and considerations. While language is likely a dominant factor for ELLs, student English language proficiency does not explain all the variation found in content performance. Additional background factors also play a role in predicting student performance, namely parent educational level and family income level. Opportunities to learn and display knowledge of both content and academic language are also potentially important predictors of student performance. These factors are not, to our knowledge, included in the extant data sets supplied by school districts, nor are they factors that are easily measured and quantified. While previous research has been useful in answering some of the questions about assessment validity, serious limitations have curtailed the ability to draw useful conclusions about the assessment of ELL students. These limitations are laid out more specifically in Abedi, Bailey, and Butler (2000), as:

a. *The lack of uniformity in defining ELL students.* Terms such as ELL, FEP, LEP, RFEP, and bilingual are used in the national dialogue about students who are acquiring English as a second language. Unfortunately, these terms are often operationalized differently across school sites within a district, across districts, and across states, causing difficulties with respect to data interpretation. For example, some districts and states have redesignation criteria that are based on different measures or different cut scores. Furthermore, students are not redesignated at the same time during the school year across districts and states. Therefore, student designations may not be accurate at the time research data are compiled or collected.

b. *The lack of comprehensive data sets.* Often existing data files do not include important data elements such as student ethnicity, parent education, and family income because the data were not collected for research purposes. In addition, item-level data are often not available.

c. *Limitations regarding the aggregation of small numbers of ELL students across districts.* While we are interested in ELL/EO comparisons at the national level, the variability in student background variables and the designation criteria across districts do not allow us to combine data sets in order to make large-scale comparisons. This issue is even more critical when studying assessment issues by sub-groups of ELLs.

d. *The limitations of language assessments.* A major weakness in the study of ELL assessment is the lack of a standard instrument that can be used to assess English language proficiency in a manner that is parallel to the way language is used on the content assessments.

This last limitation brings us to the crux of the purpose of developing ALP assessments. The number and degree of difficulty of lexical, syntactic, and discourse demands students encounter on content-area tests has already been documented (Bailey 2000a; Stevens et al., 2000). However, there is a mismatch between these language demands and the lesser demands found in ELD tests (Stevens et al., 2000). The content of currently available commercial language proficiency tests is not adequate to measure the level of language proficiency necessary for taking standardized achievement tests and for full participation in the mainstream classroom.

3.2 The Language Prerequisites of National Content Standards

National standards for the different content areas have been created by national organizations such as the National Research Council and the National Science Teachers Association. We focused our analysis on the National Science Education Standards published by the National Research Council (1996). These science standards comprise six chapters of standards for science teaching, professional development, assessment, content, education programs and systems. The chapter that focuses on science content was examined for the language that students are expected to comprehend and use in science classrooms.

These national standards for science content are organized around what students are expected to know. There are eight standards presented in grade

clusters; namely K-4, 5-8 and 9-12, with the exception of the standards for *unifying concepts and processes* which cut across all grades in acknowledgement that this knowledge continues to develop over time. We focus on the 5-8 cluster in order to characterize the national science standards in some depth.

The 5-8 content standards include seven additional categories, *science as inquiry, physical science, life science, earth and space science, science and technology, science in personal and social perspectives,* and *history and nature of science.* We examined the implicit and explicit assumptions about the prerequisite language necessary for students to meet the standards. Language of the standards that describes linguistic behaviors required of students was identified. These language descriptors of the standards provide some measure of what students are expected to do in service of cognitive operations, e.g., processing information in a specific manner or for a specific purpose. Thus, the fundamental abilities and concepts that underlie the *science as inquiry* category have students *focus; clarify* questions; *inquire; design* an investigation; *make* observations; *organize* and collect data; *take* measurements; *use* appropriate methods (e.g., math) and tools (e.g., computers); *interpret; summarize* and *describe* data; *report* on inquiries by writing, drawing and graphing; *communicate* scientific explanations; *describe* and *explain* findings; *identify* cause and effect; and *critique* and *consider* alternative explanations. The additional six categories document the content material (facts, scientific processes, etc.) that students are expected to know rather than characterize in linguistic terms how students will demonstrate knowledge.

3.3 The Language Prerequisites of State Content Standards

The implicit and explicit statement of the prerequisite language made in the content standards of four states with large proportions of ELL students (California, Florida, New York, and Texas) provides further specificity about desired language proficiency at the different grade levels or clusters. As mentioned earlier, the language of standards reveals assumptions about what students must be able to do with language in the different content areas at the different grade levels. For example, an analysis of state content standards revealed that state science standards share common language functions and descriptors. Examination of the manner in which the standards for the four states refer to the language students will need at the elementary level revealed that students must be able to *analyze, compare, describe, observe,* and *record* scientific information. At the middle school level, students in all

four states were required to *compare, explain, identify,* and *recognize.* At the high school level, the four states share just three language-related descriptors of desired student language use: *describe, explain,* and *recognize.*

These verbs provide a greater or lesser degree of explicitness in terms of what language is expected from students. That is, beyond the use of the verbs themselves, in order to compare science concepts and objects, students will likely need to know related vocabulary (e.g., other verbs such as *contrast*), already know or know rules for creating comparative adjectives (e.g., *bigger, wider),* be able to use syntax that sets up parallel constructions to make a comparison (e.g., phrasing such as *the green one is better than the yellow one…because the green one is clean and the yellow one is dirty*), and understand the purpose of comparison discourse in the academic setting (e.g., functions to explain science concepts, justify selections, argue a point of view, etc.).

