Ensuring Accuracy in Testing for English Language Learners

Council of Chief State School Officers • SCASS — LEP Consortium

Rebecca Kopriva

Acknowledgments

This *Guide* is the product of the Council's State Collaborative on Assessment and Student Standards (SCASS) LEP consortium in collaboration with the Project to Improve Achievement in High Poverty Schools of the Resource Center on Educational Equity. The Consortium activities are coordinated by Julia Lara and John Olson. Support for this project comes from the member states and the United States Department of Education.

This document was developed under the direction of Julia Lara, LEP SCASS Coordinator. Although the text of this *Guide* was written by Project Consultant Rebecca Kopriva, many people provided valuable suggestions regarding the format and contents of this report. We want to especially extend our gratitude to the following people: Doris Redfield, Independent Consultant; Valena Plisko, Sharon Saez, Barbara Coentry, Elois Scott, Edina Kole, Rebecca Fitch, Jeanette Lim, David Berkowitz, all at the United States Department of Education; Maria Medina Seidner, Texas Education Agency; Mary Ramirez, Philadelphia Public Schools; Sonia Hernandez, California State Department of Education; John Olson, CCSSO; and Wayne Martin, CCSSO.

Finally, the ease with which this document can be read is solely attributed to the skillful editing work performed by Anne Turnbaugh Lockwood.

The preparation of this document was partly financed by funds provided by the U.S. Department of Education (U.S. Department of Education, Grant Number R215U980002-98). No official endorsement by the U.S. Department of Education or the Council of Chief State School Officers is intended or should be inferred. The views and opinions expressed in this report are not necessarily those of the Council of Chief State School Officers, the SCASS LEP project staff, or the states participating in the SCASS LEP Consortium. This *Guide* is not intended to and does not add to or subtract from any otherwise applicable requirements contained in federal or state law. The *Guide* is intended to provide suggestions that school systems may find helpful to address the issue of inclusion of LEP students in large scale assessments.

ISBN 1-884037-63-1

Contents

Council of Chief State School Officers

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. The Council of Chief State School Officers seeks its members' consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public. Through its structure of standing and special committees, the Council responds to a broad range of concerns about education and provides leadership on major education issues.

Because the Council represents the chief education administrators, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. The Council of Chief State School Officers forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
202-408-5505
Fax: 202-408-8072
www.ccsso.org

The Council's Assessment Center

The Council's State Education Assessment Center was established to provide an information base on education in the United States, especially from a state perspective. The Center works to improve the breadth, quality, and comparability of education data, including state-by-state achievement data; descriptive data; indicators of quality in areas such as math and science; and performance assessment of students, teachers, and education leaders.

In collaboration with state education agencies, the federal government, and national and international organizations, the Center contributes to a set of useful and valid measures of education geared, when appropriate, to education standards. The Center also supports efforts by states to use standards and assessments to improve instruction through collaborative activities of states and others within the education field.

The Council's Resource Center

The Council's Resource Center on Educational Equity was established by chief state school officers to provide services designed to ensure equitable, high-quality, and developmentally appropriate education for all students, especially minorities, females, students with disabilities, limited English proficient students, and low-income students. The Resource Center conducts research and policy formulation, develops reports and other materials, operates grant and other action programs, provides capacity-building technical assistance to state education agencies, holds working conferences, and monitors federal and state civil rights and education programs that focus on disadvantaged students.

State Collaborative on Assessment and Student Standards (SCASS)

The State Collaborative on Assessment and Student Standards (SCASS) was created in October 1991 to encourage and assist states in working collaboratively on assessment design and development in a variety of subject areas. The State Education Assessment Center of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. A total of 45 states and two extra-state jurisdictions participated in ten projects during this project year of July 1998-June 1999.

States gain most from the primary products of the SCASS projects. All products are determined and designed by the participating states based on their particular needs. Some SCASS projects pool their resources to conduct cutting-edge research while others commission papers or write reports to define clarify or interpret assessment-related issues. Some projects develop guides to help educators understand and use assessments or build training programs for creating and using portfolios. Three projects currently are creating CD-ROMs to package large quantities of assessment items and materials in order to provide them to state personnel, teachers, and other educators in a user-friendly way.

Because these projects are collaborations among collections of states, costs for developing assessment items and implementation materials are amortized over the number of states involved. Other benefits of SCASS projects are professional development opportunities for participating teachers from member states and collaborating with the US Department of Education on several assessment-related collaboration tasks of interest to the Department.

The SCASS Assessing Limited English Proficient Students Consortium

The Assessing Limited English Proficient (LEP) Students Consortium develops procedures and materials to assist states towards more appropriate assessment of English Language Learner (ELL) students, including research on effective programs for ELL students, language proficiency measures, and other materials related to measuring academic achievement. The project is administered jointly by staff from the Council's Resource Center on Educational Equity and the State Education Assessment Center.

The members, participants, and consultants of the SCASS LEP group are engaged in the development of various products designed to support standards-based assessment for LEP students. Several products have been completed to date, which include: *A Guide to Scoring LEP Student Responses to Open-Ended Mathematics Items* (1998) and *A Guide to Scoring LEP Student Responses to Open-Ended Science Items* (1999). These guides were designed to be adapted in response to various large-scale assessment needs. They contain linguistic training guidelines, examples of students' work, a discussion of issues related to the accurate development of assessments appropriately geared toward English language learners, and a glossary of terms. These publications are tools for training scorers of LEP students' responses to open-ended mathematics and science items. Both can be used as training tools to increase the accuracy with which scorers evaluate work completed by LEP students.

A Conceptual Framework for the Valid and Comparable Measurement of All Students is a recent paper that presents a conceptual framework and research plan for test standardization. This document is the outcome of a series of meetings convened by CCSSO during the Winter of 1998 with technical experts, educational researchers, inclusion advocates, and curriculum experts on this topic. The purpose of the meetings was to initiate a discussion targeted to changing the theoretical framework used to design and construct assessments. Participants examined the discrepancies between a "one size fits all" approach and how children learn, process, reason and respond in terms of the definition of standardization. The recommendations generated at this meeting and synthesized in the report will serve as the basis for further research. This document will be listed on the CCSSO Web Site (*www.ccsso.org).*

Limited English Proficient (LEP) SCASS project staff and consultants are developing additional products designed to help state and district assessment developers design inclusive assessments. For example, the LEP Performance Sampler will provide a means for state assessment specialists to review samples of work produced by LEP students at three performance levels (advanced, proficient, and basic) in mathematics and science, and at one grade level (fourth).

Executive Summary

Ensuring Accuracy in Testing for LEP Students: A Practical Guide for Assessment Development has one broad purpose: the improvement of large-scale academic assessments for students commonly referred to as Limited English Proficient (LEP) or English Language Learners (ELLs). Throughout this *Guide*, we refer to students as both LEP and the more current term, ELL. This *Guide* was written because large-scale assessments are studded with problems that affect not only how equitably the achievement of LEP students can be measured, but also how effectively LEP students' mastery of content is assessed. Limited English Proficient students may understand and know much more than they are allowed to demonstrate under the confines of large-scale tests designed for mainstream use with students sharing common cultural experiences.

This *Guide* is designed to provide practical, research-based information for individuals and groups developing and administering tests of academic achievement for LEP students, including:

- Test publishers
- State and local educational assessment unit (SEA and LEA) staff
- Personnel involved in the academic evaluation of diverse school populations
- Curriculum specialists, both state and district
- Local educators at both the district and building levels
- Experts on pedagogical approaches to schooling
- Experts on the teaching and learning of special needs students, diverse populations, and populations from different types of cultural and geographic centers
- Parents and a variety of public and community stakeholders.

This Guide is designed to provide practical, research-based information for individuals and groups developing and administering tests of academic achievement for LEP students.

It also is intended to provide a lucid, easily accessible knowledge base for policymakers that clarifies some of the complex issues that they face — as well as recommendations for action — as they consider the needs of LEP students. As policy is formed at the national, state, and local levels, it must be informed on a continuous basis by solid evidence of the effectiveness of school programs, instruction, and curriculum. This *Guide* intends to contribute to that knowledge base.

Many educators and policymakers are well-informed about a myriad of substantive educational issues, but are not informed sufficiently about the complexities that result when well-intentioned individuals measure LEP students' academic achievement with instruments designed for the mainstream student population. The testing process is one issue; the ways in which the results are used is quite another. When high-stakes decisions are made using test results that may be flawed, students' educational futures hang in the balance.

Even well-informed educators may not realize fully the effects of culture upon the ways in which LEP students problem-solve and interpret test items. Some accommodations may be well-intentioned, but wrongheaded; another combination of accommodations may be advised in certain situations and not the wisest choice in others. This *Guide* considers the issue of accommodations, but emphasizes that the actual development and construction of a test, if properly conceived, can promote the accurate assessment of LEP students' achievement in U.S. schools.

The *Guide* begins with a discussion of several factors that affect LEP students' access to large-scale tests. In this discussion, the author addresses problems and offers solutions over and above the content issues that have been the focus of the bias literature to date. This viewpoint can help bridge the gap between the testing community and the language minority community, meet some of the needs and interests of both communities, and substantively increase the accuracy of information collected about the achievement of LEP students.

Structure of the *Guide*

To expand the knowledge base of educators working with LEP students — who may or may not have substantive background with assessment as it pertains to LEP students — the *Guide* includes considerable information about the testing process itself. Assessment personnel, however, also can benefit from the information in this *Guide* that specifically explains how issues related to LEP status influence the equitable evaluation of this population. Many of the points included in this *Guide* also may prove useful for evaluating a broader group of language minority students in the U.S. — those who speak a language other than English at home and who may or may not be classified as English Language Learners (or LEP).

Existing measures should be supplemented to avoid test bias in large-scale assessments.

As a practical consideration, the *Guide* identifies many significant points throughout the test construction and implementation phases where appropriate intervention should occur so that better and more equitable assessment tools can be developed. It also includes many practical recommendations about the "hows" and "whys" of strategic intervention in the test construction/implementation process. While some of the issues inherent in testing students over an entire state or large district differ from the problems of those who develop assessments for a school or program-wide evaluation, many of the points here should be applicable to a broad range of assessment situations. However, note one important limitation — although some states currently are attempting the standardization of portfolio procedures to achieve comparability in scores across classrooms, portfolio assessments per se are not included in this discussion.

The Psychometric Context of the *Guide*

The underlying philosophy that permeates this *Guide* is succinct: Existing measures should be supplemented to avoid test bias in large-scale assessments. Many of these measures have been instituted by the psychometric community over the past 20 years. These measures include bias or sensitivity reviews as well as the development and use of differential item functioning (DIF) statistical procedures. A detailed discussion of the technical merit of assessments for LEP students comprises a large portion of this *Guide,* because many invalidity problems are pervasive; they focus not on content, but on how mastery is evaluated.

Another key belief that permeates this *Guide* is that LEP students should be included equitably whenever possible in large-scale assessment. Too frequently, well-intentioned educators believe they are protecting the interests of their LEP students by shielding them from mainstream, large-scale assessment. Instead, they hold them to a lower standard and do not provide the wherewithal for them to participate fully in the type of education that should be every student's right in a democratic society.

Rather than discouraging LEP students from participating in large-scale assessment, this *Guide* urges an increase in the percentage of ELLs who can participate with some sense of authenticity in the testing process. As these numbers increase, it will become more possible to determine the effectiveness of programs which include LEP students — and thus, truly improve the quality of education for all students attending the nation's schools.

Introduction

Purpose of the *Guide*

The impetus for the development of this *Guide* arose from state and federal requirements regarding inclusion of all students in standards-based reform efforts. Federal-level requirement under the 1994 reauthorized Title I of the Elementary and Secondary Education Act of 1965, and Title VI of the Civil Rights Act of 1964, addressed the issue of LEP student inclusion in statewide assessments. Provisions under Title I of ESEA specified that LEP students were to be included and assessed "to the the extent practicable in the language and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students' mastery of skills in subjects other than English" [§ 1111(b)(3); 34 C.F.R. 200.4(b)(7)].

In addition, to the extent statewide or district assessment may result in an educational benefit, the exclusion of LEP students from participation in these assessments or failure to provide LEP students with accommodations may raise Title VI issues. Also under Title VI, if LEP students are excluded from a particular statewide or district assessment based on educational or psychometric justification, the districts have an obligation to collect comparable information about these students' academic progress.

This *Guide* has several goals, but is guided by a central, overarching focus: the improvement of large-scale academic assessments for students whose first language is not English and who are learning English in schools throughout the United States. Through practical, research-based information, we intend to increase the knowledge base of individuals and organizations who develop and use large-scale academic assessments with Limited English Proficient (LEP) students, otherwise known as English Language Learners (ELLs).

Specifically, this *Guide* is targeted to:

- Test publishers
- State and local educational assessment unit (SEAs and LEAs) staff
- Personnel involved in the academic evaluation of diverse school populations
- Curriculum specialists, both state and district
- Local educators at both the district and building levels
- Experts on pedagogical approaches to schooling
- Experts on the teaching and learning of special needs students, diverse populations, and populations from different types of cultural and geographic centers
- Parents, policymakers, and a variety of public and community stakeholders.

We designed this *Guide* to advance the equitable evaluation of LEP students' academic achievement using instruments designed for the mainstream student population. We begin the guide by highlighting several factors that affect accurate access to tests for many minorities, including limited English proficient students. In this discussion, we address problems and offer solutions over and above the content issues that have been the focus of the bias literature to date. We believe this viewpoint can help bridge the gap between the testing community and the language minority community, meet some of the needs and interests of both communities, and substantially increase the accuracy of information we collect about the achievement of LEP students.

Our approach throughout this *Guide* is systemic and multidimensional. We identify many significant points throughout the test construction and implementation phases where appropriate intervention should occur so that a better and more equitable assessment tool can be developed. We also offer many practical suggestions about the "hows" and "whys" of strategic intervention in the test construction/implementation process. We realize that some of the issues inherent in testing students over an entire state or large district differ from the problems of those who develop assessments for a school- or program-wide evaluation, but many of the points here should be applicable to a broad range of assessment situations. One important limitation: Although some states currently are attempting the standardization of portfolio procedures to achieve comparability in scores over classrooms, portfolio assessments per se are not included in this discussion.

In our discussion throughout the remainder of the *Guide*, we focus on assessments intended to evaluate LEP students' level of mastery in specific content areas. These assessments should not be confused with tests which evaluate basic skill proficiency in English literacy. While there certainly is an overlap between literacy skills evaluated on an English language proficiency test and tests which assess a student's knowledge in English language arts, the language arts content standards of all states define sets of knowledge and skills that are significantly broader than domains defined by English proficiency tests.

Our primary intent throughout this *Guide* is to supplement existing measures to avoid test bias in large-scale assessments.

The Psychometric Context of the *Guide*

Our primary intent throughout this *Guide* is to supplement existing measures to avoid test bias in large-scale assessments. These measures have been instituted by the psychometric community over the past 20 years and include bias or sensitivity reviews as well as the development and use of differential item functioning (DIF) statistical procedures. Traditional bias or sensitivity review procedures focus on items or passages that contain offensive or stereotypical language or contexts. Differential item functioning procedures flag items where performances vary by group. While reviews and the use of DIF procedures are important, we believe that these measures on their own cannot detect many major sources of item and test problems that contribute to systematic sources of invalidity for LEP students. Many of these invalidity problems are pervasive; they focus not on content, but on how we evaluate mastery.

We address these problems in this *Guide* both in the discussion that follows and the set of recommendations that concludes the discussion. However, we do not suggest that once tests are improved, all differences between groups will disappear. Unfortunately, differences in access to educational opportunities are just as real as individual differences in achievement. When LEP students are included in assessments, reliable results can assist educators in identifying when LEP students need additional resources and/or programs to meet challenging standards.

Increasing Testing's Technical Rigor and Access for LEP Students

In Nancy Cole's (1993) introductory address to an Educational Testing Service conference on differential item functioning, she underscored the magnitude of differences in approach and language between two well-meaning communities concerned about accurate testing for minority students: the technical psychometric community and the non-measurement stakeholders, including the language minority educational community, educators, and the general public. These differences have made it difficult to develop assessments for diverse populations and can impede the current progress of testing equity.

These two communities, however, share a common goal — to increase the technical rigor, especially validity, of tests for English language learners, while still being able to collect comparable and generalizable information across population subgroups. To that end, we provide recommendations specific to upgrading and evaluating score accuracy for LEP students within the performance standards, item and form development, administration, response, scoring, data analyses, reporting, and use.