Looking across the four states at one grade level (sixth),[5] we find that the language used in both California and Florida state standards is confined to what students should *know* or *understand.* The standards reflect the content students need to master, but there is no reference for how students should be able to demonstrate knowledge or the type of language students should be able to use to do this. New York and Texas on the other hand, organize the standards by the science content students need to know as well as how students need to demonstrate the knowledge with sample tasks. Indeed both of the latter states provide a greater variety of descriptors in terms of language use than the former two states. For example, New York addresses the language associated with Earth science at the sixth grade by requiring students to be able to *explain* seasonal changes on Earth. Students must be able to create an Earth model and *describe* the arrangement, interaction and movement of the Earth, moon, and sun. Also, students need to *plan* and *conduct* an investigation of the sky to *describe* the movement of celestial bodies. Students need to *explain* how the air, water, and land interact, evolve, and change, and *describe* volcano and earthquake patterns, and the rock cycle and weather and climate changes. This is evident when students *graph* temperature changes and *make* a record of earthquakes and volcanoes and interpret the patterns. The New York State standards provide sample tasks for teachers to use to judge whether a skill is evident in a student. For example, students are asked to construct an explanation based on

[5] This was the grade closest to our chosen example grade (fifth) for which we were able to make suitable comparisons given that the grade clusters utilized by the four states cut across the elementary and middle school levels in some instances.

their observations of a melting ice cube using sketches and a written description of what happens. We propose capitalizing and expanding on these efforts by New York State to inform the development of specific ALP prototype tasks.

3.4 The Language Prerequisites of K-12 ESL Standards

ESL standards are also examined for assumptions about requisite language abilities for students to be considered advanced English language learners. At present we have examined the Teachers of English to Speakers of Other Languages (TESOL) (1997) standards that represent a national set of standards to which many state ELD standards are aligned (New York State). Future work will need to look at the individual state ESL/ELD standards to determine how close the alignment between the TESOL standards and those of the states really are. The TESOL standards are clustered in three grade groups. For the purposes of this report, we have collapsed across clusters because most of the language descriptors cut across all three grade clusters. The standards make a distinction between language identified by the TESOL standards for social and personal interaction (i.e., Goal 1) and standards for classroom interaction for academic purposes (i.e., Goal 2). There is some overlap with a number of descriptors (35) that relay the kind of language necessary for each of the goals. These include *ask, clarify, express, imitate, listen, negotiate, participate, request,* and *respond,* which are required across all three grade clusters as well.

In both the social and academic contexts, students are expected to know a wide array of additional language functions: however, these are far more extensive in the academic standards (64 language-related descriptors in the academic language standards and 30 in the social language standards). Verbs unique to the academic language standards that cut across the three grade clusters include *analyze, contrast, define, elaborate, hypothesize,* and *justify*. Verbs unique to the social language standards that cut across the three grade clusters include *communicate, describe, elicit, engage, restate, recite,* and *talk.*

A final important caveat here is that any effort to align an ALP assessment to national standards, individual state content standards, or ESL standards is undermined by the absence of data to support the choice of language repertoire, content coverage, and levels of difficulty that they adopt. For example, the TESOL standards identify *describe* as the kind of language needed to meet social language standards but not the academic language standards. However, this language use

was seen to be used for academic purposes in observations of science classrooms (Bailey et al., 2001; Bailey et al., 2002). Justification for adoption of specific standards has thus emerged as a concern. McKay (2000) has called for a sound theoretical base for the construction of standards in the ESL arena. Schleppegrell (2002) following suit has made a critical analysis of the California ELD standards showing discrepancies within them. Standards may be a reflection of ideals for school reform by a particular group of educators rather than being a source of evidence for what and how students are being taught. Thus, we caution against taking standards at face value, but rather recommend using them as a helpful starting point for identifying the kinds of language possibly found in academic settings.

3.5 Language Expectations of Teachers

Very little study of the linguistic expectations of mainstream teachers has been made and even less is known about the language expected of ELL students as they are reclassified and enter mainstream classrooms. However, it is conceivable that teachers have definite expectations for the linguistic sophistication of their students. Hicks (1994) states:

> When asked to provide a comment or answer a question during a whole-class discussion, children are generally expected to provide an explanation or description, or sometimes to recount a personal experience. Fundamental to the expectations of a system of formalized schooling is the notion that children can engage in socially appropriate kinds of discursive activity (p. 225).

One possibility for the paucity of empirical information in this area is as Wong Fillmore and Snow (2000) point out—speakers, in this case teachers, are rarely explicitly aware of their language expectations of others. Anecdotally, from our own experiences gleaned from presentations to teachers on AL instruction, we find that there is much confusion about what is meant by AL and language expectations in general. As Wong Fillmore and Snow go on to argue, teachers (ESL and regular education) need more professional development in the area of language and linguistics in order to better serve student language needs in the classroom. This is an issue to which we return in section 4.3.

What little systematic empirical work is available suggests that teachers do at least pay close attention to language facility as a prerequisite to being able to succeed in school (Boyer, 1991), even while they may have difficulty articulating concretely what their expectations are. In one of the few studies of explicit teacher expectations for age-appropriate language skills in EO students, Hadley, Wilcox,

and Rice (1994) found that preschool teachers expected more talking from children than kindergarten teachers. The latter expected students to be quiet and pay attention, especially during teacher-directed activities. While this study provides important insights into expectations for when student talk should occur during the transition to a more academic environment in the early grades, it does not investigate teacher expectations for level of language sophistication, nor how teacher expectations differ as children move into the higher elementary grades and beyond.

Solomon and Rhodes (1995) were, however, able to survey 300 ESL teachers on their perspectives of academic language. This survey at least provides us with some idea of teacher expectations in the higher grades. These teachers identified academic language in terms of the language needed by students to effectively participate in classroom activities that included both discrete aspects, such as specific vocabulary (e.g., greater than, plus, minus, equals and label), and parts of speech (e.g., definitions of nouns, verbs, etc.), as well as the major functions of language in the classroom, such as summarizing, categorizing, comparing, and contrasting.

We suggest further study of teacher expectations of mainstream and ELL students across the grade levels to help determine the language demands ELL students face in the opinions of the teachers who currently teach them or who eventually will.

3.6 Language Demands of the Classroom

By describing the language embedded in oral and written classroom discourse, particularly characterizing the AL of mainstream content-area classes that can serve as a baseline for AL expectations, it will be possible to create a framework for English language assessment that is aligned with the language of classroom instruction and curricula materials. Available work comes primarily from the area of atypical language development, where researchers have found it difficult to articulate norms for language development in school-age children. This difficulty is due perhaps to the individualistic nature of language development after the pre-school years (i.e., the range of particularistic AL contexts to which children are exposed by their choices of content classes in later grades, Nippold, 1995).

In the past, few studies examined the language of actual lessons over an extended period of time in order to carefully document and conceptualize academic language as it actually occurs (Solomon & Rhodes, 1995). Recently, researchers have

at least begun to make sample selections of language in the classroom in both the oral and print domains. For example, Schleppegrell (2001) contrasts such linguistic features as lexical density and clausal structures requiring nominalization and sentence embedding in spoken discourse and a textbook selection. By illustration, she finds the textbook language to be much more lexically dense and to use nominalization strategies (e.g., *the formation of sedimentary rocks*) rather than pronominalization more commonly found in the oral discourse (e.g., *they, it*).

In a series of observations of fourth- and fifth-grade mainstream science classrooms, Bailey et al. (2001) developed matrices to illustrate teacher and student talk as they intersected different contexts of instruction (science concepts, vocabulary, application instruction) with language functions, repair strategies, and classroom management. In each of the three instructional contexts, teachers used primarily four language functions—description, explanation, comparison, and assessment—and two repair strategies, clarification and paraphrasing. Overt instruction of specialized vocabulary occurred more often than overt instruction of nonspecialized vocabulary and frequently took the form of giving examples. The use of definitions repeatedly required teachers to clarify, and the use of synonyms often required teachers to paraphrase whole sentences. In addition, some lexical items were found to have precise scientific meanings as well as non-academic discourse usage. Failures in communication were repaired with strategies of clarification and paraphrasing. Paraphrasing, however, was not always used for repair. Rather, it could be used to circumvent a possible breakdown in communication in both science concept instruction and academic vocabulary instruction.

Student talk data revealed five predominant functions of language in the classroom—explanation, description, comparison, questioning, and commenting. As with teacher talk, students were mainly explaining, describing, and comparing scientific concepts. They also asked the teacher questions of scientific substance and added commentary in isolated cases. In terms of repair strategies, students both requested and provided clarification, and showed understanding of classroom management by their appropriate responses to teacher questioning.

Print material selection by teachers and writing products by students were also examined. Print selections had varying levels of linguistic demand as measured by syntactic complexity, sentence length, and amount of academic vocabulary. The materials also contained multiple language functions and embedded academic vocabulary (both nonspecialized and specialized) similar to that observed in teacher

talk. While some materials required the student to react in writing, most simply provided information. The skills and level of academic language found in student writing varied across samples (i.e., essays, test answers, and letters). In general, students used different styles of writing for different tasks (e.g., using a distinctive dictionary style in writing definitions.) Students also used multiple language functions within their writing, primarily explanation and description. In addition, academic vocabulary, mostly specialized, appeared in student samples. Students exhibited varying skill in utilizing and adapting academic language from print material in their essays.

3.7 Evaluation of the Six Evidentiary Bases

The six bases from which we draw evidence of AL will not be given equal weight in the operationalization of ALP due to variation in the quality of information that they may yield. Specifically, given the need for validation of content and ESL standards, and the impressionistic nature of teacher expectations, the expert panel (personal communication, July 22, 2002) made the recommendation that CRESST focus on textbook analysis and analysis of classroom discourse in extant datasets as they become available. It was argued that this type of data will yield more consistent and higher quality data than standards and expectations data. Current work underway with the standards will continue to inform our effort, but the main focus will be on textbook and teacher language data. Further systematic examination of textbook language features and teacher language will provide a fuller picture of the language demands across both content areas and grade levels. Textbook analysis continues with the investigation of language functions first examined in our classroom observation research, with the additional analytic approach of Schleppegrell (2001) and others to also reveal the range of clausal complexity, lexical density, proportion of specialized academic words, etc., by content area and grade. In addition, differences in linguistic demands and features across genres (e.g., expository, evaluative, directive, etc.) have been noted. In section 4.1, we illustrate how preliminary findings with fifth-grade science textbooks generated by these types of analyses can feed into ALP test specification efforts. These attempts to capture the linguistic characteristics of a range of genres go further than the existing ELD assessments available that have tended to narrowly focus on narrative skills in their attempts to sample extended discourse skills.

4. Broad Application to Education

In this section of the document, we focus on the creation of a framework for application of ALP in the areas of assessment development, curriculum development, and teacher professional development. Though the main focus will be assessment development, the other areas are included for completeness and to begin to outline a comprehensive treatment of ALP in educational arenas.

4.1 Assessment Development

The *Standards for Educational and Psychological Testing*, published by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) (1999), provide a comprehensive set of guidelines for test development and use and form the basis of recent efforts by a number of agencies and organizations providing standards for assessment practice in K-12 education and the testing of individuals of diverse linguistic backgrounds more specifically. The CRESST *Standards for Educational Accountability Systems* (Baker, Linn, Herman & Koretz, 2002) take the joint AERA, APA, and NCME standards as a given and offer 22 guidelines for good testing practices, one of which pertains directly to assessment practices with ELL students, and four others of which offer pertinent direction. Standard 12 focuses on issues of test validity with ELL populations who may differ from the general student population, on whom tests are typically normed. ALP assessments and tasks will need extensive norming with students who are representative of the ELL students who will be expected to take the tests or accomplish the tasks. Validity in the test development process is crucial in this circumstance so that fair and equitable decisions can be made about the educational programming provided to ELL students. Baker, Linn, and Herman (2002) have argued that it is the responsibility of the states to request validity studies and data from the commercial test developers with whom they contract.

CRESST Standards 4, 7, 9, and 11 are also pertinent to ALP assessment development and use because they (a) provide direction on issues of inclusion of all students (as now required by NCLB Act [2001a]), (b) suggest the use of multiple tests, especially important with ELLs, in order to measure progress across a range of different contexts that may only occur in different testing contexts, (c) point out validity issues regarding the multiple purposes to which tests are often put, such as

placement, progress, and diagnostic purposes, and (d) require that the technical quality of assessments be high. This is especially pertinent now that federal sanctions may be made against schools that do not demonstrate adequate yearly progress in ELD for ELL students, as well as content area knowledge for all students.