The Challenge: Reform and Equitable Testing

The national, statewide, and local education policy agendas articulated over the last several years are unanimous: All students must be held to high academic standards. Several federal, state, and local mandates emphasize that all students, including LEP students, must be included in large-scale, mainstream assessments used at the state or local levels that are designed to evaluate student progress as measured against state or local standards. However, many initiatives historically have excluded large numbers of LEP students from large-scale exams given to most students (Lara and August, 1996). As a result, these students are excluded from their accountability systems for at least two or three years — often much longer. Furthermore, researchers and policymakers agree that the exclusion of LEP students from regular testing has an unintended but important effect — these students also are excluded from many of the educational benefits of the standards movement (August and Hakuta, 1997). To the extent that large-scale assessments may result in an educational benefit, the exclusion of LEP students from participation in large-scale assessments would raise concerns under Title VI of the Civil Rights Act of 1964, which prohibits states and school districts that receive federal financial assistance from discriminating in the operations on the basis of race, color or national origin.[1]

To reap the benefits of these educational reforms, LEP students must be included in mainstream, large-scale academic testing, and must be provided with the optimum quality of teaching and other educational resources to meet challenging standards. Both the testing community and the language minority community need to assess

[1] The U.S Department of Education's Office for Civil Rights (OCR) enforced Title VI. OCR investigates complaints fielded by individuals or their representatives who believe that they have been discriminated against because of their race, color, or national origin. OCR is available to provide technical assistance regarding the application of Title VI requirements to statewide and district wide testing of LEP students. On December 8, 1999, OCR issued a draft document entitled "The Use of Tests When Making High-Stakes Decisions for Students: A Resource guide for Educator and Policymakers." This draft document is designed to provide policymakers and educators with a useful, practical tool that will assist in planning and implementation of policies relating to the use of tests as conditions of conferring educational opportunities to students. OCR expects to issue the final document in the early Summer of 2000.

LEP students in ways that document reliably and equitably what these students know and can do.

Typically, states and districts have a well intentioned reaction to the task of including more LEP students in large-scale assessments — they provide competent conceptual translations of paper-and-pencil tests in use throughout the school system. But in many cases this is problematic. Translations from English to the student's first language (also known as the home language), usually assume a degree of literacy in this language which may be absent (DeAvila, 1997). Translations also pose many psychometric and linguistic difficulties. Many experts (see August and Hakuta, 1997) agree that it is advisable to test in the language of instruction — English in most situations throughout the United States. However, there are instances in which LEP students are literate in the home language but may not receive instruction in that language. In this circumstance, students should have the opportunity to be assessed in that language if it will provide the most accurate and reliable assessment of what they know. Clearly, it is a challenge to capture accurately the extent of student knowledge when:

- Some of that knowledge may have been learned in the student's country of origin and some in U.S. schools
- The student's knowledge of English is developing and emergent
- The student may or may not have the capacity to read or accurately demonstrate his/her mastery through written texts in either English or the home language.

Generalizations about any one of these subgroups may mask important background characteristics that can serve as valuable information for designing curricular interventions for these students.

Demographics of the LEP Population

According to the most recent SEA reports, the number of LEP students in public schools has increased in almost all states. As of the 1996-77 school year, the total number of LEP students in the U.S. was approximately 3.4 million. This figure represents a 7 percent increase since 1995-96.[2] Many experts have noted that this total number is an undercount of the actual LEP population since many school districts do not report their LEP student enrollments and the operational definition of LEP varies across states and districts.

With the exception of Florida, the greatest percent change in K-12 enrollment since 1995 has occurred in southern and midwestern states. For example, Alabama experienced a 22 percent increase; Arkansas, 20 percent; Tennessee, 39 percent; North Carolina, 32 percent; and South Carolina, 36 percent. In the Midwest, Kansas reported a 26 percent increase in enrollment; Nebraska, 28 percent; and South Dakota, 39 percent. Therefore, smaller school districts in these states are heavily affected by the enrollment increases. Nonetheless, 79 percent of all LEP students are concentrated in seven states, with California containing the largest proportion (41

[2] National Clearinghouse for Bilingual Education, 1998. *Summary Report of the Survey of the States' Limited English Proficient Students and Available Educational Programs and Services,* 1996-77.

[3] The other states are Arizona, New Mexico, Florida, Illinois, New York and Texas.

[4] It is important to note that a proportion of Native American students are also LEP. American Indian and Alaska native students in both BIA/tribal and high Indian enrollment public schools were more likely than Indian students in public schools with low Indian enrollment to speak a language other than English in their homes (one percent) or be limited English proficient (two percent).

[5] Berman P. et al. (1995) *School Reform & Student Diversity: Case Studies of Exemplary Practices for LEP Students.* Washington, D.C.: National Clearinghouse on Bilingual Education.

[6] National Research Council. (1998). Educating Language Minority Students. August, D. and Hakuta,K. (eds.). Washington, D.C.: National Academic Press.

percent) of the total LEP enrollment.[3] Not only are LEP students concentrated in certain states but they are enrolled primarily in large urban school districts and their surrounding metropolitan area. Thus, schools in the metropolitan areas and their urban core are the most impacted by the challenges presented by these students.

Meeting the instructional needs of LEP students successfully is a daunting challenge because this population is linguistically, culturally, and racially diverse. Nationwide, the vast majority speaks Spanish (75 percent), with Asian language speakers comprising approximately 13 percent, and the remaining languages 12 percent.[4] As a group, LEP students tend to be native-born and recent immigrants (40 percent)[5]; speak over 100 languages; are likely to be poor; and are primarily enrolled (53 percent) in grades K-4.[6]

However, there is much variability across these major ethnic/linguistic categories by country of origin, cultural background, socioeconomic status, and level of education prior to immigrating to the United States. For example, the National Research Council's panel report, *Improving Schooling for Language Minority Students* (August and Hakuta, 1997), indicated that 35 percent of families that spoke Asian/Pacific Island languages had incomes below $20,000 compared to 57 percent for Spanish speakers. Differences also exist within groups. The first wave of Southeast Asian refugees was comprised of highly educated people who, in spite of lack of proficiency in English, could support actively their children's learning at home and ease the transition to the new environment. Subsequent immigrants were less educated and perhaps less equipped to advocate for their children's education.

Some LEP children come from countries with an official language and a variety of regional dialects. For example, Chinese immigrants, if literate, are proficient in the Chinese script no matter what part of China they immigrate from. However, immigrants from mainland China are likely to be Mandarin speakers; those from Hong Kong speak mostly Cantonese; and those from Taiwan speak Fujianese. Moreover, there are two versions of the Chinese script — simplified and traditional. Traditional script is mostly used in Hong Kong and Taiwan while the simplified version is used in mainland China. Depending on where students come from, they may use characters of people originating in Taiwan and Hong Kong, or the simplified characters used by those educated in Mainland China. Clearly, in trying to provide instructional supports to these students, teachers need to understand these background characteristics. Thus, generalizations about any one of these sub-groups may mask important background characteristics that can serve as valuable information for designing curricular interventions for these students.

Approximately 80 percent of LEP students attending public schools are enrolled in programs designed to meet their needs. Some schools use the home language to teach content as the child learns English (bilingual education); others focus exclusively on teaching English and phase in content instruction as English language development increases (ESL); while others combine these two approaches by providing some level of native language support while the child acquires English. At the middle and high school levels, sheltered content instruction is often used to mediate students' understanding of the content areas. Much has been written about the degree to which these approaches are effective in raising the academic achievement of LEP students and teaching English. This report will not focus on this work. However, it is important to note that no matter what language program alternative is selected, students are expected to learn English and meet the standards set forth in the content areas.

Getting an accurate and reliable picture of LEP student academic performance at the national or state level is problematic because the measures currently used to assess LEP student performance in the content areas are often inadequate. Until better measures are developed, policymakers must use multiple sources of information about LEP students' performance in order to assess their educational status. Nonetheless, several reports indicate that LEP student achievement is below expected standards of performance. A longitudinal study of students receiving Title I services (Prospects, 1995) reported that LEP third grade students did not perform as well in reading and math as other third graders. Other indicators of student academic status, such as in-class grades, support the test data findings. Additionally, a recent report of LEP student performance from the 1995 NAEP for Reading Assessment (1999) show that average composite scores on the fourth grade reading assessment were lower for both males and female LEP students than for their non-LEP peers. However, there is an exception in the pattern of performance indicated above. Several researchers have conducted within-group analyses of LEP students and have found that first generation LEP students outperform their native born peers on several tests of academic achievement. These findings reinforce the previous assertion regarding the variability across the limited English proficient student population (Portes, 1996).

Summary

To date, there are few mechanisms in large-scale testing processes that ensure that assessments will yield accurate and defensible information about LEP student achievement. Emerging research focuses on the linguistic structures of items; on some administration accommodations for LEP students, students with disabilities, and poor readers; and also on increasing accuracy in scoring for special needs students. However, while some essential procedures — such as the establishment of bias review panels and the use of DIF statistics — have been adopted in recent years by test developers to reduce bias, widespread misunderstanding continues to persist about what exactly mainstream tests actually measure relevant to the achievement of LEP students.

Chapter 1

Issues of Alignment, Inclusion, and Participation

Content standards are written primarily at the state and/or district levels to define priorities and expectations for students' knowledge and performance. Most evaluation of student progress that relates to content standards is ongoing throughout the teaching and learning cycle. Many states and/or districts also require large-scale testing across classrooms as an independent evaluation of student achievement and programmatic effectiveness. This large-scale testing typically is handled through the administration of a single test or through administration of a more complex system of tests and other data collected for every student or for a sample of students. In either case, the assessment(s) are selected from a test publisher or developed, sometimes from ground zero and sometimes using an established test as a core with new items created for the state or district: If the assessment(s) are to be used to measure student progress toward the state's and/or district's content standards in an effective manner, they must be aligned suitably to the content standards.

Many states also require large-scale testing across classrooms as an independent evaluation of student achievement and programmatic effectiveness.

What Is Alignment?

The meaning of alignment is frequently misunderstood. Alignment suggests that the results of the test(s) can be used with confidence to evaluate whether programs effectively deliver necessary services, whether teachers practice their craft to their maximum ability, and whether students learn the content specified as important in the state or district's content standards — for all students involved in the programs. The results must be reliable over time and for all students. They also must be valid, accurately measuring what they say they will measure for each student. Since the test items are samples of everything important in the standards, the results also must be a valid and reliable measure of this content. In other words, the results must be generalizable to the entire content standards domain for all students taking the test.

In the *Handbook for the Development of Performance Standards,* Hansche (1998) clarifies what it means to align content standards to pre-built tests (also known as off-the-shelf tests), or to newly built custom-built tests, so that results can be generalized to the content standards domain. The same advice is relevant for achieving generalizable results for custom-built tests. She writes — a common misconception is that when pre-built or off-the-shelf assessments are matched to content standards and curriculum, most "aligners" will get a good match. Many of the test items/tasks, usually around 80-90 percent or even 100 percent, match what is specified in their content standards. What doesn't happen, is the *rest of the match,* which, it turns out, is critical. An important question to ask is how much of the *content standards and curriculum* is assessed using a pre-built system? That match may reflect only 50 or 60 percent of content standards and curriculum.

- What about the parts of the content standards/curriculum that are not assessed at all with the pre-built system or even a custom-built system?
- Is the relative emphasis of content covered by the pre-built test the same as the intended emphasis in content standards?
- Is the difficulty level of the test consistent with the desired performance standards?

> Real alignment means that the assessments match the depth (difficulty levels) and breadth (content and weighting) of the content one-to-one (italics original, p. 22-23).

The Problem: Alignment for All Students

When we work to align content standards with assessments, we need to keep an important question paramount: Do the tests and any other evaluation methods which are being developed and/or selected measure the same content (depth, breadth, and relative emphasis) for our diverse learners as effectively as they measure the same content for our other students? When we refer to diverse learners, we include:

- Students who are English language learners
- Students with disabilities (emotional, physical, cognitive or behavioral)
- Students with learning, processing, and/or response styles which differ from the mainstream.

Whenever separate evaluation indices are used for some students, significant difficulties arise.

Unfortunately, most of the activity to date related to aligning assessments with content standards to ensure equity has been rhetorical and not reflected in action.

Currently, most large-scale testing across content areas favors students who:

- Read well and easily
- Express themselves well in written form
- Do not rely heavily on kinesthetic, tactile, spatial, and/or auditory information to solve problems
- Perform well for sustained periods of time, and
- Adjust easily and fit into a multitude of schedule and format decisions born from the various stakeholders' competing priority structures.

Most "special needs students" — identified as such or not — face substantial challenges in some of these areas. They may know just as much as their test-favored peers but are unable to demonstrate their mastery due to the nature of large-scale tests.

When Should LEP Students Be Included in Mainstream Assessments?

The agency that is responsible for buying or developing the assessment and subsequently administering it decides when to include LEP students in that assessment. This decision typically is made by a state or local school district. As a general rule, state and district personnel need to remember that all students in academic programs should be included in mainstream assessments if academic standards are intended to apply to all students.

Whenever separate evaluation indices are used for some students, significant difficulties arise. Are the two indices really measuring the same constructs? How are the scores compiled so appropriate program evaluation can include information from all students? All too often, information about students waived out of large-scale testing is not included alongside information about those who take the mainstream assessments. This procedure becomes very problematic when information from large-scale tests is used to evaluate programs which should be effective for all students in the charge of schools, including students who have no voice, such as those who have been waived.

For the most part, current large-scale tests have not included a wide range of LEP students in their validity and norming samples. However, it is still important to include LEP students in large-scale assessments, with accommodations as necessary, as more inclusive assessments are built. Until validity data are available which specify what the tests are measuring for LEP students (with and without accommodations), the test results should be used tentatively and triangulated whenever possible with other information.

Furthermore, agencies should expect test publishers to collect the validation evidence for the full range of limited English proficient students, with tests administered to them under conditions which yield accurate information. This includes disaggregating validity data by level of English proficiency or other salient indicators of LEP status, when accommodations are not required. It also includes collecting validity data on accommodated conditions. Additionally, publishers should begin to include the range of LEP students in a development of their norms and their criterion data.

Participation of LEP Experts in the Development of Assessments

Experts with substantive knowledge of LEP students need to be involved in the development of assessments intended for these students. These experts also should have experience understanding how to adapt teaching environments to ensure optimum learning for LEP students.

In recent years, test publishers have drawn upon the knowledge of practicing teachers and other educators at many points throughout the development of their tests. Agencies, such as state or district departments of education, also have included local educators and other educational stakeholders in the decisionmaking, review, and development of their testing systems. Unfortunately, experts who work with special needs populations, including experts in schooling LEP students, have not been included in sufficient numbers in the test development and review process. For the most part, their participation is requested during bias reviews when they are charged with the task of detecting offensive material by quickly reviewing large amounts of items and related material.

Experts on LEP issues who bring the most to the development process have a deep understanding of the content standards of their state or district and understand the cultures and strengths of limited English proficient students. They could include:

- Educators working with migrant students
- Educators working with students who are newly arrived to the U.S.
- Educators from classrooms where students are learning English as well as content
- Educators from classrooms where LEP students are placed after they have mastered a certain level of English

- Educators working in L1 (first language) or bilingual (first language and English) classrooms.

It is also important to stratify educators by geographical and urban/rural differences.

We recommend that experts on issues related to the education of ELLs be involved at each step in the development or review cycle. This means that their active participation should be solicited for:

- Developing test specifications
- Reviewing and writing items and rubrics
- Trying out items in their classes
- Evaluating forms for coverage and accessibility
- Making decisions on all testing materials based on data from pilots and field tests
- Scoring, reporting, and making decisions about test use.

Evidence must document that the test is accurately measuring the same content standards for LEP students as it is for other students.

Recommendations for Alignment, Inclusion, and Participation

Alignment

1. Evidence must document that the test measures acccurately the same content standards for LEP students as it does for other students. Examples of the types of documentation will be discussed below; however, alignment of content standards to tests should not be automatically assumed for LEP students, even though it has been documented for other students.

Inclusion

2. LEP students should be included in large-scale assessments with accommodations as necessary.

3. Because current large-scale tests have not been validated for a wide range of LEP students, their results should be used tentatively and cross-validated with other information about the students' mastery.

4. Educational institutions should expect publishers to collect validity information for LEP students. This includes disaggregating validity data when accommodations are not required, and collecting validity data on accommodated conditions.

5. Publishers should begin to include the range of LEP students in the development of their norms and their criterion data.

Participation

6. Meaningful participation of a range of educators of LEP students should be expected throughout development of new assessments, or reviews of existing assessments.

Chapter 2

Alignment of Performance Standards to Content Standards

Content standards typically specify a framework for what schools should teach. They identify the curricular priorities of the state or district. In order to evaluate progress toward the specified content standards, the state or district defines what levels of mastery are adequate. States and districts also distinguish levels of performance that approach the "adequate" level as well as one or more levels that surpass it. Performance standards must be aligned properly to the content standards and to the assessments.

Are Performance Standards Equitable for LEP Students?

When considering performance standards in the context of LEP students' achievement, two related questions should be kept in mind: Do the performance standards allow a variety of performances (e.g., oral explanations, live performances, and/or the use of pictures and charts), and do they allow for performances in languages other than English? Do these non-text and non-English performances receive the proper value as suitable communication tools?