The Council of the Great City Schools and The National Clearinghouse for English Language Acquisition (2002) have recently offered a new set of standards expressly for the assessment of ELLs. The *Assessment Standards for ELLs* document contains seven standards. We briefly review five that appear most relevant for ALP assessment, some with greater or lesser degrees of overlap with the CRESST assessment standards. Standard 1, like CRESST Standard 4, makes inclusion of ELL students in content assessment a priority. Standard 2 calls for assessments that will allow students to demonstrate their language proficiency and academic achievement. Standard 4 raises the issue of norming all assessments used with ELL students, although it is not clear to us at this stage of inquiry on whom ALP assessments should be normed—redesignated ELL students or English-only students. The problem with norming assessments on redesignated ELL students is that the redesignation criteria are made through use of the kinds of ELD assessments we have judged inadequate for the task of determining student proficiency in the language of academic settings. Standard 5, like CRESST Standard 7, calls for multiple measures to be used so that decisions are not made on the basis of one test type. Finally, Standard 6 advocates individual longitudinal data collection on ELL students. Indeed this is a requirement now laid out in NCLB Title III (2001b) so that states will track progress at the student level. This seems particularly pertinent given the complexities of attempting to draw meaningful inferences from data aggregated at the school or district level due to high rates of ELL student mobility.

When done well, test development is a complex process that begins with determining what construct or constructs and associated skills are to be assessed. The process generally begins with a needs analysis that helps set the parameters for how the test(s) is to be used. A framework document is then developed to characterize the construct being tested. The framework draws from the research literature in the relevant fields and identifies gaps that may require additional research to help solidify the content base for next steps. The construct articulated in the framework must then be operationalized for actual test development. That is, the content in the framework document must be synthesized and translated into a

working format/paradigm (facilitated by the creation of matrices) that will lead to test specifications, which in turn will guide task development.[6] The documentation of this process provides the validity foundation for the test(s) being produced (Bachman, 1990; Davidson & Lynch, 2002).[7]

The needs analysis for the current effort has identified shortcomings (discussed in Sections 1 and 3.1) of widely used ELD tests. Performance on those tests does not provide a sufficiently broad indication of whether students have the English language skills necessary for success in the school setting; thus, there is a need at both the conceptual and practical level to move beyond the assessment of general/social language use exclusively to the assessment of ALP as well. Within school systems, a range of decisions must be made about ELL language ability from intake decisions (whether students should receive ELD services or go directly to the mainstream classroom), to diagnostic evaluation, to monitoring progress, to redesignation decisions and readiness screening for content assessments. While this range of decision types is too broad for a single instrument, a framework document that articulates an ALP construct can serve as a base for the development of multiple instruments for different purposes and can also provide the content base for curriculum development and teacher training. Each individual application would set parameters for specific needs.

The framework document for the CRESST test development work will broadly define the construct of ALP drawing on the data from the evidence bases discussed in Section 3. At this time, we hypothesize that in this context, the student's ability to understand and use English across the four modalities and across content areas will be important. We envision a grade-cluster approach (K-2, 3-5, 6-8, 9-12) that will capture the developmental nature of ALP. Further, we anticipate combining modalities (reading/writing, listening/speaking, etc.) to capture the integrative nature of the modalities in the school setting (Gibbons, 1998). In addition, we acknowledge the interrelationship of language and cognition in the academic contexts being examined. While our focus is specifically on language, language is used for a purpose often to achieve a function, to explain, to interpret, etc., and thus

[6] Following Davidson and Lynch (2002), we use *task* to mean both individual *items*, such as a multiple-choice item and constructed-response *tasks*, such as producing a writing sample.

[7] See Butler, Weigle, Kahn, and Sato, 1996; Butler, Weigle, and Sato, 1993; Kahn, Butler, Weigle, and Sato, 1995; and Weigle, Kahn, Butler, and Sato, 1994, for the description of a process for the California Department of Education that led to specifications and prototype tasks anchored to the Adult English-as-a-Second-Language Model Standards for California.

is interwoven with cognition such that it is nearly impossible to exclude one from consideration of the other. For us the challenge is to create prototype ALP tasks that respect the interwoven nature of language and cognition but that have as their goal the assessment of language ability.

Our intent is for the framework document to provide guidance for moving from the conceptual to the ideal to the practical. We recognize that while operational constraints, both financial and logistical, are inherent in every testing situation, it is the ethical obligation of those who do test development and test administration to be principled in accommodating operational constraints. Minimally there must be a clear link between test tasks and the constructs being assessed. There must be adequate sampling of the skills being tested to allow for reasonable inferences about student ability. Administration and scoring must be carried out systematically by trained personnel. The unique needs of ELL students must be addressed in terms of test accessibility concerns, such as test presentation format, administration and response conditions, cultural interferences, etc. (see Kopriva, 2000 for further discussion). All of these factors influence the validity of inferences drawn from assessments (e.g., Messick, 1989).

Our goal is to provide a framework with guidelines for establishing linkage to example specifications and prototype tasks such that teachers, agencies, and organizations needing to develop ALP assessments will be able to take our working characterization of ALP as a point of departure for developing their own test specifications and tasks.[8] The example specifications we provide will include the components shown in Box 1, which reflect earlier specification work by Butler et al. (1996). (See also Davidson & Lynch, 2002; Popham, 1978; and Turner, 1997, for important discussions on the development and use of test specifications.)

[8] *Prototype tasks* are defined here as tasks that have been tried out and revised and ultimately determined to produce ratable samples of student language (productive skills) or indications of student comprehension (receptive skills) and can serve as models for task writers.

Box 1: Test Specification Components for Task Development

1. General description—indicates behavior or skill to be tested.
2. Prompt attributes—detail what will be given to test taker, including directions.
3. Response attributes—describes in detail what test taker will do.
4. Sample item—explicit format and content patters for item or tasks that will be produced from the specs.
5. Specification supplement—additional information including rating/scoring procedure, time allotment, etc.

Box 2 provides an example of how components of the evidentiary framework we described in Section 3 can be utilized to provide information to feed into test specifications. We can use results of the data analysis conducted on the textbooks to (a) inform the writing of test specifications (i.e., determine what characteristics of language demands need to be exemplified in test items) and (b) provide guidelines for evaluating and refining the language of the prototype ALP tasks themselves. Specifically, we will be able to align tasks to the data generated in the textbook analyses by conducting the same analysis on newly created ALP tasks. These tasks can be edited until language features and properties similar to the textbooks are obtained.

Box 2: Example of Using an Analysis of Textbook Language Demands to Feed into Test Specifications and Later to Inform Validation of Test Prototypes

The preponderance of fifth-grade science textbook selections analyzed across a variety of types and genres (e.g., expository reading passage, directives for lab activities, assessment) have a mean sentence length of between 10 and 18 words, and about 40% of these sentences required the processing of complex syntax (e.g., embedded clauses). Thus fourth-grade prototype ALP tasks will need to reflect these linguistic features. Fourth-grade ELL students comprehending this type of language can be rated as advanced/proficient. Items reflecting the language of each grade (or grade cluster) will enable us to determine at what grade level equivalent a student is comprehending language.

Box 3 provides an example of test specification components applied to the creation of a draft prototype ALP task for the science content area in fifth grade. This task is designed to measure student ability to comprehend an oral description (input), infer meaning from linking different parts of the description (a cognitive operation), and subsequently produce an explanation of what has been heard

Box 3: Example of Test Specification Components Applied to a Draft Prototype ALP Task

Domain:

Oral Language: Comprehension of *Description* (input) and production of *Explanation* (output).

General description:

The task will test the test taker's ability to *listen and comprehend* the language of description and in turn *produce* the language of explanation.

Prompt attributes:

The test administrator will read aloud to the test taker a short passage with specified attributes that give sentence length and complexity, breadth and depth of vocabulary, etc., as determined by textbook and classroom discourse analysis (see also Box 2). The passage and explanation test question will be crafted to elicit the language of elaborated explanation. The task will have an academic theme or focus, but all information to provide an accurate response to the prompt will be included such that no specific content-area knowledge outside the prompt will be required.

Response attributes:

The test taker will respond orally and will produce the necessary language to achieve the goals of the task, which include (a) demonstrating understanding the language of description via responses to a series of comprehension questions, (b) using cognitive processes to infer relevant information from the descriptive passage, and (c) producing a fully elaborated explanation in response to the explanation question (see scoring guidelines under specification supplement below).

Sample item/task read aloud by test administrator:

I am going to read you a short passage and then ask you some questions about it.

A teacher specifically told a group of students to carefully place their experiments in a safe location in the classroom. One student placed his glass bottles very close to the edge of his desk. When the teacher turned around she was angered by what she encountered.

Box 3: Continued:

Who told the students to place their experiments in a safe location? (comprehension question)

Where did one student place his experiment? (comprehension question)

Who was angered? (comprehension question)

Explain as much as you can why the teacher was angry? (explanation question)

Specification supplement (scoring guidelines):

(a) Test taker will need to accurately answer comprehension questions about the description heard (scored correct/incorrect regardless of language sophistication and fluency), (b) test taker will need to infer that the teacher in the prompt was angry because she saw that the student put his experiment in the wrong place, and (c) test taker will need to use the language of explanation (vocabulary, syntax, and discourse) to demonstrate that understanding to the tester.

Rubric for scoring explanations:

Level 1 - Response is characterized by an incomplete and/or incorrect answer.

Example response 1a:
The teacher was angry.

Example response 1b:
The teacher was angry because the student put the bottle on his desk.

Level 2 - Response is characterized by a generally correct answer but the test taker has failed to elaborate how the inference (the bottle is in a dangerous position and could fall easily) was drawn.

Example response 2a:
He didn't follow directions.*

Example response 2b:
The teacher was angry because the student did not follow directions.*

Box 3: Continued:

Level 3 - Response is characterized by use of appropriate language to demonstrate a fully elaborated explanation. The test taker is able to infer that the teacher in the prompt was angry because the student put his experiment in the wrong place. The test taker demonstrates the use of the language of explanation to demonstrate that understanding (e.g., use of conditional tense for hypothetical events).

Example response 3:
The teacher was angry because the student did not follow directions. He put his bottle very close to the edge of the desk, which is a dangerous place because the bottle could fall and break.

*Note that in casual conversation, the explanations in Response #2a and #2b would be considered adequate. This highlights the difference between social uses of language and academic uses of language that hold speakers accountable for their claims, requiring them to verbally construct an argument citing evidence or logical conclusions to back up such claims. Moreover, these responses may be acceptable in many classrooms. Teachers may not require students to elaborate on their explanations in a way that overtly demonstrates to the teacher the necessary inferencing processes or steps in logical thinking.

(output). This example also illustrates how information procured in the textbook analysis should inform the character of the linguistic features (e.g., sentence length and complexity) of the oral language prompt. The brief description of events that comprise the oral prompt has the science classroom as its theme but does not require knowledge of specific science content or concepts. However, comprehension of the task does assume the general (basic) knowledge that gravity will cause the glass bottle to drop to the floor if it falls off the table and that glass is a brittle matter that can shatter on contact with a hard surface. The series of comprehension checks can be used to indicate whether or not a student has initially understood the description. The student's performance on these comprehension questions must be taken into account when interpreting the student's performance on the explanation question. These checks will help assure a valid score of ability in explanatory language use by ruling out lack of comprehension as a source of interference. Moreover, examples and trial items before the presentation of the test item can be used to demonstrate to students that fully elaborated explanations (as exemplified in Box 3 by Example Response 3) are expected in order for a student to obtain a maximum score on such a task.

We consider this a good draft prototype because it is both authentic and parsimonious. We believe it is authentic to the kinds of language students encountered during observations of fifth-grade classrooms. Among a number of speech functions including comparison and explanation, students heard teachers provide descriptions of scientific terminology and describe scientific tasks and task-related materials. In turn students were asked to give their own explanations of scientific concepts (Bailey et al., 2001). Evidence of the need for familiarity with these particular language functions is further garnered from preliminary analyses of fifth-grade science textbooks. The inference that students must draw in this prototype ALP task is a cognitive operation that is frequently required in school (cf. observations of classrooms; the national, state, and ESL standards; and textbooks and standardized tests). Inference requires the student to do something with language, namely make new meaning from already given meanings (see Box 3 for further details). The draft prototype task is parsimonious because it capitalizes on two modalities of language required of students in the classroom while serving to capture both input and output dimensions of language (see Figure 1, Bailey et al., 2001). Specifically, the draft prototype task requires both listening comprehension, that is the input from the teacher, and oral language production, that is the output from the student.