Allowing a Variety of Performances

The 1994 NAEP mathematics performance standards at grade four, in Table 1, are an illustration of very common assessment practices. Each standard at each level of performance (basic, proficient, advanced) is written in two paragraphs. The first paragraph identifies sophistication of process and understanding related to mathematics; the second outlines specific evidence required in order to determine the level of sophistication.

The first paragraph explains the level of performance, e.g.:

- *Basic:* [Students] should show some evidence of understanding mathematical concepts and procedures.
- *Proficient:* [Students] should consistently apply integrated procedural knowledge and conceptual understanding to problem-solving,
- *Advanced:* [Students] should apply integrated procedural knowledge and conceptual understanding to complex and non-routine real-world problem-solving.

The second paragraph provides examples of specific types of performances, for instance:

- *Basic:* [Students should be] able to use — though not always accurately — four-function calculators, rulers, and geometric shapes.
- *Proficient:* [Students] should have a conceptual understanding of fractions and decimals.
- *Advanced:* [Students] should be able to solve complex and nonroutine real-world problems.

Table 1 ○ 1994 NAEP Performance Levels for Grade 4 Mathematics

Basic: Fourth-grade students performing at the basic level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content strands. Fourth graders performing at the basic level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve simple real-world problems in all NAEP content areas. Students at this level should be able to use — though not always accurately — four-function calculators, rulers, and geometric shapes. Written responses that receive a *Basic* rating are often minimal and presented without supporting information.

Proficient: Fourth-grade students performing at the proficient level should apply integrated procedural knowledge and conceptual understanding consistently to problem-solving in the five NAEP content strands. Fourth graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how these solutions were achieved.

Advanced: Fourth-grade students performing at the advanced level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem-solving in the five NAEP strands. Fourth graders performing at the advanced level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. *Advanced* students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. Their interpretations should show deep understanding, and they should be able to communicate their thoughts clearly and concisely.

The second paragraph also discusses communication strategies, which include:

- *Basic:* [The student's] written responses are often minimal and presented without supporting information.
- *Proficient:* [The student's] written solutions should be organized and presented both with supporting information and explanations.
- *Advanced:* [Students] are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should be able to communicate their thoughts clearly and concisely.

Clearly, the ability to communicate the student's level of understanding is a central component in determining the mastery level of students in mathematics, and communication in writing is part of what is expected. It is noteworthy that writing is the specified strategy for communication at the *Basic* and *Proficient* levels, while at the *Advanced* level no strategies for communication are specified. Instead, students apparently have flexibility in how they can demonstrate what they know.

Why is this important? First, these mathematics standards limit the number and kinds of recognized ways in which some students must communicate knowledge and skills. Second, the standards tie the quality of written discourse to the quality of mathematics knowledge. That is, the standards say that the students at the *Proficient* level must show not only evidence of a higher quality of mathematics but they must perform with a greater proficiency in writing than their peers who are at the *Basic* level. The apparent attenuation of how students can demonstrate mastery and the tie between levels of evidence in a content area and level of writing are both especially distressing for limited English proficient students. Even if educators recognize exceptions, it is not uncommon for these types of associations to be made implicitly, which often puts students with limited abilities in writing at a disadvantage.

Weighing Non-Text and Non-English Performances Properly

It is reasonable to expect students to know how to understand and communicate, through reading and writing, within specific content areas. What is questionable is that these requirements be linked *inextricably* to mastery in content areas other than language arts.

Sometimes there is no explicit relationship in the performance standards between literacy and mastery in subject areas other than language arts. However, all too often, assessments are designed and administered in such a way that students can not do well unless they have these skills. It is important that students be given the opportunity to demonstrate what they know. To this end, performance standards should specify the goal. If the goal is to see students demonstrate mastery, then the standards should recognize explicitly that students demonstrate this mastery in a number of ways.

Sometimes measuring literacy within content areas might be one component of the total evaluation of mastery in a subject such as mathematics or science. Assessments can be designed so that a certain section evaluates students' literacy in specific subjects. This allows LEP students to perform well on some aspects and not as well on others and still be able to achieve mastery.

Sophistication of writing skills should not be a gatekeeper that keeps LEP students from attaining higher performance standards in subjects other than writing or language arts.

The apparent attenuation of how students can demonstrate mastery and the tie between levels of evidence in a content area and level of writing are both especially distressing for limited English proficient students.

Recommendations for Aligning Performance Standards to Content Standards

1. Unless explicitly identified in the content standards, quality of written discourse should not be tied to the quality of knowledge and skills in subject areas other than language arts.

2. LEP students should be able to demonstrate their skills and knowledge. In subject areas other than English language arts, they should be given the opportunity to demonstrate knowledge in ways that do not depend on literacy in English. This demonstration would probably involve non-text performances and/or performance in the students' native language.

3. If literacy within content areas is part of the total evaluation of mastery in a subject, it may be appropriate to identify performance criteria that specify literacy as one component of the evaluation. This would satisfy the requirements of literacy within subjects, and still allow students who have difficulty in writing or reading to demonstrate their knowledge about the subject in other sections of the evaluation.

Chapter 3

Test Specifications and the Accessibility Framework for LEP Students

Test specifications outline what the test covers for each content area at each grade level tested. If the test has more than one form, there usually is a schemata for determining form coverage as well as test coverage. These specifications form the bridge from content standards or other subject matter domain descriptions (such as those traditionally used by test publishers to describe and define the specific content area) to what items will be included in the test. Typically, the standards or domain descriptions take the form of curriculum guidelines or frameworks which are developed and used by state and local education agencies. For that reason, specifications are translations from curricular priorities into what must, should, and might be evaluated by the assessments.

Developing test specifications from content standards or domain descriptions is one of the first procedures in building a custom or semi-custom test for an agency. Test specifications (or test outlines of some type) also may be developed from state or local content standards to evaluate how well these frameworks are aligned with existing tests. Often publishers will present agencies with their perspectives about ways in which their pre-built tests "map" or "align" with the agency's curricular priorities (as stated in their curriculum frameworks). While these maps can be useful, it is critical that agencies develop their own specifications, based on their own sense of educational priorities, as they align tests to their standards or frameworks.

Whether tests are being selected or developed, the key points from the standards should be reflected in the test specifications accurately.

Whether tests are selected or developed, the key points from the standards should be reflected accurately in the test specifications. It is important to note that specifications vary in "grain size" or specificity within and across topics. Generally, agencies do not want the specifications to be turned into highly specific behavioral objectives, but these specifications must be precise enough that they retain the integrity of the standards on which they are based. In both cases (pre-built or custom-built) the number and relative weight of particular items or sections of items is a key consideration for alignment with the content standards.

Ensuring that test specifications are properly aligned to content standards is especially important because these specifications provide the framework for the development of all items and forms. They are the tool that determines holes in the test's coverage as well as the appropriate amount of coverage of important concepts in the curriculum. If the test specifications show that the tests and curriculum coverage is acceptable, then it is assumed to be acceptable throughout the reporting, test interpretation, and use stages.

Because test specifications play such a central role in test development and review, accessibility to the tests needs to be evaluated at the macro-level (namely, the overall interpretation of the standards by the test). To ensure validity for diverse populations, including LEP students, alignment at the macro level pulls together much of the validity work that is accomplished at the micro level (that is, validity of each of the items for LEP students). Evaluating accessibility at the macro level provides much-needed information: whether the micro-accessibility coverage is adequate or needs improvement. It also provides a periodic formative evaluation perspective from which to make mid-course corrections about what work under development and what work in the other phases of the test development and review process needs to be increased or reduced.

Robert Linn (1993) emphasized the need for tighter test specifications. He pointed out that these specifications should take statistical information about the irrelevant functioning of items by certain groups into account, as well as provide a more detailed schemata than typically has been the case — especially prior to the widespread use of standards. He saw this vehicle of tighter test specifications as a powerful tool. It can connect diverse information by guiding decisions about the topics that should be covered on tests, and about how items will be selected or rejected.

Components of Test Specifications

Test specifications provide a blueprint for how the content standards will be covered by the test. Content coverage is specified in terms of breadth (knowledge and skills) and sometimes depth (complexity and thoroughness of performance). These form the basis for what the test measures. In this section, we recommend that content coverage must be specified also in terms of its accessibility. Accessibility test specifications would highlight if the same content is being measured for all students.

Test specifications provide a blueprint for how the content standards will be covered by the test.

Breadth

Breadth consists of valued knowledge and skills within content areas that warrants coverage by test items. Knowledge and skills are defined by topics to be covered and usually also by process. Sometimes breadth includes identifying percentages of items at particular levels of difficulty within topics.

With the advent of the latest round of educational reform and the development of content standards by the national associations in various content areas, such as mathematics and language arts, a new emphasis is placed on process or types of knowledge and skills that are expected from students. Knowledge and skills can be rote, such as the memorization of facts or multiplication tables, or they can be conceptual — students' understanding of the relationships between multiplication and addition. Process also may involve applying problem-solving skills using rote and/or conceptual knowledge as the content.

Depth

Embedded in many content and performance standards is an expectation of performance capability best identified as depth. If we refer back to the NAEP fourth-grade performance standards in mathematics (Table 1), we see examples of depth at the different proficiency levels.

These examples of depth extend the expectations of performance beyond a mere list of topics. Typically, depth coverage has been assumed by item difficulty, the process aspect of breadth, and/or by coverage using various types of items (that is, multiple choice, short answer, or extended open-ended response items; these usually are seen as producing information about depth from low to high, respectively). However, these indices are insufficient. Items may be difficult because the fact is obscure, not because a more sophisticated performance is required. Likewise, a multiple choice "word problem" may require more depth of understanding and integration of diverse material than a highly structured, highly scaffolded, and extended response item. In order to ensure alignment between tests and expectations of depth which have been identified and specified in standards throughout the country, the aspect of depth needs be clarified in test specifications.

The Delaware Department of Education developed a mechanism for ensuring depth coverage as well as breadth coverage. They define breadth as a matrix formed by both the content and process dimensions. In addition to this matrix, they use a second matrix that focuses on depth. In this matrix, they define depth as a function of four variables:

- Approximate time required to complete an item
- Item scaffolding (ranging through three levels from step-by-step task guidance to no hints or guiding questions)
- Level of generalization (in three levels from highly specific items to items which require generalization)
- Complexity of process (in three levels ranging from managing only a limited amount of information to processing or considering multiple pieces of information and/or procedures simultaneously).

Each item is coded on all four variables. In assembling forms and tests, depth balance across items is guided by the content standards. It was developed from the original work of Rigney and Pettit (1995) who suggested criteria for assessing the depth and quality of student work in portfolios which were assembled as large-scale evaluation tools.

Accessibility

As stated previously, test coverage can be conceptualized in terms of breadth (content and process) and depth. However, are test items accessible to all students? Are the tests measuring the same breadth and depth for all students? Are provisions adequate so that all students have the opportunity to understand what is being asked and to demonstrate what they know? Most, if not all, standards are written to include all students in academic programs. If evaluation of breadth and depth coverage in terms of accessibility is assumed, but not explicitly specified in test specifications, grave errors in opportunity and interpretation of scores are likely to result.

Explicitly defining accessibility specifications for tests explains the publisher and/or agency's commitment to accessibility and how they plan on attaining it. This is true whether tests are being developed or if they are being reviewed, selected, or retrofitted.

The following section identifies one way agencies or publishers might approach the development and implementation of accessibility test specifications.

The Test Specifications Accessibility Framework for LEP Students

The Test Specifications Accessibility Framework for LEP students can be found in Table 2. It intends:

- to ensure that LEP students can access the requirements of each item in an assessment, and

Table 2 ○ ELL Accessibility Framework

The ELL Accessibility Framework is designed to guide the development of assessments and evaluate the accessibility of completed tasks.

Coverage

- ELL accessibility coverage focuses on ensuring that language minority students are able to access what is required on the tests, and are able to demonstrate what they know. For this population this means, for the most part, being able to confound the tests with respect to the students' English reading and writing limitations, and minimize cultural interference.
- Test accessibility is viewed as a composite evaluation of how each item in a form and test is made accessible through a combination of the item presentation, and the administration and response conditions available for that item on for the entire test. As appropriate, rubric accessibility is also examined.
- Accessibility coverage is determined by completing the Accessibility Matrix. For multi-form tests, one matrix will be completed for each form.

Density Expectations

- Maximum accessibility should be expected. This is accomplished by:
- Greatly reducing the reliance on restricted modes of communication.
- Offering multiple access avenues that take advantage of multiple student strengths.

Density opportunities should be equally spread throughout all aspects of breadth and depth coverage.

- to ensure that they will be able to demonstrate what they know.

Accessible test coverage is a function of how the item is presented on the assessment, the administration conditions, and the response conditions. Accessibility is also important in scoring, analyses, reporting, and use. However, these latter elements build upon accessible test coverage, and do not define it.

The Test Specifications Accessibility Framework for LEP Students is used in two ways. First, it guides the development of assessments. Second, it evaluates whether the accessibility work which has been done during development is adequate. The evaluation occurs during the validation part of test development or when agencies review and select off-the-shelf tests.

The Framework is divided into two sections. The first section focuses on defining accessibility coverage. This is done by completing the Accessibility Matrix. The second section specifies density criteria: how accessible is accessible enough. This latter section is parallel to the section in the traditional content test specifications which specifies weights and percentages of coverage for each topic or content/ process cells, and sometimes includes area subsections within the larger cells.

Coverage: The Accessibility Matrix

The Accessibility Matrix of the LEP Accessibility Framework can be found in Table 3. The Accessibility Matrix demonstrates how each item in a test form is made accessible through a combination of the item presentation, the test administration conditions, and the response conditions. The Matrix guides the accessibility success of the development work discussed later in this *Guide*. Once matrices are completed for a new or pre-built test, for each content area and grade, they will be used to evaluate the adequacy of accessible test coverage during test development or the selection of an off-the-shelf test.

Table 3 ○ ELL Test Accessibility Matrix

The completed matrix will demonstrate if and how each item in a test, and therefore each test, is accessible for English language learners.

- As a test specification tool, the matrix guides the development of items to ensure they are each accessible to English language learners.
- As a test evaluation tool, the matrix demonstrates the items which are accessible and how, so that reviewers can determine if this coverage is adequate.

Directions

For each item, the Item Presentation, Administration, Response, and Rubric columns should be completed. Entries in the columns should explain how the item is accessible by using the information from the respective chapters of Ensuring Accuracy in Testing for English Language Learners (CCSSO, in press). For instance, under the presentation of items, reviewers might specify plain language editing, use of simple visuals, and access to tools and resources (this includes a list of which tools are available and should be useful for the particular item).

The Access column should indicate if the reviewers judge that the items appears to be accessible by indicating Yes/No. If no, reviewers can recommend what should be done to enhance accessibility. The total number of items which are found to be acceptable on the final matrix should guide reviewers in determining the adequacy of the test for English language learners.

The matrices are completed by filling out the appropriate information by item, for each form of the test. Item information includes rubric information, in the case of extended response items. Those completing the matrices are asked to list how each item is accessible by completing the presentation, administration, response, and rubric access columns. Typically, entries in the columns follow the information from the chapters found in this *Guide*. For instance, Chapter Four focuses on accessible presentation of items. Examples from Chapter Four which can be included in the matrix include specifying plain language editing, use of visuals, and/or open access to tools and resources. They may specify which tools and resources are available or should be useful for a particular item.

ITEM	PRESENTATION	ADMINISTRATION	RESPONSE	RUBRIC	ACCESSIBLE

Density Expectations

This section of the LEP Accessibility Framework answers the question, "How accessible is accessible enough?" We recommend maximum accessibility. That is, agencies responsible for the development or selection of the assessment should expect that all students will be able to access virtually all the items. This should be done both by expecting maximum access to item requirements, and by clearing pathways so all students can demonstrate what they know.

While it is unreasonable to assume that all items will have visuals or performance activities, it is quite reasonable to expect that the items adhere to several of the recommendations outlined in the chapters on item writing and accommodations. This suggests thinking about broadening accessibility in some ways for all items, such as clarifying expectations and using plain language editing. It also suggests opening up as many items as possible so that students with different strengths can access the item requirements. For instance, students with highly developed kinesthetic intelligence benefit from having access to a range of tools and resources throughout the test; visuals are useful for visual learners; and contextual information, if organized well, can offset cultural and prior experience issues.

Density of accessible items is determined by evaluating the completed Test Accessibility Matrix for a test (see Table 3). Evaluators may complete a matrix by selecting a random set of items for a multi-form test which is being reviewed. In either case, evaluators are then asked to determine if the conditions are such that the items are sufficiently accessible for LEP students, and to place their judgments in the last column of the matrix. The total number of accessible items relative to the total number of items per test provides the percentage of density.

We urge test developers to set density expectations prior to developing tests or selecting final test forms. We urge those who are reviewing, selecting, or retrofitting tests to identify current levels of density, decide on an agreeable density level, and work with the publishers to meet this level in the selected test. This target may be phased in over a couple year period.