In addition to the specifications for task development, in operational settings, specifications would also be produced for test assemblers. The test assembly guidelines would indicate the types and number of items needed to test specific skills and would include as well the ordering and formatting information for actually producing forms of the test. In our framework document, we will address sampling issues based on evidence from our research.

Since assessments provide different kinds of educationally useful information, be that screening, formative, or summative (Gottlieb, 2001), different types of ALP assessments will be necessary to fulfill all the tasks to which assessments are generally put. The framework document will therefore be designed to allow test developers to identify the specific purposes of the assessments they want to create (e.g., a screening measure for initial placement, monitoring individual student progress, creating diagnostic tools to inform instruction). Thereafter, test developers can move to create the types of tasks that are appropriate for the specific purposes of a given assessment.

4.2 Curricular Development

We offer here a brief and inexhaustive review of the few curricular initiatives that have explicitly addressed ALP. One relatively well-known curriculum is the Cognitive Academic Language Learning Approach (CALLA) devised by Chamot and O'Malley (1994). This comprehensive approach to integrating academic language throughout the different content areas provides teachers with a handbook of instructional strategies for explicitly teaching the nonspecialized academic language encountered across content areas, as well as a number of self- and peer assessments for students, and diagnostic assessments for teachers to use. Project CAPE in San Antonio School District, Texas, has implemented CALLA with ELL students and is the recipient of an OBEMLA grant to now study student outcomes as a result of CALLA. In another initiative, a project aimed at high school and college level students, Academic Language: Assessment and Development of Individual Needs (ALADIN), explicitly trains students in how to process and learn from the language of college lectures and other academic presentation formats encountered in the college experience (Kuehn, 2000).

Although many methods and approaches for teaching ESL were originally designed for instructing adults, some have been adapted for use with K-12 students and intersect with general instructional pedagogy (see Echevarria & Graves, 1998, for examples of adaptation). Content-based language teaching/learning models currently used with ELLs offer teachers the opportunity to help students develop and enhance their second language skills within meaningful academic contexts. Schools typically choose from a variety of language programs (e.g., ESL pull-out classes, bilingual classes, sheltered content/structured immersion, adjunct classes) that are designed to teach language skills concurrently with academic content (e.g., Snow & Brinton, 1997).

A refreshing approach to monitoring the quality of classroom instruction comes with the Sheltered Instruction Observation Protocol (SIOP) developed by Echevarria, Vogt, and Short (2000). This approach is less a curriculum, per se, than it is a way in which teachers can self-reflect on their own language input to ELL students. Not only can this impact instruction but also teacher professional development.

Because all K-12 students must acquire academic language to some degree to be successful in the U.S. educational system, perhaps a second language approach to

teaching AL would benefit native speakers and ELLs alike. Kuehn (2000) arrived at the same conclusion with high school seniors and college students needing to acquire AL for the purposes of learning in higher education settings. Students who are proficient in English are expected to acquire academic language as they move from grade to grade without necessarily receiving direct instruction specifically in AL skills, and they do so presumably with varying degrees of success. ELLs, on the other hand, often receive direct instruction in language, though not necessarily in AL, through the specially designed programs discussed previously. An implication of thinking about AL as a second language is that mainstream teachers could benefit from professional development in proven pedagogical approaches for teaching ESL in an attempt to assure that all students learn AL. Teacher professional development in the area of AL is the strand to which we now turn.

4.3 Teacher Professional Development

The application of ALP to teacher professional development obviously overlaps with both AL assessment and curricula development. Teachers need to know about the AL construct to be in a better position to both assess and teach ELL students, and as we have speculated, perhaps all students. As Wong Fillmore and Snow (2000) point out, graduate courses that teach an understanding of language development have been sorely missed in teacher education programs to date. All teachers, not only ESL teachers, need to have a basic understanding and knowledge of linguistic features and processes. Without this background, teachers will not be able to adequately assess ELD, nor target their teaching practices to student language needs. An example of work in the area of professional development and AL includes Wellington (1994) who, while not referring to teachers of ELL students, nevertheless calls for a conscious awareness of language used by teachers in the science classroom in such a way that they can build up student understanding of words from simple names or labels to more challenging abstract concepts.

5. Concluding Remarks

The results of the work proposed in this design document should contribute to educational practices in the assessment of ELL students in multiple ways, among them the following goals (see also recommendations made by Abedi, Bailey, et al. 2000):

1. The identification of an empirically based ELL assessment validity threshold for defining the academic language proficiency of ELLs.
2. The establishment of a much-needed set of principled procedures for implementing accommodations as an outgrowth of an established validity threshold for academic language proficiency.

5.1 Identification of an ELL Validity Threshold

An important outcome of the development and implementation of an ALP assessment will be the identification and/or recommendation of a threshold level of proficiency that would indicate when ELL performance on a standardized content test would be valid from a linguistic standpoint. The existing language tests do not appear to provide adequate specificity about student language at the upper range of proficiency (e.g., Butler & Castellon-Wellington, 2000) and thus are not likely candidates for establishing such a threshold. However, the notion of identifying a threshold of language proficiency is viable with an ALP test that will provide a clear indication that the language complexity of the content assessment is not a barrier to student performance. In order to establish a validity/language proficiency threshold, we propose the development of standards for defining proficient academic language ability in ELL students. As mentioned in section 3.1, one stumbling block to both research and policy with ELLs is the lack of uniformity in how school districts and states operationally define these students through their designations, such as LEP, FEP, RFEP, and bilingual. The lack of uniformity is due in large part to the different approaches states take to making their designations. An ALP test that allows for clear, objectively defined parameters for ranges of linguistic performance would help remove this stumbling block and make articulation of ELL performance uniform. These efforts would specify academic language proficiency characteristics aligned with the type of language used on content assessments and standards documents, as well as that found in teacher expectations and classroom talk and print exposure. The threshold should be drafted based on extensive study of the academic language requirements for successful school performance and will require participation of language experts as well as policymakers.