We urge test developers to set density expectations prior to developing tests or selecting final test forms.

Recommendations

1. **Accessibility of a test is determined by item by completing the Accessibility Framework and its matrices.** Accessibility is viewed as a composite of how items are presented to students, how they are administered to students, and how students are allowed to respond. For instance, in some cases how an item is presented may make it sufficiently accessible. In another case, one or more administration procedures may make it accessible, or accessibility may be a function of a combination of presentation and how students are allowed to respond.

2. **Density expectations should be spread equally throughout breadth and depth of a test to ensure adequate accessible test coverage.** Accessibility should not be focused within only one or two topics or at one depth level.

3. **It is not acceptable merely to include one or two token "accessibility" items, or items that allow only one or two types of access.** Admittedly, it is impossible to ensure perfect access because students have different strengths and limitations. However, heightened access results when tests loosen their reliance on certain modes connected to the limitations of many LEP students.

Chapter 4

Presentation Accommodations: Accessibility in Writing Items

Item writing is an iterative process, with several significant steps. To produce a test that is valid and accessible to all students, including LEP students, the items themselves must be as valid and accessible as possible. Therefore, in order to increase the assurance that large-scale mainstream testing instruments measure what they purport to measure for the LEP student population, it is important that accessibility be produced and evaluated at the item level, as well as at points where test level concerns are the issue.

We make an artificial distinction in this *Guide*. This chapter will focus on the writing of items while the following chapter will focus on the development of rubrics for the constructed response items. Actually, when constructed extended open-response items and performance tasks are written, the rubrics or scoring guides should be written at the same time. While some test developers or agencies use generic rubrics which are built independent of specific items, it is still important for item developers to be aware of ways in which the evaluation of student work as defined by these rubrics affects the items being written.

We do not differentiate in this chapter between the writing of different types of items (short answer, multiple choice, extended open-response, and various kinds of hands-on performance tasks) except when noted. Items should be accessible, no matter what type they are. Certainly, different kinds of items pose special issues, including accessibility issues, and some problems of access are relevant to most or all items.

Literacy sophistication should be kept to a minimum in assessment items, since the purpose is to impart a clear stimulus or signal to which the test-taker is expected to respond.

General Considerations When Writing Accessible Items

The sophistication of literacy should be kept to a minimum in assessment items, since the purpose is to impart a clear stimulus or signal to which the test-taker is expected to respond. One exception is the deliberate use of certain vocabulary germane to a subject area which should be learned as part of learning the concepts the vocabulary represents. For instance, students in geometry should be expected to know what pentagons and hexagons are. The other exceptions, of course, are in the evaluation of mastery in language arts. Reading passages are selected or written to contain the appropriate level of reading sophistication for the grade level being assessed, and writing rubrics specifically define and evaluate the sophistication of the written responses. However, even in language arts, items themselves can be written for clarity and evaluated for complexity.

There are several fields and bodies of work where clear, effective communication is central: technical writing, advertising, linguistics, computer software technology, English as a second language, and the field of learning disabilities. Some learning and cognitive psychologists as well as language arts specialists also focus on this issue. These fields share certain common elements relevant to responses in student testing. These elements include defining what makes language clear or plain, what types of surroundings enhance or constrict the effective communication of the language — including text format, visuals, and other information which is given or afforded the reader — and what modes of information presentation and receiver response transmissions are most effective in which situations.

The influence of culture is often misunderstood and underestimated for LEP students. This is particularly relevant for language minority students who have lived in and experienced a culture diverse from that of the U.S. It is also an issue, to varying degrees, for students who grow up in the U.S. but within environments divergent from U.S. mainstream culture. At times, item writers do not consider fully how expectations related to diverse experiences or value systems affect students if they come from very different cultures.

The parameters defining effective communication and use of cultural assumptions in items should hinge on what knowledge or skill is supposed to be measured in an item. While it is reasonable to expect that parameters of time and funding costs will affect testing products, communicative and cultural misunderstandings should not distort tests to the extent that the integrity of what is being measured is compromised.

At times, item writers do not consider fully how expectations related to diverse experiences or value systems affect students if they come from very different cultures.

Nine Elements in Writing Accessible Items: Recommendations for LEP Students

It is beyond the scope of this *Guide* to conduct a thorough review of all the common accessibility elements which could affect how items are drafted. However, we provide an overview of nine of the most important and relevant elements combined with sets of recommendations for each element which discusses selected intervention strategies for LEP students. The nine elements we discuss include:

- Defining what items measure
- Clear expectations
- Use of plain language
- Use of plain format
- Use of simple visuals
- Access and contextual information
- Access to tools and resources
- Access and performance activities
- Additional issues of native language and culture.

Defining What Items Measure

Item construct statements explain what specific knowledge and/or skills the items intend to measure. One or two bulleted phrases are often sufficient, particularly when they define what is being assessed in multiple choice and short answer items. The statements also should specify any prior knowledge or skills expectations which are defined as necessary building blocks to what is being measured.

Test developers typically do not write statements about what each item is supposed to measure. The practice is more prevalent when extended open-response items or performance tasks are under development because the measurement requirements for these items and tasks is often more complex than the requirements for multiple choice or short, constructed response items.

Statements about what is measured are critical for those championing accessibility because these statements explicitly define the validity intent, or exactly what is supposed to be measured, for each item. It is important to know what specific content and/or skill(s) are supposed to be assessed in items, so it can be evaluated if other skills or content are required which have no relevance to what is being measured. In this way it will become clear which conditions, if any, can vary, so that the integrity of the construct will remain intact but the irrelevant factors will be minimized for students who need it.

Recommendations for Item Statements

1. We recommend that statements be written for each item on a test. Many publishers and contractors will balk at this expectation (mainly because it is not part of their regular procedures, and they usually have vast banks of completed items with no statements). The statements also specify exactly what the item asks students to know and/or do, and item developers (publisher/ contractor staff or educator-writers) have typically not been held responsible at that level. But once this expectation has been included in the development routine, it actually takes very little time and often makes item-writing easier.

2. Agencies may work with their publishers to phase in this requirement over a couple of years when new or retrofitted tests are being developed. If existing tests are reviewed or selected, we recommend that statements are requested for a random sample of items prior to review.

3. Statements should be written in such a way that it is clear not only what is being measured, but what is not being measured by the item. This will help writers know what part of the items can be made more accessible for LEP students.

Clear Expectations

One of the most important elements in writing a "good" item is to be very clear in the item about what the student is being required to know and/or do. Sometimes it is easy for educated adults to make assumptions about what is expected, which often produces misleading results about student mastery when students have to "guess." It is especially problematic when the student's experiences are different from mainstream culture and its expectations, as would be the case with a large percentage of English language learners. An example of an item which explains how students can respond can be found in Figure 1. Specifically, we recommend that there are three aspects of items that can be reviewed for clarity.

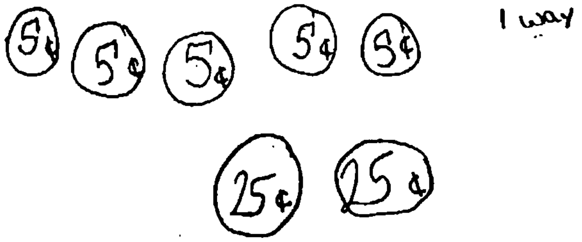
Recommendations for Clarifying Expectations in Items

1. Student response options must be clearly stated. For instance, if students are being asked to explain themselves, it should be made explicit that pictures, diagrams, charts, and/or algorithms will be acceptable if it is appropriate. Likewise, if communication via writing is expected, this also should be made clear.

Figure 1 ○ The Vending Machine

Maria wants to buy a 75-cent snack from a vending machine. The machine takes only nickels, dimes, and quarters. Maria has 7 nickels, 5 dimes, and 2 quarters.

Part 1
Show all the different ways she could pay for the snack. You may use words, diagrams, or charts.



5 dimes plose 1 more 25¢ = 75

2. If the quality of the students' writing, and/or the presentation of information in the charts, pictures, and other visuals will be evaluated, then it must be explicit that these will be evaluated. The evaluation criteria for the presentation of the material should be made clear to the student, either through the directions, as part of the item, and/or in rubrics provided prior to the test. This is over and above the criteria about the evaluation of the subject matter covered in the item.

3. Any additional requirements or constraints should be spelled out. If students are expected to anticipate the concerns of the "audience" in a constructed response item, then the type of audience for the item must be specified explicitly so that the type of concerns will be clear. Students should be told specifically to anticipate, state, and respond to the concerns. While writers might think certain expectations are obvious, if they are not explicit in the item, then they are subject to honest misinterpretation in the responses.

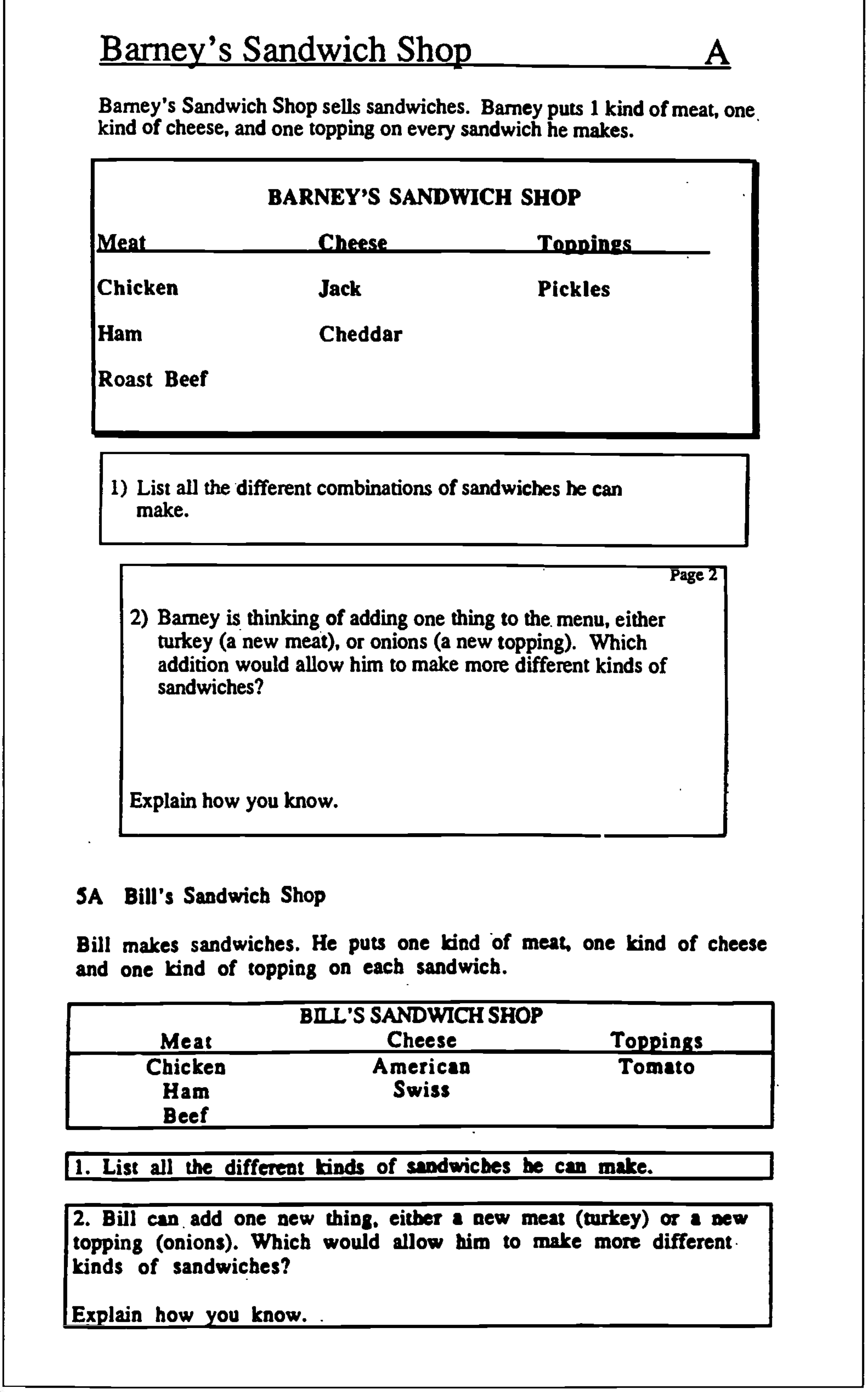
While writers might think certain expectations are obvious, if they are not explicit in the item, then they are subject to honest misinterpretation in the responses.

Using Plain Language

In order to guard against unnecessary language complexity, test publishers routinely route most or all of their item-related text through procedures which flag readability problems, and subsequently edit items and some text passages. However, these screens do not appear to be sufficient for a significant portion of students, including many LEP students (LaCelle-Peterson and Rivera, 1994; Hanson, Hodgkins, Kopriva, Saez, and Brooks, 1996; 1997; Abedi and others, 1995; 1997).

We make seven recommendations intended to improve the accessibility of assessment text material. These apply not only to the items and contextual passages, but also to the directions and any other text in which the language is not salient to what is being measured. We use the term "plain language" to distinguish

Figure 2 • The Sandwich Shop



Barney's Sandwich Shop A

Barney's Sandwich Shop sells sandwiches. Barney puts 1 kind of meat, one kind of cheese, and one topping on every sandwich he makes.

BARNEY'S SANDWICH SHOP

Meat	Cheese	Toppings
Chicken	Jack	Pickles
Ham	Cheddar	
Roast Beef		

1) List all the different combinations of sandwiches he can make.

Page 2

2) Barney is thinking of adding one thing to the menu, either turkey (a new meat), or onions (a new topping). Which addition would allow him to make more different kinds of sandwiches?

Explain how you know.

5A Bill's Sandwich Shop

Bill makes sandwiches. He puts one kind of meat, one kind of cheese and one kind of topping on each sandwich.

BILL'S SANDWICH SHOP

Meat	Cheese	Toppings
Chicken	American	Tomato
Ham	Swiss	
Beef		

1. List all the different kinds of sandwiches he can make.

2. Bill can add one new thing, either a new meat (turkey) or a new topping (onions). Which would allow him to make more different kinds of sandwiches?

Explain how you know.

the editing procedures from a perception that modifications reflect a dumbing down of the item requirements. We emphasize that editing clarifies expectations and should make the item requirements clearer and more accessible for many students.

Figure 2 provides an example of Plain Language and Plain Formatting. The first item illustrates the original item; the second item is a revision with plain language and plain formatting changes.

Recommendations To Improve Accessibility of Text Material

1. **Item sentences or stems must be kept brief and straightforward, with a simple sentence or phrase structure.** No additional clauses or phrases should be used, and the same sentence structure (e.g., subject with optional adjective first, followed by a verb, and then an object with optional adjective and/or adverb) should be retained throughout the assessment, for all items, as possible.

2. **Consistency in paragraph structures should be employed.** The range of organizational structures in paragraphs needs to be restrained and held constant throughout the test, to the extent possible. For instance, paragraph structures may all begin with the topic sentence, include two explanatory sentences, and end with a closing or summary sentence. While complexity and variation in structure might represent good writing, they make tests more problematic for some English language learners, as well as some other students.

It is common for these students to be unable to recognize the item's requirements (e.g., what the item is asking them to do) when the item is presented in more complex stems, sentences, or paragraphs. This is minimized when stems and sentences are consistently of one structural type and when paragraphs are structured in the same way.

The focus in "plain formatting" assessments is to minimize or offset sensory overload.

3. **The present tense and active voice should be used as much as possible.** The rationale is the same as the rationale regarding simple structures in stems, sentences, and paragraphs. Present tense and active voice typically are learned first, when LEP students are learning English in U.S. classrooms. Other tenses and voices very easily can camouflage the measurement intent in items for ELLs, as well as for other students who are challenged by reading or language.

4. **Rephrasing or rewording ideas and sentences should be kept to a minimum.** Paraphrasing words or ideas should be recognized explicitly, and either not done or done with all the original words immediately following in parentheses. This means using the same words to refer to the same phenomenon, concept, person, place, thing, action, or modifier rather than using a variety of words. While repetition is considered poor writing, use of variant terms makes comprehension significantly more difficult.

5. **Pronouns should be used in a very judicious manner.** The same word should be used repeatedly, rather than using pronouns to refer to people, things, or ideas. The complexity engendered by multiple pronouns is needlessly confusing. When pronouns are used, they should be followed immediately by the term they reference in parentheses, as necessary and possible.

6. **High-frequency words are the best choice.** Word frequency estimators are published and available. Sometimes frequency is measured in context: for instance, pepperoni is considered a higher frequency word than anchovies when pizzas are the topic. For LEP students, high-frequency issues are sometimes mediated by cultural experiences.