5.2 Formation of Principled Procedures for the Implementation of Test Accommodations

While we have not focused specifically on test accommodations in this document due to the emphasis we wanted to place on ALP, there have been important research efforts directed at determining the effects of accommodation on test performance. One line of research in this area has been the simplification of the language of test items (e.g., modifying the language to enable students to more readily understand the content being tested). Research in this area has led to mixed results (e.g., Abedi, Courtney, & Leon, 2001; Rivera & Stansfield, 2001), with evidence that some ELL students benefit from a simplified form of content assessment, whereas others may not. One explanation for why there has been no unambiguous finding that simplification leads to better test results for ELL students is the criticism that language is not always easy to process if simplified (e.g., Wong Fillmore & Snow, 2000). Fragmenting extended discourse into shorter chunks can lead to difficulties when temporal and causal conjunctive adverbs (e.g., however, moreover, etc.) are removed from text.

Where we see our initiative to measure ALP having an impact on the area of accommodations with ELL students is in the creation of a set of principled procedures for implementing those accommodations. This would be an outgrowth of an established (i.e., validated) threshold for academic language proficiency. For example, if an ALP assessment determines that a student is having problems with AL at the vocabulary level (i.e., has not reached a threshold for academic vocabulary, yet to be determined), then providing a dictionary as a form of test accommodation is meaningful. However, if a student's performance on an ALP assessment reveals challenges with extended discourse features of English, then no amount of thumbing through a dictionary is going to make the testing situation equitable or the student's score valid. While this example is perhaps overly simplistic, it is illustrative of the types of guidelines for the implementation of test accommodations that can come from assessment of ALP.

References

Abedi, J., Bailey, A. L., & Butler, F. A. (2000). General discussion and recommendations, *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (J. Abedi, A. L. Bailey, & F. A. Butler, Project Directors; Final Deliverable to OERI/OBEMLA, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Courtney, M., & Leon, S. (2001). *Language accommodations for large-scale assessment in science* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Mirocha, J. (2000). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data, *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (J. Abedi, A. L. Bailey, & F. A. Butler, Project Directors; Final Deliverable to OERI/OBEMLA, Contract No. R305B960002; pp. 3-49). 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., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

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

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda.* (Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students, Board on Children, Youth and Families, Commission on Behavioral and Social Sciences and Education, National Research Council, Institute of Medicine). Washington, DC: National Academy Press.

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

Bailey, A. L. (2000a). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners, *The validity of administering large-scale content assessments to English language learners: An*

investigation from three perspectives (J. Abedi, A. L. Bailey, & F. A. Butler, Project Directors; Final Deliverable to OERI/OBEMLA, Contract No. R305B960002; pp. 85-105). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L. (2000b, Fall/Winter). Learning to read makes language learners of us all. *Center X Forum, 1*(1), 1,9. [Available from UCLA Graduate School of Education at http://www.centerx.gseis.ucla.edu/forum/]

Bailey A. L. (forthcoming). From *Lambie* to *Lambaste*: The conceptualization, operationalization and use of academic language in the assessment of ELL students. In T. Wiley & K. Rolstad (Eds.), *Academic language*. Mahwah, NJ: LEA.

Bailey, A.L., & Butler, F.A. (2002, May). *Ethical considerations in the assessment of the language and content knowledge of English language learners K-12.* Paper presented at the Language Assessment Ethics Conference, Pasadena, CA.

Bailey, A. L., Butler, F.A., Borrego, M., LaFramenta, C., & Ong, C. (2002, Spring). Towards the characterization of academic language in upper elementary classrooms. *Language Testing Update*, Vol 31, 45-52.

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

Baker, E., Linn, R. L., & Herman, J. (2002, Spring). Special issue: No Child Left Behind. *The CRESST Line*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E., Linn, R. L., Herman, J., & Koretz, D. (2002, Winter). *Standards for educational accountability systems* (Policy Brief No. 5). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Boyer, E. L. (1991). *Ready to learn: A mandate for the nation*. Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Butler, F. A., & Bailey, A. L. (2002, Spring). Equity in the assessment of English language learners K-12. *Idiom, 32*(1), 1,3. [Available from New York State TESOL at http://www.nystesol.org/]

Butler, F. A., & Castellon-Wellington, M. (2000). Students' concurrent performance on tests of English language proficiency and academic achievement, *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (J. Abedi, A. L. Bailey, & F. A. Butler, Project Directors; Final Deliverable to OERI/OBEMLA, Contract No. R305B960002; pp.

51-83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing*, Vol. 18, No. 4, 409-427.

Butler, F. A., Stevens, R., & Castellon-Wellington, M. (1999). *Academic language proficiency task development process* (Final Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., Weigle, S. C., Kahn, A. B., & Sato, E. Y. (1996). *California Department of Education adult English-as-a-second-language assessment project: Test development plan with specifications for placement instruments anchored to the model standards.* Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Butler, F. A., Weigle, S. C., & Sato, E. Y. (1993). *California Department of Education adult English-as-a-second-language assessment project* (Final Rep., Year 1). Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Cazden, C. (2001). *Classroom discourse: The language of teaching and learning (2nd ed.).* Portsmouth, NH: Heinemann.

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

Council of the Great City Schools & National Clearinghouse for English Language Acquisition & Language Instruction Educational Programs. (2002, March). *Assessment standards for English language learners* (Draft Executive Summary). Washington, DC.

Cummins, J. (1980). The construct of proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics: Current Issues in Bilingual Education,* 81-103.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters, Ltd.

Cunningham, J. W., & Moore, D. W. (1993). The contribution of understanding academic vocabulary to answering comprehension questions. *Journal of Reading Behaviors, 25*, 171-80.

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven, CT: Yale University Press.

Echevarria, J., & Graves, A. (1998). *Sheltered content instruction: Teaching English-language learners with diverse abilities.* Needham Heights, MA: Allyn and Bacon.

Echevarria, J., Vogt, M., & Short, D. J. (2000). *Making content comprehensible for English language learners: The SIOP model*. Needham Heights, MA: Allyn & Bacon.