7. Words with double meanings or colloquialisms should be omitted or defined in the text. Both words with double meanings and colloquialisms are used far more than we realize. It is important that someone skilled in recognizing these words and phrases review all text. A few examples of double meanings in science include earth (the planet) vs. earth (soil); plate (as tectonic structure) vs. plate (hot plate); and fault (geologic feature) vs. fault (error). When words with double meanings or colloquialisms are used, they can be defined in the text itself or in parentheses immediately following the word or phrase.

Using Plain Format

The focus in "plain formatting" assessments is to minimize or offset sensory overload. For many students who struggle because of linguistic challenges and visual orientation issues — as well as students who suffer from anxiety, processing, and/or activity challenges — several ways in which we package tests can be extremely distracting and/or overwhelming. This population includes many LEP students who struggle with overload because they must translate back and forth in their heads between their native language and English. English language learners also benefit from formatting that breaks up text or presents other sources of stimulation beyond text. These recommendations can be beneficial for LEP students of all ages, not just elementary school students.

Recommendations for the Use of Plain Format

1. Visuals are often important, but they should be kept simple and to the point. These are useful for LEP students because they provide another source of information about the item's expectations.

2. Large print forms should be available for LEP students. While these typically have been produced for the visually impaired, a number of teachers of learning disabled students report that these forms help students of all ages focus on the text. Several teachers have suggested that other students — for example, some English language learners and students not from any identified special population — have experienced the same benefits from large print forms as well. Several font sizes have been recommended, with 18 point often cited as useful and manageable. This font size appears to be beneficial for students, not too costly to produce, and not terribly bulky for teachers to handle along with other test forms.

3. At least some forms should omit the use of item columns, limit the number of items per page, and/or provide students with a template to use. The goal of each of these format adaptations is to reduce textual clutter, allowing students to focus more easily on any particular item or section. Item columns may be efficient, but teachers report that they are a central source of overload and confusion for some students, especially those who have challenges with written text (Hodgkins, 1997). This certainly includes some English language learners, particularly those whose English literacy proficiency is minimal and those students who do not have much experience with text.

Limiting the number of items per page usually means using the entire page, but spreading out a lesser number of items than usual with plenty of empty, white space between each one. Use of a template serves the same purpose as the white space when the number of items are limited per page. Similar to surrounding items with empty space, templates (usually pieces of cardboard with holes cut out of them about the size of an item) restrict or focus a student's field of vision on a particular item. However, the templates are only helpful if students are accustomed to working with them.

4. Sample templates should be part of advance sample materials which teachers should be encouraged to use in class, but as few as possible should be used in an assessment. If formats include item columns, small print, and/or many items per page, they are particularly useful. However they can be confusing ("which one to use?") or embarrassing ("I'm being treated like a baby") unless care is given to the manner in which templates are presented.

5. Lines or "boxes" which frame text or answer space should be used in a judicious way. These formatting tools are often useful in organizing text in assessments. Care must be taken, however, to avoid distraction with fancy borders, and their purpose must be clear. For instance, in one assessment students were asked to "provide your answer below." The item was framed in a box to draw the students' attention to it and white space below was where students were supposed to compose their answer. However, several students tried to squeeze their answers into the item box.

Teachers of LEP students suggest that simple visuals be used as frequently as possible.

Another common "line" problem highlights the advantages and disadvantages of placing lines in the booklet on which students can write. Spacing and numbers of lines can restrict expression or imply that a lengthy response is expected, whereas leaving white space may be too ambiguous. Educators should guide the spacing requirements: the number of lines and/or amount of white space should be part of the review discussions.

6. Forms that include these recommended adaptations might be given to a wide variety of students who request them, including teacher requests for students. This may include not only students identified as limited English proficient but others who can benefit from the adaptations. Some of the recommendations can be accomplished on all forms; some, such as large print forms, will be produced as an alternative to a smaller print form with the same items. When format-adapted forms are published, the varied formats should occur across all or several forms of the test to avoid the effect of certain students taking only certain forms.

Use of Simple Visuals

Teachers of LEP students suggest that simple visuals be used as frequently as possible to facilitate the understanding of what is being asked/presented in a specific item or group of related items. However, the illustrations or charts should not be so complicated as to distract from the intention of the item.

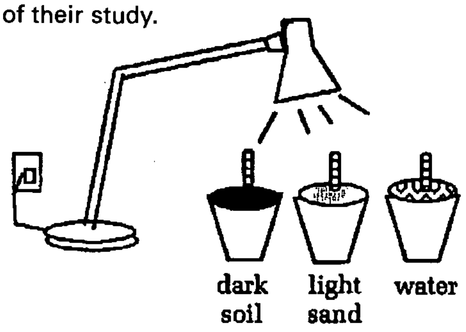
An example of an effective visual appears in a local assessment in California. The item, *Where on Earth?,* can be found in Figure 3.

This visual is effective because of the bendable nature of the lamp and clearly separate beakers with distinctly different textures to represent water, soil, and sand. This provides a visual grounding and sequencing for students that parallels the statements and expectations in the text of the item. The thermometers underscore the heating aspect of the question. The words "dark" and "light" are used to describe

Figure 3

Where on Earth?

Mrs. Flores' class has been learning about the position and composition of the Earth and the interaction between the earth and the sun. The following experiment was used in part of their study.



Using everything you know about light, heat and water, how do you think this activity helps explain the uneven heating of the Earth's surfaces?

the soil and sand and to facilitate accurate understanding of the item. It is also important that the picture contains no other information that could distract students.

Recommendations About the Use of Visuals

1. **Visuals should be used to facilitate the understanding of what is being asked or presented in an item or group of items.** Remember *why* the visual is being used.

2. **Visuals should mirror, or parallel, the item statements and expectations.** This mirroring enhances the items by providing two sources of grounding and sequencing (the same information in the text and in the illustration or chart).

3. **No "supplementary" or unnecessary information should be placed in the visual to distract students from the requirements in the item.**

4. **Each major part of the item should be represented in the visual.** While additional, unnecessary information is confusing to students, omitting important pieces of the item also can be misleading.

5. **Simple text can and should be used in the visuals that correspond to important words in the item.** As in the example above, using "dark" and "light" in the illustration is important not only because the words clarify which beaker holds soil and which holds sand, but because they are the same words as in the text and they act as a bridge to connect the item text and the illustration.

Contextual Information

Contextual information consists of introductory or explanatory text which is part of items or blocks of items. For instance, to introduce an eighth-grade 1996 NAEP science performance activity and several related items about salt solutions, an introductory section placed the exercise within the context of water in natural ecosystems.

The section also explained what would occur during the exercise. Throughout the activity, directions were included prior to different steps of the performance, with explanations about how to complete the various charts and other preparatory information. These portions were all in addition to the items themselves.

There are advantages and disadvantages to including contextual information, particularly for LEP students. To some extent, effects are mitigated by how the contextual information is conveyed to the student. The solution, for the most part, is to maximize the advantages contextual information affords, and minimize its disadvantages.

It is important for educators to remember that items which include contextual information are often measuring different parts of constructs than items which are set in a more sterile environment. If LEP students are to be evaluated with the same kinds of assessments as other students, the issue then becomes "How should LEP students be tested with items which contain contextual information?," not "Should LEP students be tested with items which contain contextual information?"

Using tools and resources means abandoning the attitude that they allow students to "cheat" or feign knowledge.

Recommendations for Using Contextual Information

1. We recommend, in general, that contextual information be included in assessments for LEP students. Inclusion is recommended because:

- The information usually places items within a context that is often more real for students
- Contextual introductions can help "even the playing field" by providing background information rather than relying on prior knowledge
- Step-by-step explanations used throughout a complex assessment exercise provide a sequencing framework.

2. The disadvantages related to including contextual information for LEP students should be minimized. This can be done by following the guidance presented in this document, and integrating it into how the contextual passages are presented. The primary disadvantages are:

- The explanations and introductions are often lengthy, and reading intensive;
- Students are typically required to read the information on their own as part of the test.

Test directions are not routed routinely through existing readability indicators, much less the plain language and format recommendations discussed here. The same is true with contextual introductory and explanatory sections. Information is often lengthy, providing a vivid contextual frame which is entertaining to those of us who read well and easily. But to LEP students, the same information may be confusing, exhausting, and cumbersome. Frequently, the full explanation and set of directions is given all at once before an activity and/or block of items, and students are expected to remember or refer back to the directions while they move through the exercise. This is doubly difficult for those with linguistic challenges, and unnecessarily complicated. As a result, the intensity with which students must deal with the language can significantly distort how students respond to the items.

3. General considerations for minimizing the disadvantages include:

- Performance activities should be included if at all possible which give substantial access to those with kinesthetic, tactile, spatial, and related strengths

- All text, including directions and introductory text, should be routed through plain language and format procedures
- Directions and explanations should be read aloud, as students follow the text in their booklets
- Directions should be spread throughout the exercise
- The amount of text the students were expected to read on their own should be kept to a minimum
- Visuals should be incorporated if possible
- Pretest discussions before a set of items and/or a performance activity should enhance access.

Access to Tools and Resources

Limited English proficient students, as well as many other students, rely on their kinesthetic, tactile, spatial, and other strengths to help them learn. They also rely on these strengths to understand and process the requirements and problem-solving expectations of test items. Incidentally, these are strengths commonly used in the workplace: Engineers and designers regularly make models and use 3D software to conceptualize problems and define solutions.

We define tools and resources as physical aids, such as mathematics manipulatives, or informational documents of some type. In all cases, the tools and resources which are discussed here refer to age-appropriate aids which do not compromise the integrity of what assessment items are intended to measure. This means that the tools and resources should not give some students an advantage in understanding what items are requiring or how students are allowed to demonstrate what they know. However, earlier arguments have discussed that typical paper-and-pencil tests disadvantage some students, including many LEP students. Therefore, it may be appropriate to allow students to use selected tools and resources to offset the disadvantages of tests which have substantial literacy expectations.

Using selected tools and resources during the testing process has become more commonplace, such as including paper tiles or rulers in test booklets for students to punch out and use for specific items in mathematics. While encouraging, this progress is not sufficient.

Cautions and Costs

Obviously, it is possible to overdo. For instance, computers can allow students access to vast amounts of information, depending on software, programs, and hardware capacity. Using computers to provide standardized sets of information and manipulation capabilities is a good idea. However, it is possible to distract and over-stimulate students. We encourage educators and assessment experts to find a balance which is richer and more accessible than current testing environments, but not uncontrollably inundated with technology.

Some agencies have taken the position that any tools that the content standards encourage schools to acquire and use can be allowed during the assessment. Depending on the choice of tools and resources, there should be at least one set per classroom. The cost is born by the districts, as long as time and security arrangements are effective. Staff from the test publishing company, Advanced Systems in Measurement and Evaluation, have said that to ensure security, there can be a maximum of two testing shifts which immediately follow each other per subject area/grade. This, of course, has implications for resource levels.

Recommendations for Using Tools and Resources

1. **Test-givers should allow reasonable access to a set of tools and resources throughout the test, and remove the specific tool or resource when items explicitly measure that specific skill.** When this practice occurs, students are allowed to access their strengths to solve problems, except when those skills are being evaluated.

For instance, mathematics or science items which measure computation or estimation skills can be done in a block when calculators are not allowed. Other than that, calculators should be allowed for all items: whether students need them to calculate information (these items then might measure students' ability to determine which operations to use), or whether students use them as a ruler or building block. The same use would apply to dictionaries and other resources. While this sounds daunting, once completed, policies and many resources can be recycled over time.

Some item-drafting problems are more specific to English language learners.

Using tools and resources means abandoning the attitude that they allow students to "cheat" or feign knowledge. This is why it is so important that everyone, *beginning with item writers and test developers,* knows *specifically* what items measure and what they do not measure.

Withholding resources from an entire test because of selected items is the same as allowing them for all items, regardless of what items seek to measure.

2. **Local content standards should define what tools and resources are useful and allowable.** The integrity of the standards should not be compromised. Students should not be allowed to use resources directly related to what is being assessed.

3. **Reasonable access means that most items (not just one or two per test) will be accessible to students with a range of diverse strengths.** This includes reasonable access to tools and resources which may not be directly related to an item, but which are also not compromising the integrity of the item. This allows students some freedom to solve problems in what we might consider to be unconventional ways.

4. **Who decides which tools and resources to use and when?** It is recommended that these decisions be made by key individuals who include:

- Content-matter specialists who are very familiar with the content standards and local needs
- Specialists who are involved in the development or selection of the assessments
- Specialists from special needs, gifted, and vocational programs who have first-hand knowledge of local resources and how to access the strengths of these students
- Test contractor's representative
- Funding agency representative with political and funding knowledge.

5. **Tools and resources can be assembled from a variety of places.** Besides producing disposable tools as part of the assessment materials, other sources include:

- Trade information from businesses and government agencies
- Public television and radio programs

- Computer companies or computer departments at universities where resources can be found which can be adapted, and/or people might volunteer or be paid honoraria to lend their expertise
- Mainstream teachers, teachers of special needs students, gifted students, and those in vocational and creative experimental programs.

6. There are many issues and questions to balance and consider when making decisions about resources. We recommend that considerations such as these identified below be discussed thoughtfully with a range of stakeholders, including educators of LEP students. For instance:

- Item or testing environments can be too complex and extensive. The goal is to find a balance which is richer and more accessible than the environments of today, but not uncontrollably inundated.
- Who pays for the resources and tools, and how extensive a set should be provided? Should there be enough for one set per student, one set per classroom, or one set per four students? Some of the tools/resources are expensive; is it worth the cost when students will probably only need and/or use them for a few questions?
- If students are allowed to share tools and resources, and/or if testing is done in shifts, how is security maintained?
- Time and scheduling constraints are very real and need to be addressed.
- Local educators need public relations help from the client agencies and publishers to be able to explain the reasons, rationale, and the balance of plus and minuses behind these types of enhancements.

Performance Activities

The pros and cons of including performance activities in large-scale assessments are a source of debate, mostly in terms of how they can provide enhanced opportunities for demonstrating mastery versus funding, time, and security constraints. However, the advantages of including performance activities in order to allow students access to item requirements cannot be overstated. These activities can provide extended opportunities for students to use their kinesthetic, tactile, spatial and other related strengths beyond what is possible on even the best paper-and-pencil test.

The recommendations outlined below provide suggestions which increase accessibility for LEP students and appear to be reasonable for publishers and agencies responsible for test administration.

Recommendations for Using Performance Activities

1. Performance activities do not need to be lengthy, complex, or expensive to be worthwhile. Fifteen minutes or so is often sufficient.

2. Performance activities can precede a set of items, occur as part of the assessment, or be used by students to respond to item requirements. Structuring the use of tools and resources often entails some kind of activity. Students might collect and record data by measuring their desks, shoe laces, or even their arms and legs, and then answer items having to do with estimation, interpretation, and extrapolation.

Class discussions might be useful starting points. A structured 15-minute conversation about the Persian Gulf war could lead into items that ask students to provide individual answers about the nature and implications of wars in general, and which require specific evidence from conflicts in or including the United States during its first 100 years as a country. The items would not cover the same materials as the discussion, but the conversation stimulates student thought about wars in ways that extend beyond simple considerations of weapons. Including music as a tool enhances item access in any subject area, whether used in mathematics or used to underscore the intent of a reading passage in language arts. Developing a short activity associated with the music tape, perhaps tapping to the music, discussing art as a conveyer of information, or asking for one or more item responses to be "written in music" (not necessarily formally) are also examples of productive performance activities.

3. Several local experts can provide ideas and evidence about the importance of including this opportunity in assessments. Experts who work with students identified as gifted, as well as teachers of special needs students (across the ability continuum), should be able to provide ideas and documentation to wary constituencies about who benefits from performance activities. Test publishers experienced with performance activities, subject matter specialists who are involved with educational reforms, and several noted cognitive and educational psychologists, can help provide credible empirical information regarding the relationship between performance access and scores.

Sometimes confusion arises simply because LEP students are not completely proficient in English.

Additional Issues of Native Language and Culture

Some item-drafting problems are more specific to English language learners. These include the impact of students' home language or limited proficiency in English, expectations in items which assume prior common experiences if a student grows up in the U.S., and expectations in items which assume a common value system to that typically endorsed in the U.S.

This section provides examples of where confusion might arise in items, based on language or culture. While not recommendations, per se, the examples which follow highlight areas in which test developers need to be cognizant. Using the recommendations from the other eight sections in this chapter, items can be drafted which alleviate types of problems outlined here.

Language Issues

Items can be written that unintentionally are linguistically confusing to English language learners. These may include the use of words or phrases which mean something different when translated from the student's first language. Also, confusion can result from rules of syntax or word order that differ in a student's home language. Yet another common source of student confusion comes from words that mean something different in English than in the student's home language.

Different symbols are used to indicate numbers or numbers operations in different parts of the world; for instance, periods are sometimes used instead of commas in mathematics in writing numbers; 3.001 is three thousand and one, but we would read it as three and one thousandth. In England the "3.5" is often written with the decimal point at the midpoint between two numbers rather than at the bottom: 3.5). This could be confused with the symbol for "dot product" in the American mathematical notation for multiplication. In another example, a "billion" is translated numerically into 1,000,000,000,000 in some Latin American countries. We would call this number a "trillion."

There is often confusion related to different monetary systems. Use of monetary words, such as dollar or peso, may mean different amounts depending on the country.

Sometimes confusion arises simply because LEP students are not completely proficient in English. The use of homophones can be confusing: hole vs. whole, break vs. brake, passed (read as past) vs. past (time), and reflex vs. reflects. Students may misunderstand the meaning of words because of typographical conventions unique to writing in specific academic disciplines. For example, words learned and used formally in certain science disciplines are also words in English which mean something else entirely. While fluent English speakers learn to listen to the context to ascertain the proper meaning, this is unnecessarily confusing for LEP students. Some examples discussed earlier include earth (the planet) vs. earth (soil), plate (as tectonic structure) vs. plate (one eats from), and mole (a chemical unit) vs. mole (a birthmark) and mole (a rodent).

Cultural Issues

Throughout this *Guide*, we caution readers not to underestimate the impact of cultural differences on the accurate assessment of student achievement. A test item is written to measure a certain skill, but the context in the item can alter the measurement intent for students from certain backgrounds or with different cultural/linguistic experiences. Sometimes the issue is item-specific and must be clarified or rethought; occasionally problems arise when scoring unexpected responses.

Two sets of cultural expectations seem to have a primary impact on how a student understands the requirements of an item: (1) expectations in items that assume prior experiences that are common if a student grows up in the U.S., and (2) expectations in items that assume a value system common to the value system typically endorsed in the U.S. These cultural expectations become especially problematic when a student's experiences or values are distinctly diverse from those typically experienced by the mainstream population in the U.S.

In some cases, items expect students to recognize certain things or events, or to have had certain experiences growing up. One example is a set of items which uses a parade as the context. In some parts of the world, parades are limited to political agendas, and espouse very different sets of feelings and perceptions than U.S. parades, even though many U.S. parades are patriotic in origin (for example, a parade on July 4th). Sometimes items ask a question using an object common in Western cultures but less common in other cultures (for instance, telephone booths or vending machines).

Problems that spring from diverse cultural value systems often can be overlooked, sometimes because item writers are unaware that the item is affecting a value that is not universal. One example is the assumption that, if requested, a student will provide an extended discourse about a particular topic, which often assumes they will predict the concerns or issues of the audience and include responses to those points in his or her answer. Some cultures do not value arguments or differing points of view, an explanation of something when they know the reader already knows the answer, or answers to questions or concerns before they are expressed. Other examples include assumptions that the student will understand the value of competition, or the value of individual rights.

To solve such problems, sometimes it is important to change the context or wording of an item, or to clarify expectations for responses to a problem whose context is set in a cultural experience. Minimizing language issues depends on how the items are evaluated by LEP students, by experts familiar with the students' primary languages, and by experts in the teaching of content and English.

Chapter 5

Writing Accessible Rubrics

Rubrics are scoring guides that accompany and evaluate how well students respond to constructed response items. There are "generic" and "item-specific" rubrics. Generic rubrics guide the evaluation of responses of a number of items in a general way, while item-specific rubrics discuss specific evaluation issues with respect to a given item. Rubrics should be developed concurrently with constructed response items; rubrics, as well as the items, should be reviewed and evaluated to determine if their expectations of performance are accurate and adequate for LEP students.

It is not unusual for test developers to begin by framing generic rubrics, and then build item-specific rubrics from the general model. Sometimes the generic rubric is retained and supplementary notes are written about the specific issues inherent in each item. Even when item-specific rubrics are used, more detailed notes often are developed to discuss item-related evaluation points. In either case, notes become an integral part of the rubric materials. For this reason, this chapter focuses on not only the rubrics per se, but also on any and all notes which accompany the rubrics and which are used to evaluate student responses to open-ended items.

It is not unusual for test developers to begin by framing generic rubrics, and then build item-specific rubrics from the general model.

What does it mean to write accurate and accessible rubrics? Such rubrics allow students to demonstrate what they know about what the item requests. Unless the item measures writing skills, per se, students should have flexibility in the ways in which they respond. In short, rubrics should be properly aligned with their respective items and they should be properly aligned for all students.

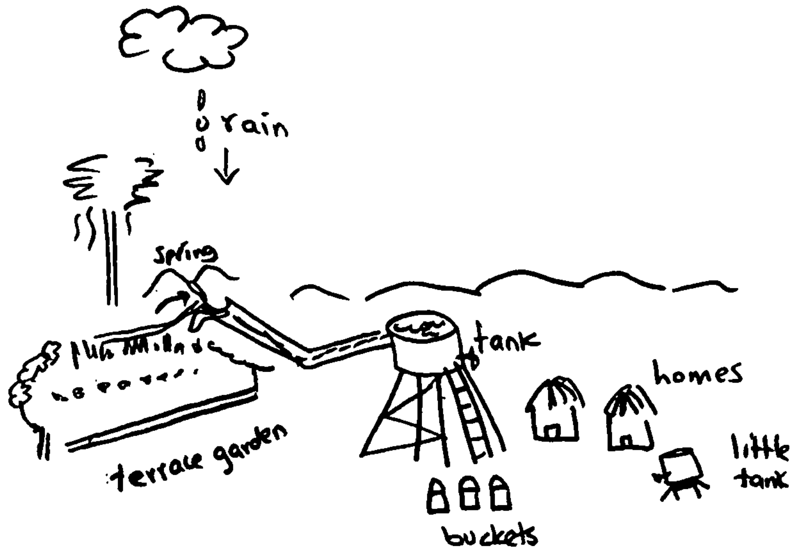
Why is proper alignment critical? Unintended literacy assumptions in rubrics, e.g., that a student will articulate what s/he knows through writing, can often limit LEP students' ability to demonstrate what they know. In addition, cultural expectations can limit access for English language learners. While these confounding elements will not be identified specifically by the items, it is not unusual to find expectations about literacy or cultural perspectives written into rubrics.

It also is not unusual to find these expectations about literacy or cultural perspectives written unevenly into rubrics throughout the score points. In the NAEP mathematics performance standards, discussed earlier in this *Guide*, literacy expectations occurred in the bottom two performance levels (Basic and Proficient). Sometimes, construct-irrelevant expectations such as these occur more frequently at the higher score levels where students are expected to respond with a certain level of sophistication (e.g., literacy sophistication or evidence of deeper, subtle knowledge of a value, experience, or perspective which is more common in the mainstream U.S. experience or belief system than in experiences or beliefs in other cultures). If LEP students know the content that is assessed, they should not be hampered by cultural or literacy limitations or differences. Rubrics must support this perspective in a consistent and even manner. Since cultural issues often are more difficult to pinpoint or identify than literacy issues, we discuss two brief examples related to culture.

Sometimes students understand the construct being measured, but their responses reflect prior experiences that may be very different from those experienced by students growing up in mainstream U.S. culture. For instance, a national assessment of educational progress recently included an item that asked students to define elements that would promote a healthy lifestyle. The test developers deter-

Figure 4 ○ Water Distribution

Draw and label a picture showing the source and distribution of water in a community.



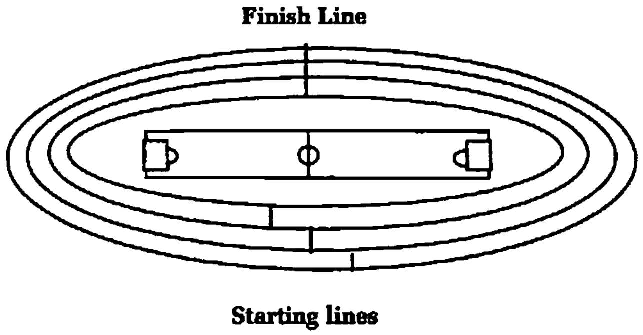
Sometimes the cultural issue reflects a misunderstanding based on diverse cultural value systems.

mined that proper diet and exercise are key to a healthy lifestyle; these were identified as the key parameters on the rubric. Scorers were given examples to help them determine whether specific answers were correct at different score points. Many LEP students from developing countries independently responded to this item by highlighting such issues as the need to have clean water, enough food to eat, and food free from toxins. Given the explanations of the parameters in the rubric notes, these answers were not given credit. We can argue that those LEP students understood "health," and that the points they raised do promote health and a healthy lifestyle.

Figure 4 illustrates a student's response about building a community water distribution plan. The original rubric and the associated notes expected a response based on sophisticated plans used in U.S. cities and towns. The response from this LEP student reflects his experience — unexpected, but many would argue it is a correct response given the item.

Sometimes the cultural issue reflects a misunderstanding based on diverse cultural value systems. For example, a recent mathematics item asked students to create a fair race course (see Figure 5). The rubric criteria expected students to create a race course in which all of the contestants have to run equal distances. However, some students interpreted "fair" to mean that all contestants have an equal chance to win. This was especially true in cultures which do not emphasize competition. As a result, these students created a race course in which the slower contestants ran shorter distances. Answers which reflect an accurate understanding about the relationship between speed and distance (the mathematics construct being measured) should be scored accordingly, regardless of the student's interpretation of fairness.

Figure 5 ○ Racing Track



Recommendations for Accessible Rubric Development for LEP Students

In order to produce rubrics that are accessible for LEP students, we recommend that five steps be followed during rubric development:

1. **Educators with particular knowledge of the literacy and cultural issues of LEP students should influence the development of *all* rubrics and all rubric notes.** These educators should draft some of the rubrics and notes, informally trying them out in (and getting feedback from) classrooms where there are LEP students. They also should be involved in the in-house sharing of ideas as rubrics and notes are built. These educators must be familiar with different sets of issues relevant to the home country cultures of the LEP students being tested locally, and a range of first language issues and English proficiency levels. Examples of work from LEP students should be included in the notes.

2. **Rubrics should be drafted to ensure that content measurement expectations are not confounded with literacy and cultural expectations, unless literacy or certain experiences or beliefs specifically are chosen for evaluation.** Rubrics should be properly aligned with construct statements of the items. If no construct statements exist, experts on LEP students will need input from content experts to determine if literacy or cultural expectations are part of the constructs that are being measured — and then write the rubrics and notes accordingly. Sometimes rubrics and notes need to be edited to remove any irrelevant literacy or cultural expectations. At other times, additional language clarifies the rubrics.

3. **Proper alignment between item and rubric expectations must occur consistently and evenly throughout all the score points so that students are not discouraged from meeting criteria at each and every score point.**

4. **Each item and rubric should provide multiple, explicit ways that allow students to respond.** This is consistent with the multi-modal approach that occurs in the classroom. LEP educators are proficient at encouraging a range of ways for students to demonstrate knowledge in the classroom, and therefore they will be very helpful working with test developers to broaden their ideas of ways in which this type of information can be gathered within the testing context.

5. **Most, if not all, rubrics should be free from improper literacy and cultural interference.** It is not acceptable to say a test is accessible because a mere one or two of the items allow non-text responses.

Improper literacy expectations usually arise when developers (a) rely too heavily on written responses to demonstrate mastery, or (b) expect a level of sophistication of written response beyond the expectation that students somehow successfully communicate to the reader what they know. Instead, rubrics should allow non-written information from students to augment or replace written responses. This information should include non-text paper and pencil responses such as charts, diagrams, pictures, and algorithms. If possible, students may also respond orally or perform the solution to a task (such as actually producing a chemical solution which would solve a chemistry problem).

Our intent is to separate cultural issues from content expectations. Students should be evaluated on the content expectations, regardless of the unexpected ways they might handle the contextual surrounding of the item. The differential impact of student experiences or value systems needs to be anticipated before tests are implemented. Item writers, along with reviewers, need to be alert to the influence of culture on responses. This influence needs to be written into the rubrics and notes so that scorers will know how to give proper credit to LEP student responses.

Chapter 6

Promising Administration and Response Accommodations

What test administration and response accommodations are beneficial for use with English language learners? Obviously, the central reason for any accommodation for LEP students is the increased accuracy of results about their mastery of content. No matter what the accommodation, the integrity of the subject matter construct must be retained.

A thorough discussion of administration and response accommodations is beyond the scope and purpose of this *Guide*. Instead, we focus on:

- Administration and response accommodations with particular promise for LEP students
- The rationale for using these accommodations
- An explanation of those LEP students who would benefit the most from specific accommodations
- Possible complications associated with the use of accommodations.

In her presentation at the American Educational Research Association's Annual Meeting (1998), Martha Thurlow summarized the current status of accommodation use in state achievement tests for students with disabilities. Her comments are relevant also to the status of LEP students taking the same kinds of tests. While some accommodations are more popular than others, there is little consistency governing which accommodations are allowed. There is also little rationale for accommodation patterns when scores of accommodated and non-accommodated students are to be aggregated. There is no cohesive logic that governs when and under what conditions tests taken under accommodated conditions are scored or "counted," even in a separated form from the non-accommodated results. The marked lack of research about the uses and effects of accommodations has resulted in guidance stemming mostly from political or cost-efficiency sources, rather then from the wisdom of best practice documented by empirical research.

Results from the Council of Chief State School Officers' 1998 annual survey of state testing directors point to the confusion that surrounds the selection of LEP students for inclusion in mainstream testing, which accommodations are allowed, if LEP students' scores are reported, and how the results from the accommodated students should be used.

Usually the agency responsible for building and/or administering the assessment determines which LEP students should receive accommodations. Since commercially published tests typically are not constructed to include a broad range of LEP students (which means that many English language learners have not been included in piloting or norming samples, especially under accommodated conditions), publishers might advise waiving students who would benefit from accommodations or from being included without accommodations in mainstream tests. This guidance from publishers may run counter to the standards-based expectations in states and districts that base program and possibly student-level decisions on test data assumed to be inclusive and comparable. Tindal (1998) describes three accommodation

Usually the agency responsible for building and/or administering the assessment is responsible for determining which LEP students should receive accommodations.

decision models which explain criteria for inclusion, comparability, and their subsequent implications.

We recommend a combination of approaches — including most LEP students in large-scale mainstream testing, and working simultaneously with publishers to build technical information about the validity and reliability of accommodated results. Initially this is awkward, because publishers have done little or no research and development to include broader ranges of students in tests, including both English language learners and students with disabilities. As technical information about accommodated scores is being accumulated, we recommend including the range of LEP students in the large-scale testing, using accommodations as necessary. These results initially can be reported with appropriate tentative caveats until definitive technical information is available. Note that it is imperative that systematic documentation about the validity of the scores begin to be collected immediately. It is not acceptable to record caveats year after year, when accumulated evidence could be collected, at most, over about a three-year period.

The need for flexibility is particularly intense when students are new to this country and/or have little proficiency in English.

Construct Validity Considerations

As we have said many times, the same academic constructs must be measured in assessments for all students. While it may seem obvious, a science test should evaluate science, not the literacy proficiency of students.

Whether or not the constructs might be altered in an accommodated assessment (that is, when students take a test using one or more presentation, administration or response accommodations) is sometimes more difficult to discern than one might expect. Psychometrics has a long empirical history which demonstrates that, on average, changing anything in the assessment, even the placement of items, affects how students respond. By extension, it affects the constructs that are being measured. While this is true, given large heterogeneous student samples, the findings from this broad type of sample do not speak to the needs of subgroups of students who, because of accessibility barriers, cannot provide accurate information about their achievement levels under traditional testing circumstances. Emerging research in psychometrics focuses on understanding more precisely what tests and items actually measure, and then understanding what conditions can be used to measure those constructs — while simultaneously meeting the needs of all students.

Framework for Selecting Appropriate Options: Matching Student Needs and Accommodation Strengths

The need for flexibility is particularly intense when students are new to this country and/or have little proficiency in English. Flexible testing conditions remain important and relevant as long as literacy and/or cultural issues serve as a barrier to the accurate assessment of the students, whether or not these students are schooled in mainstream classrooms or formally identified as LEP.

Limited English Proficient students should receive test administration and data collection accommodations as long as these accommodations prove useful in producing more accurate information than would be the case if the students took the assessments in the traditional manner. While no legislation exists to guide how this might be determined for English language learners, the federal Individuals with Disabilities Act amendments of 1997, Section 504 of the Rehabilitation Act of 1973, and Title II of the Americans with Disabilities Act of 1990 contain requirements, guidelines and procedures for determining accommodations that must be provided to students with disabilities. Essentially, these decisions are made on a student-by-student basis. Those making the decisions include specialists (in this case, specialists in the education of LEP students), parents, the student's teacher(s), and the student. They are guided by experience, measurement constraints, and parameters dictated by district, state, and/or federal policies.