Gibbons, P. (1998). Classroom talk and the learning of new registers in a second language. *Language and Education, 12*(2), 99-118.

Gottlieb, M. (2001). Four 'A's needed for successful standards-based assessment and accountability. *NABE News, 24 (6),* 8-12.

Hadley, P. A., Wilcox, K. A., & Rice, M. L. (1994). Talking at school: Teacher expectations in preschool and kindergarten. *Early Childhood Research Quarterly, 9,* 111-129.

Hicks, D. (1994). Individual and social meanings in the classroom: Narrative discourse as a boundary phenomenon. *Journal of Narrative and Life History, 4*(3), 215-240.

Johns, A. M. (1997). *Text, role, and context: Developing academic literacies.* Cambridge: Cambridge University Press.

Kahn, A. B., Butler, F. A., Weigle, S. C., & Sato, E. Y. (1995). *California Department of Education adult English-as-a-second-language assessment project* (Final Rep., Year 3). Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Kern, R. (2000). *Literacy and language teaching.* Oxford: Oxford University Press.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Kuehn, P. (2000). *Academic Language Assessment and Development of Individual Needs. (A.L.A.D.I.N.) Book One.* Boston, MA: Pearson Custom Publishing.

MacSwan, J., & Rolstad, K. (in press). Linguistic diversity, schooling, and social class: Rethinking our conception of language proficiency in language minority education. In C. Paulston & G. Tucker (Eds.), *Essential readings in sociolinguistics.* Oxford: Blackwell.

McKay, P. (2000). On ESL standards for school-age learners. *Language Testing, 17*(2), 185-214.

Menyuk, P. (1995). Language development and education. *Journal of Education, 177*(1), 39-62.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-103). New York: Macmillan.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.

National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. Shavelson & L. Towne (Eds.), Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nippold, M. A. (1995). Language norms in school-age children and adolescents: An introduction. *Language, Speech, and Hearing Services in Schools, 26,* 307-308.

No Child Left Behind. (2001). *No Child Left Behind.* Conference Report to Accompany H.R., 1, Rep. No. 107-334, House of Representatives, 107th Congress, 1st Session.

No Child Left Behind. (2001a). *No Child Left Behind.* Title I: Improving the academic achievement of the disadvantaged. 107th Congress, 1st Session, December 13, 2001. (Printed version prepared by the National Clearinghouse for Bilingual Education). Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.

No Child Left Behind. (2001b). *No Child Left Behind.* Title III: Language instruction for limited English proficient and immigrant students. 107th Congress, 1st Session, December 13, 2001. (Printed version prepared by the National Clearinghouse for Bilingual Education). Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.

Popham, W. J. (1978). *Criterion referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Rivera, C., & Stansfield, C. W. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Scarcella, R. (in press). *Key issues in accelerating English language development*. Berkeley, CA: University of California Press.

Schleppegrell, M. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431-459.

Schleppegrell, M. (2002, May). *Grammatical and discourse features of the target genres in California's English language development (ELD) standards.* Paper presented at the UC LMRI 2002 Annual conference, University of California, Berkley, CA.

Short, D. (1994). Expanding middle school horizons: Integrating language, culture, and social studies. *TESOL Quarterly, 28*(3), 581-608.

Snow, C. (1991). Diverse conversational contexts for the acquisition of various language skills. In J. Miller (Ed.), *Research on child language disorders* (pp. 105-124). Austin, TX: Pro-Ed.

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

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

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

Turner, J. (1997). Creating content-based language tests: Guidelines for teachers. In M. A. Snow, & D. M. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language and content.* (pp. 187-200). White Plains, NY: Longman.

Weigle, S. C., Kahn, A. B., Butler, F. A., & Sato, E. Y. (1994). *California Department of Education adult English-as-a-second-language assessment project* (Final Rep., Year 2). Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Wellington, J. (1994). Language in science education. In J. Wellington (Ed.), *Secondary science* (pp. 168-188). Great Britain: Mackays of Chatham PLC.

Wong Fillmore, L., & Snow, C. (2000). *What teachers need to know about language.* ERIC Clearinghouse on Languages and Linguistics. Retrieved August 1, 2000, from http://www.cal.org/ericcll

Academic vocabulary is distinguished from the "everyday" vocabulary that is used to communicate on a less formal level outside of the classroom.20 For example, the non-specialized academic vocabulary words examine and cause contrast with the everyday vocabulary words look at and make.Â Bailey, A.L. and F.A. Butler, An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document. 2003, CRESST/University of California, Los Angeles: Los Angeles. Bertelson, P., et al., Metaphonological abilities of adult illiterates: New evidence of heterogeneity. An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document.Â This document provides an approach for the development of an evidentiary framework for operationalizing academic language proficiency (ALP) for broad K-12 educational applications in these three key areas: curriculum development, instruction, and teacher professional development. Expand. 112. An evidentiary framework for operationalizing academic language for broad application to Kâ€“12 education: A design document (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Bennett, R. (2011). Formative assessment: A critical review. Assessment in Education 18(1), 5â€“ 25. Black, P., & Wiliam, D. (1998).Â Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. Working Papers on Bilingualism, 19, 121â€“9. Cummins, J. (1981). An evidentiary framework for operationalizing academic language for broad application to Kâ€“12 education: A design document. Los Angeles: University of California.Google Scholar. Bailey, A. L., & Huang, B. H. (2011).Â Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. Marketing Letters, 9, 21â€“35. doi:10.1023/A:1007911903032 CrossRef Google Scholar. Berendes, K., Dragon, N., Weinert, S., Heppt, B., & Stanat, P. (2013). HÃ¼rde Bildungssprache? Eine AnnÃ¤herung an das Konzept â€œBildungssprachеâ€ unter Einbezug aktueller empirischer Forschungsergebnisse [Academic language as an obstacle? Request PDF | An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document. CSE Report | With the No Child Left Behind Act (2001), all states are required to assess English language development (ELD) of English language learners (ELLs) |Â An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document. CSE Report. January 2003.Â My rationale in designing the CDF-construct has been that analysis of classroom discourse (together with textbook analysis) will yield more consistent and higher quality insights about educational practice than curriculum standards and expectations alone (as argued also by Bailey & Butler 2003).