The decisions about when to use accommodations for LEP students and which ones to select depend on several factors, of which six are central. These are:

- The student's level of proficiency in English
- The student's literacy in his or her home language
- The language of instruction
- The amount of schooling the student received in his or her home country
- Cultural issues
- Accommodations that are used in the classroom as part of instruction

Below we discuss 12 promising accommodations for test administration and response. A combination of these accommodations may be the most appropriate for a particular LEP student. Under each accommodation, we briefly outline recommendations for appropriate use. The 12 accommodations are:

Administration Accommodations

- Primary language assessments
- Side by side assessments in L1 and L2
- Use of L1 or L2 dictionaries and glossaries
- Oral administration of directions in L1 or L2
- Oral administration of the assessment in L1 or L2
- Extended time
- Additional breaks
- Modifications to the test setting

Response Accommodations

- Responding without writing
- Written response in L1
- Oral response in L1 or L2
- Using computers

Administration Accommodations

Primary Language Assessments

This accommodation focuses on forms in languages other than English; these accommodations are used to measure the same set of content standards and the same constructs within those standards, as the mainstream assessment in English.

Certainly, primary language forms are relevant particularly in subject areas other than language arts. However, they should not be dismissed in language arts if the skills and knowledge are referencing language art development per se, and not defining it within the context of the English language.

This option is not viewed universally as an accommodation. In fact, if the primary language test(s) or form(s) are not built to be parallel to the English form(s), it is not an accommodation but a different assessment — most likely measuring different construct domains. In general, using two different tests presents significant comparability problems and should be avoided if results from the two tests are in any way going to be used together or to reference the same set of standards.

Parallel forms in English or in one or more languages other than English can be:

- Built from the ground up along with the forms in English
- Translated properly from the English
- Translated in a manner in which the integrity of the content and items remains the same, but culturally unfamiliar words or situations from the English version are replaced with parallel portions to better suit the culture of those who take the primary language assessment.

In each situation, the technical quality of the non-English forms must be consistent with that of the assessment in English. Unfortunately, it is not difficult to find poorly done translations in use that may also have not undergone proper piloting and validating with the range of students expected to use them.

Test developers should use the same plain language techniques when building primary language as well as English forms.

Recommendations for Appropriate Use

Educators need to consider two primary factors when contemplating primary language assessments: the degree of literacy in the first language, and the language of instruction.

1. Paper-and-pencil assessments in the home language are only useful when students are literate in their home language. This cautionary note extends to assessments which are multiple choice as well as those which require extended written responses. While many LEP students are orally proficient, at least conversationally, in their home language, we should not assume they will be literate in their home language unless they have had steady, consistent, and in-depth instruction in these specific skills.

Typically, primary language forms require a high degree of academic literacy in the student's home language or (L1). If any kind of readability indicators are used at all to guide the L1 literacy levels in these forms, they are usually of the same type used to evaluate English readability in the mainstream assessments, which are not very effective. At best, the readability expectations in most primary language forms is expected to be one or two grade levels below the test's target. For a tenth-grade high school student, this means s/he must have the L1 literacy sophistication of an eighth- or ninth-grade student. Administering a primary language assessment to high schoolers is inappropriate unless these students have received advanced literacy instruction in their home language, either in the U.S. or in their home country. This also should be a consideration for LEP students at other grades.

2. Test developers should use the same plain language techniques when building primary language as well as English forms. This opens up primary language assessments to a broader range of LEP students. However, it is still important to assess whether these students possess proper levels of literacy in their home language before a "plain-languaged" primary assessment is administered.

3. Assessments for LEP students are generally the most effective when the subject-matter knowledge and skills under assessment have been taught to the student in that language, either in the U.S. or abroad. Clearly, many of the constructs have academic language and skill components which the students have learned, and often only recognize, in the language of instruction.

4. For the newly arrived student it may be most appropriate to administer the test in L1 orally, even though the language of instruction is English. The student who experiences mixed instruction may benefit from side-by-side assessment.

The factors of language, instruction, and time in primary language and English-speaking schools present real dilemmas for LEP students. This is particularly true when students are newly arrived in the U.S. and have severe literacy limitations in English, but understand more in their English-instructed classrooms than they can read or express on a test. It also is an issue for students who have had some instruction in their primary language and some in English, perhaps first in their home country and then in U.S. schools.

Side-by-Side Assessments in L1 and L2

Side-by-side assessments are paper-and-pencil tests where the written portion, and possibly the entire test, is duplicated in a primary language (L1) and English (L2). These are usually formatted so that the same text, in L1 and L2, appears on either the left or right side of an open, double-page test booklet. It is also possible to replicate this format on a computer or use a computer to recall one form or another as needed.

This accommodation allows students to check their understanding of the item requirements and/or response choices (in the case of multiple choice) in the language which is less dominant, and/or in the language which is not being used for instruction. Preliminary research suggests that students primarily use one text, shifting to the other when they are unsure of a word, phrase, or situation.

Recommendations for Appropriate Use

1. This accommodation appears to be useful when students are taught in English, are literate in their home language, and have a degree of literacy in English. In this case, the English text is often the main text for the student. It also might be useful for those students who have received instruction in both their primary language and English, for instance when they come to this country in fourth grade and have been schooled in their home country for years prior to their arrival in the U.S. In the latter case, either the primary language or the English version may be the predominate text. In the case of side-by-side assessments with constructed response items, students may respond in English, L1, or a combination.

2. This accommodation, like many discussed here, is effective when students have had a chance to experience and use it in their classrooms before the test is

administered. This should include ongoing experience as part of classroom learning. In the case of this specific accommodation, test-taking preparation using this format is useful. Students must be accustomed to keeping track of their place as they move back and forth between texts, and they must be comfortable with the volume of text.

While this option has been provided recently in some large-scale assessments and has not been particularly popular, this may be because the LEP students taking the exam were screened beforehand to possess a high degree of literacy in English. They may not have needed the accommodation, were shy about requesting it, had not experienced it as an accommodation in their classrooms on an ongoing basis, or had sufficient test preparation using this approach.

3. The following conditions should influence the decision not to use this accommodation:

- Students are confused by duplicated information or affected by too much stimulation, particularly written stimulation
- Students have not been given proper test preparation for this accommodation
- Students have been given proper test preparation, but remain uncomfortable in a testing situation unless they can read carefully all text which has been placed in front of them. In this case, fatigue, anxiety, and/or boredom could offset the potential benefits of the option.

As always, the quality of the publisher-provided materials needs to meet the types of recommendations proposed throughout this document, (e.g., the evidence of plain language editing and proper piloting).

Use of L1 or L2 Dictionaries and Glossaries

This accommodation could include the use of L1 and/or L2 dictionaries and glossaries as classroom resources during the administration of the assessments. These L1 and/or L2 dictionaries or glossaries also might be built and provided by the publishers to the students as part of the assessment.

The advantage related to the use of these materials is that these are resources students usually encounter as part of instruction. However, material use is not consistent across schools, and classroom materials may include some information the test intends to evaluate. Thoughtful foresight is essential; educators need to anticipate which words, phrases, or situations need to be included in the publisher-provided resources.

Whether the resources should be written in English, the primary language, or in a combination of the two languages is another consideration. As always, the quality of the publisher-provided materials needs to meet the types of recommendations proposed throughout this document, (e.g., the evidence of plain language editing and proper piloting).

Recommendations for Appropriate Use

Use of literacy-based resources such as dictionaries and glossaries have been found to be useful for some LEP students.

1. Their success and usefulness depends largely on whether LEP students can effectively access them — that requires familiarity, "plain languaging," and/or student literacy skills in English and/or their primary language. They are not as bulky, time-consuming, and overwhelming as side-by-side assessments, but LEP students are a diverse subpopulation and predicting where their literacy or cultural barriers may occur is a difficult, if not impossible task.

2. **It may make the most sense to provide a basic, publisher set of literacy-based resources, by content area, and then generate specific guidelines by content area for additional types of classroom resources which would also be appropriate.**

3. **It is sensible to provide materials by content area, but these materials should not provide the answers when a test, or specific items, measures literacy understanding.** Examples could include not using a dictionary during a reading test where vocabulary is evaluated, or allowing a basic dictionary but limiting the use of a mathematics glossary during a portion of the mathematics assessment when certain academic words are being tested.

Oral Administration of Directions in L1 or L2

It is common to provide directions orally and in English for all students. It is also somewhat common to paraphrase directions in English to tailor them to individual classrooms, and to reread directions to individual students throughout the test who ask for assistance. What is less common, but not particularly controversial, is the oral administration of directions in the primary language(s) of students, including paraphrasing and rereading test modifications.

Recommendations for Appropriate Use

1. **Administering directions orally in L1 is useful for students who are fluent in their home language and literate in English or their primary language (if the test is in the primary language), and perhaps nervous about the test expectations.** However, it is questionable how much impact this accommodation will have if students cannot read the rest of the assessment.

2. **Adapting directions, including paraphrasing, is recommended when security of the test is not threatened.**

Oral Administration of the Assessment in L1 or L2

Oral administration of the test items themselves includes "live" administration from a teacher or trained personnel, a videotape, or an audiotape of someone reading the assessment aloud.

Recommendations for Appropriate Use

1. **Oral administration of the assessment is sensible when literacy in the language of instruction is an issue — when the assessment is not measuring literacy skills.**

2. **The biggest drawbacks to "live" oral administration include:**

- Unintentional cueing by readers
- Logistical modifications in administration at the school site which need to be implemented.

Not every student wants or benefits from the assessment read aloud in L1 or L2. Research has shown that, even when educators are trained and observers are present, "live reading" of the test in classrooms often unintentionally includes cueing. Cueing might include voice, rate of reading, or body language changes which, in some way, provide more information about what the responses should or should not be than is apparent from the printed words.

3. Effective ways to alleviate the cueing effects are being empirically validated. This includes the use of cassettes, videos, or computers with sound.

The logistical issues that surround this accommodation are substantial. Although the use of the electronic equipment usually includes earphones, the equipment must be purchased and students must be cycled systematically through the use of limited quantities of the hardware. School personnel are familiar with using all available staff during testing, and dealing with special needs students compounds the challenges. However, priorities need to be set to ease resource and staffing problems. Creative "win-win" solutions have been negotiated at many sites that allow sharing of resources and staffing (even across sites within districts), as well as the use of volunteers and volunteer resources from the community.

Conclusive research must be completed to understand when and how to use extended time in large-scale assessments, and how to distinguish productive from non-productive time.

4. A preliminary study (Kopriva and Lowrey, 1994) found that students tended to prefer oral administration in their primary language (Spanish, in this case) when they were new to the U.S., were not literate in their home language, and had little oral or literacy proficiency in English. This appeared to be true whether they were taught in English or their primary language.

5. In this study (Kopriva and Lowrey, 1994), the oral administration of the assessment in English was often preferred by those who have been taught mostly in English in U.S. schools for a period of time, had achieved a level of conversational oral proficiency in English, but were not yet very literate in English (particularly given the high level of reading literacy expected in a test).

Many who preferred the oral administration in English said they tried to follow along in their booklets.

The first group of students in the study represented those who had little English proficiency, while the second group represented those who soon might be in transition to mainstream classrooms. These particular subgroups may be the students who benefit most from oral administration. Most of these students typically would be waived from mainstream assessments, and it may be this accommodation (possibly in combination with oral response) which goes the furthest in allowing agencies to include all LEP students who are in academic programs. More work needs to be done to determine if this pattern holds for LEP students from various cultures and backgrounds, when this accommodation is no longer needed by English language learners, and the role of the text booklets.

Extended Time

Extended time refers to giving students a longer time period to complete the assessment than is normally allotted.

In order to access item requirements of an English test and assemble responses, LEP students face the increased work of decoding the language, in addition to reading the items and text for subject-matter intent. It is time-intensive to wade through a language in which a student is not proficient. In addition, one estimate reveals that parallel Spanish assessment texts (which make up the bulk of primary language assessments) are approximately 33 percent longer than their English counterparts (Rivera, 1997). As a result, mental and physical fatigue are confounding issues that further slow the process; confusion, frustration, anxiety, hopelessness, and fear compound the problem.

Recommendations for Appropriate Use

1. Given the heavier linguistic load in L1 or L2, fatigue, and frustration, test-taking is usually more time-consuming for LEP students (if they don't give up) and extended time makes sense.

Conclusive research must be completed to understand when and how to use extended time in large-scale assessments, and how to distinguish productive from non-productive time. Much of the preliminary work was confounded by research designs that eased time restraints while at the same time additional kinds of items were introduced in large-scale assessments. Students need to have classroom familiarity with a range of item types before they will be able to use additional time effectively to give us accurate information about what they know. Early work has also been confounded by long-standing patterns related to LEP students and test frustration — frequently, they find it easier to give up entirely. We should not underestimate the importance of sending messages to these students which communicate we care about obtaining accurate information from them, and that we are willing to meet them "half-way" by allowing the types of accommodations described here, including extended time.

Hafner (1999) has begun researching the effect of extended time for LEP students. However, given the lack of systematic information about extended time in the research literature to date, we do not know who benefits most or how much time is enough. Teachers tell us language minority students benefit for years from extra time, until they are fully proficient in the demands of academic English. The literature that addresses extended time for students with disabilities suggests this accommodation is central. Clearly, work needs to be completed for LEP students which unravels important issues related to test-taking behaviors and provides guidance. However, when in doubt, the best practice may be to allow for extended time.

2. The same construct should be measured whether or not this accommodation is used. Sometimes the issue of time affects what is being measured; this must be considered. For instance, writing assessments frequently ask students to complete a rough draft, and evaluations of the responses are based on expectations that what is being reviewed is a first or rough draft. Extended time may allow some students to produce a more polished document, while others still would be completing their first attempt. Implications such as these need to be weighed when extended time is considered.

Additional Breaks

This accommodation typically is associated with the same issues that encourage the use of the extended time accommodation. That is, the time and fatigue associated with the physical, mental, and emotional demands of test-taking often require that LEP students receive breaks more often.

Recommendations for Appropriate Use

1. Because of the additional demands placed on LEP students in large-scale testing situations, additional breaks appear to be a reasonable accommodation. Of course, this is particularly true when the students are given extended time.

For a long time, teachers have been concerned about ensuring that the LEP population is provided enough breaks during periods of teaching, learning and evaluation, particularly when literacy or linguistic demands are high (Farr and Trumbull, 1997). However, we do not know of any empirical research on additional breaks in large-scale assessments. Similarly, there is no systematic analysis that provides information about which students do best with additional breaks in classrooms.

2. Research needs to determine:

- The number of breaks that could be optimal and how these should be structured
- Which students find them useful
- Whether the need for this accommodation is consistent with the degree of language proficiency
- If the usefulness of the accommodation varies by culture or the similarity of English and the primary language, and other considerations by background, culture, or classroom experience.

One major purpose of altering the test administration setting is to minimize distractions or the additional stimulation of other test-takers.

Modifications to the Testing Setting

Setting modifications generally include small-group or isolated carrel administration rather than whole-class administration. Sometimes setting modifications refer to who administers the test, for instance, the LEP resource teacher rather than the student's mainstream education teacher.

Recommendations for Appropriate Use

1. Because of the increased demands placed upon LEP students — linguistic, temporal, unfamiliarity with testing procedures and emotional — setting modifications may be an appropriate accommodation for some LEP students. One major purpose of altering the test administration setting is to minimize distractions or the additional stimulation provided by the presencce of other test-takers. Since LEP students confront demands beyond those addressed by native speakers, minimizing further stimulation and distractions is important for some students and may encourage them to stay on task.

Work has not been done to determine if there are any patterns of LEP students, by nationality or English proficiency, who would benefit most from this accommodation, or if most of the variance is due to personality or learning style differences overlaid with the demands of learning in a new country, language, and/or cultural setting. Best practice in teaching, and in teaching LEP students in particular, has highlighted the importance of restraining the setting for some students if we expect to collect reasonably accurate information about what they know (Resnick and Resnick, 1994; Farr and Trumbull, 1997).

2. Sometimes the setting modification of choice relies on the person adminis-

tering the assessment rather than on change in the type of setting. Allowing a resource LEP or ESL teacher, or a trained member of the students' cultural community to administer the test to some LEP students may allow the students to focus on the test rather than on the unfamiliarity of the testing procedures or the demands of the large-scale testing situation.

3. The use of other administration or response equipment, such as a cassette or a computer, may necessitate a different type of setting. In this case, the value or disadvantage of a different setting, in addition to the advantages or disadvantages of using supplementary equipment, needs to be weighed when decisions are made about the appropriate accommodations for LEP students.

Response Accommodations

The same issues that confound test administration conditions affect the problem of LEP students' responses, especially if students are expected to write their answers.

As a response to the data collection needs of students with disabilities, some large-scale testing efforts quietly have accommodated response challenges for a number of years. Even in traditional test formats, with multiple choice or other forced-choice response situations, experience has documented that some students benefit from responding in ways other than using the "bubble sheets."

The recent expansion of item types in large-scale testing, including constructed response items and performance tasks, has presented formidable challenges for English language learners. This is largely due to heavy literacy expectations in presenting the items and tasks, and in the expectations that student responses and results will be written. While these types of items are problematic for LEP students, it is very important to understand that they tap different areas of academic mastery than multiple choice or other force-choice items. These areas of mastery are found in virtually all content standards, and need to be assessed in large-scale evaluations. Therefore, they are an integral portion of tests where LEP students are involved.

Kopriva (1994), among others, has found that LEP students often *prefer* constructed response and performance types of items. This is because these students felt they were freer to explain themselves, and did not feel constrained by the choices or possible additional linguistic misunderstandings posed in forced-choice items. Rather than resist these item types, it is important to find accommodations which allow students to respond effectively to the constructed response item requirements. Four accommodations are particularly appropriate for LEP students and are discussed briefly below.

Responding Without Writing

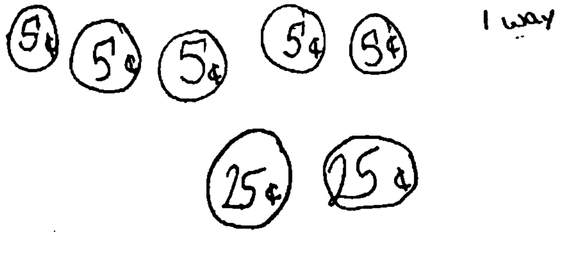
Allowing students to demonstrate what they know without writing includes both non-written responses on paper and responses not on paper. Non-written but graphic responses on paper include using pictures, algorithms, charts, or diagrams to convey what students know. Other non-written responses include oral responses and performances.

In Figure 6 on the following page, an LEP student is responding to the item largely without writing. It is especially important for those who evaluate student responses to remember that assuring the level of understanding the subject matter, in

Figure 6 • The Vending Machine

Maria wants to buy a 75-cent snack from a vending machine. The machine takes only nickels, dimes, and quarters. Maria has 7 nickels, 5 dimes, and 2 quarters.

Part 1
Show all the different ways she could pay for the snack. You may use words, diagrams, or charts.



5 dimes plose 1 more 25¢ = 75

this case elementary-level mathematics should not be confused or colored by the student's literacy limitations.

Allowing students to write their responses in their primary language has been used by several large scale assessments.

Recommendations for Appropriate Use

1. While still retaining the integrity of the construct part that is being measured in a particular item, we recommend that students be able to respond on paper-and-pencil tests in non-written forms. Many constructed response items unintentionally restrict how a student must answer the items because time and cost constraints have largely limited responses to written ones. *This attenuation has had severe consequences for the English language learner.*

If the item is not intended to measure written skills or communication of a particular subject, then allowances should be made for students to respond in a variety of ways.

The chapters which focus on rubrics and scoring discuss types of non-written paper responses in greater detail.

2. Non-written oral responses and performances are recommended, as the technology permits. Oral responses and computer simulations will be discussed below. A discussion of other types of performances are outside the scope of this *Guide,* but are generally recommended as technology can adapt them into large-scale testing situations.

3. We recommend that non-written modes of responses be given proper credit. It is important that experts on LEP issues participate in developing rubric notes, scoring materials, and appropriate scorer training, so that non-written types of responses can be scored appropriately. The chapters which focus on these phases of test development provide recommendations about evaluating the responses. This is especially important for LEP students because their knowledge of subject matter is often more sophisticated than their literacy skills. If their level of writing proficiency is tied to how their responses are scored, the scores may be lower than they should be.

4. Research needs to be conducted in tandem with experts on LEP issues that investigates how to encourage a range of types of non-written responses in paper-and-pencil tests, and how to reasonably evaluate the responses of performances. The greatest barrier to including these types of responses regularly is that test developers do not know how to build items which can handle non-written paper responses, while still requiring students to respond to the same construct. Teachers of LEP students have worked effectively under this constraint and would be a valuable resource in collecting documentation and providing guidance.

Written Responses in the Primary Language

Allowing students to write their responses in their primary language has been used by several large-scale assessments. The accommodation has been usually coupled with administering tests in the students primary language, although this is not necessary, per se. LEP experts say that written proficiency in English comes somewhat later than reading proficiency for some students, so reading in one language and responding in the other may occur.

Recommendations for Appropriate Use

1. Unless the test measures English writing skills or an item measures communication through writing in English within another content area such as science or mathematics, allowing written responses in L1 should be seriously considered.

2. The primary consideration regarding proper use is the student's L1 writing proficiency and comfort in writing in his or her native language. That is, a student must be able to communicate sufficiently in writing in the primary language for his or her level of mastery in the particular content area to be evaluated. Many students are not literate in L1, or their level of literacy is basic enough that academic constructs may be difficult to successfully communicate. Sometimes students are "rusty" — they have not used their literacy skills in their primary language very often or routinely, especially in a U.S. school context. Often, the response is actually a "combination" of L1 and English, in a variety of different forms. For instance, words from both languages might be used in sentences, English or L1 words may be "translated" into the other language with endings or other conventions familiar in that language, or words in one language may be "sounded out" using the phonetic structures from the other language.

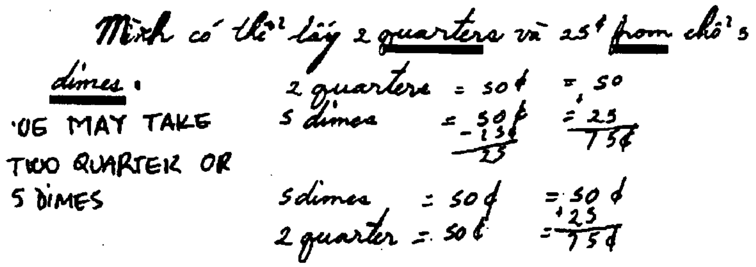
Figures 7 and 8 on the following page depict two students' responses which are a combination of their first language and English. Figure 7 is an example of "code-switching" where students alternate use of languages. Figure 8 shows students transferring what they know about the phonetics of their native language into trying to write in English.

3. Scorers should be literate in the primary language and properly trained to score these responses with the same consistency and validity as their English-speaking peers score responses in English. Since table discussions inform the scoring, it is also important that there be a table of native-language scorers for each language represented, or that the scorers be bilingual so they can participate in table discussions in English.

Figure 7 ○ The Vending Machine

Maria wants to buy a 75-cent snack from a vending machine. The machine takes only nickels, dimes, and quarters. Maria has 7 nickels, 5 dimes, and 2 quarters.

Part 1
Show all the different ways she could pay for the snack. You may use words, diagrams, or charts.



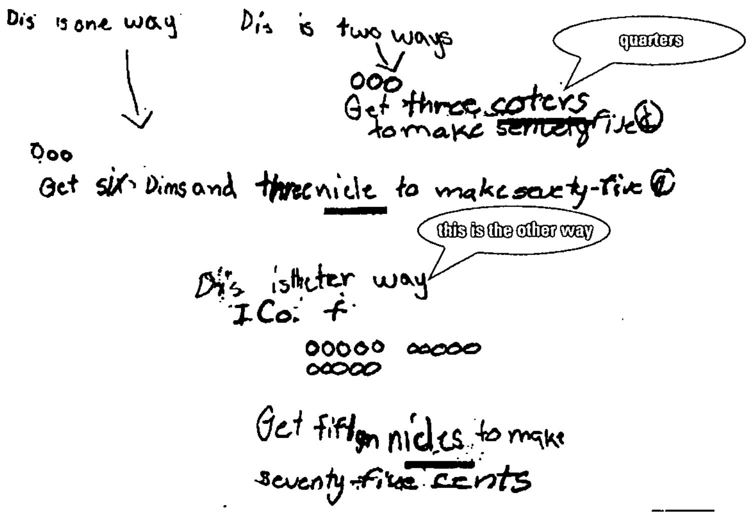
Mình có thể lấy 2 quarters và 25¢ from chỗ's dimes.

'UE MAY TAKE
TWO QUARTER OR
5 DIMES

2 quartere = 50¢ = 50
5 dimes = 50¢ − 25¢ = 25 = 25 = 75¢

5 dimes = 50¢ = 50¢
2 quarter = 50¢ + 25 = 75¢

Figure 8



Dis is one way

Dis is two ways

000
Get three coters to make ~~senety~~ five ¢

Ooo
Get six Dims and threenicle to make sevety-five ¢

Di's isther way
I Co: f

00000 00000
00000

Get fiftyn nicles to make seventy-five cents

quarters
this is the other way

Recently we have reached an understanding of the limits of the multiple choice item format in terms of what it can and cannot measure.

In many cases, responses that are a combination of English and L1 are a challenge to scorers, because they often are not scored easily by speakers of either language. The chapter on scoring discusses in more detail how to train scorers so that they can score these types of responses successfully and appropriately. While the focus of the chapter is on training English-speaking scorers to score responses from LEP students, the discussion can apply to training L1 scorers as well.

Oral Response in English or L1

Allowing or requiring students to respond orally is certainly not new and played a central part in exams favored in the U.S. before large-scale multiple choice testing became popular. In addition to the cost and time efficiency of machine-scored tests, multiple choice formats and the standardization in test administration and response conditions were initially encouraged by experts because they appeared more equitable.

Recently we have reached an understanding of the limits of the multiple choice item format (also known as forced-choice or close-ended item formats) in terms of what it can and cannot measure. As a result, many educators would like to see tests that include item formats that measure achievement in ways that are not available through multiple choice assessment. Measurement experts have developed ways of reliably scoring responses of open-ended item types, including constructed response items and performance tasks. Unfortunately, time and cost constraints have largely limited responses of these item types to written records. As stated earlier, this attenuation has had severe consequences for the LEP student.

Many linguistic and LEP experts have cautioned that effective written skills in academic content can take years to develop fully in language minority students. Until that time, this mode of communication may under-represent what English language learners know. While these students also are still learning how to verbally articulate academic content, they are often better able to effectively express themselves orally. This mode should not hamper the standardized advantages of written responses as long as consistent practices and reliability evaluation checks in scoring are used.

Oral response may be useful in some cases where a multiple choice format is being used, if literacy or cultural confusion undermine the benefits of this type of response. Kopriva and Lowrey (1994) found that correlations between written responses to open-ended and multiple choice items were consistently different for LEP students than non-LEP students. While this could be attributed solely to the literacy demands of the constructed response items, it is likely that the responses to the multiple choice items were affected as well. Investigations need to occur to determine when and if oral response makes sense for forced-choice items, particularly if other accommodations, such as oral administration of the test, are used.

Recommendations for Appropriate Use

Both student accessibility needs and resource practicality need to be considered if this accommodation is selected.

1. Oral response to open-ended items and tasks makes sense when written response skills are not what is being measured. Essentially, LEP students who lack sufficient English proficiency to allow them to communicate effectively in written form should be able to demonstrate their knowledge orally or through other response accommodations. Schools must work with students to find a mode that both effectively meets the students needs and is also reasonable for the institution.

2. One way to lessen the practical burden of collecting and scoring oral responses is to ensure that there are ample open-ended items that allow other modes of communication — besides writing — in which to demonstrate mastery. This includes the use of charts, diagrams, algorithms, or computer-simulated formats.

3. Since the practical issues of collecting oral responses can be formidable, institutions need to weigh advantages and disadvantages of various procedures.

- While some agencies allow students to dictate responses to a scribe who writes their answers, interpretation inconsistencies can be a problem (the scribe may not write down exactly what the student says), and the cost of scribes is a consideration.

- Using cassettes or videos to record responses alleviates the use of a scribe, but this presents a "real time" scoring problem — most written responses which take about 15 minutes or so to write can be scored in about one minute. Scoring from cassettes or videotapes requires scorers to listen for as long as it takes the student to respond, say the same 15 minutes or so. This is an expensive use of a trained scorer.
- Expensive and less-than-reliable scoring problems may be overcome, however, by recent advances in voice recognition technology. This technology allows students to respond into a computer which "hears" the response and transcribes it into printed text. The text then can be scored with the written work of other students. This approach is also promising because low-functioning computers (such as those which often exist at school sites) can be networked with a central computer to provide the service. Therefore, the cost for the technology is minimal. The state of Delaware, under a grant from the U. S. Department of Education, is researching this option to determine whether the approach is feasible.

As hardware and software technology advances, there will be increased opportunities to use computers.

4. Space is a second logistical issue. The oral response accommodation requires each student to be separated from other students, for test security purposes. Traditional spaces, such as listening labs, certainly can be used, and there is work being done which investigates using microphones which pick up soft voices. This consideration, however, will certainly tax the resources of schools, and other creative solutions need to be identified.

Using Computers

We have already discussed the use of computers to administer tests and voice recognition software to facilitate oral responses. Other types of computer technology also may be very useful for LEP students.

As hardware and software technology advances, there will be increased opportunities to use computers. This new technology will open up item presentation and administration possibilities. It will also play a key role in broadening response formats for students. Since the literacy challenges of English language learners are central to their performances on traditional assessments, these advances will be particularly useful for them.

Recommendations for Appropriate Use

1. Simulation software can encapsulate "performances" which demonstrate mastery. One set of examples are programs that allow students to perform science experiments or other types of "hands-on" activities in content areas such as mathematics or social studies. Work done by Shavelson and Baxter (1993) investigated simulation of a "hands-on" science activity. Their findings suggested that allowing students to perform a science experiment through a computer simulation yielded more accurate information about what students knew than did their paper-and-pencil responses. In some cases the simulation was almost as beneficial as live performances in the classroom.

2. **Graphic arts technology that allows users to "build and craft" 3D solutions, incorporating movement and music as well as video media, can provide alternative response opportunities, even if the item requirements are text-based.** Simulated figures can dance and perform response solutions that broaden the sensory access of students whose strengths are not linguistic.

3. This accommodation is influenced heavily by classroom opportunities for students to use computers and software such as those programs described above. Assessment developers may build practice modules from which students can learn and practice the programs before testing. As hardware and software resources increase district-wide and statewide, it will be important to discuss how opportunities that already exist (probably unevenly) in the schools can be broadened to provide instructional experiences to the full range of classrooms and content areas. Concurrently, these discussions and should explore how the same hardware and software can be effectively used in large-scale assessment.

4. LEP educators and experts who educate other special needs populations should take an active role in encouraging the types of response avenues engendered by the use of computer technology. While some may suggest that costs prohibit building in the types of experiences discussed here, the bigger problem may be re-thinking how to utilize the technological resources which exist in the large-scale assessment of students. This is a challenge for educators and for test publishers alike. Those charged with the schooling of non-standard populations such as LEP students (as well as students with disabilities, gifted and talented students, or creative students), have the most to gain from extending the types of assessment opportunities for their students.

Chapter 7

Expanded Bias Reviews

Bias reviews serve an important function in test development because they provide the primary opportunity for representatives from special populations (including English language learners) to review and comment on testing materials. Bias reviews become problematic because their scope often has been very narrow: reviewing materials for offensive stereotyping, wording, or subject matter. Reviewers typically are asked to review vast amounts of item and related materials in a very short time frame which necessitates, at best, an evaluation that is well-meaning but superficial. To the credit of publishers and client agencies, the recommendations from these reviews often are taken seriously and incorporated into the revisions of materials. Unfortunately, even when bias reviews are included in the current test development processes, many problems remain for LEP students and other special needs students.

Are bias reviews necessary? We believe that bias reviews are very useful, for a number of reasons discussed in this chapter. We also believe that they should be expanded.

Test publishers differ in the ways in which they organize their reviews. We do not recommend any one review process, but encourage the best elements from different procedures.

Bias reviews provide an opportunity for a wider range of special needs educators and other educational stakeholders to have input beyond those LEP experts involved throughout the test development process.

Recommendations for Expanded Bias Reviews

1. Bias reviews provide an opportunity for a wider range of special needs educators and other educational stakeholders to have input beyond those LEP experts involved throughout the test development process. These individuals include:

- Educators whose LEP students are not in the primary language groups or the most prevalent geographical settings
- Educators who represent a wide range of classrooms that school LEP students, including educators from classrooms that are less conventional than the norm
- Parents with a particular interest in equitable access to large-scale tests, such as the parents of LEP students or other special needs students
- Community representatives with a vested interest in equitable testing for all students.

2. In addition to reviewing test items for offensive content, bias reviews provide an excellent opportunity for test developers and publishers to receive guidance about test formats, wording, rubrics, non-text item accessibility, and administration and response conditions. Table 4 outlines questions for panelists to consider. Areas for consideration in bias reviews include:

- Test format issues, including page layout, item spacing, font size, visuals, and response procedures
- Test item wording, including word choices, sentence structures, and paragraph structures in test directions, test items and all contextual information
- Rubric parameters which should focus on linguistic and cultural issues

Table 4 ○ Bias Reviews: Expanded Questions for the Panel To Consider

Test Formats

- What type of test formats would be best for your student population and why?
- What format problems should the assessment avoid if possible and why?

Wording in Items and Directions

- What general advice about the wording in the items and directions can you give developers to ensure that your student population will understand what they are being asked to do?
- What is your advice about how to structure and incorporate any additional contextual information which students will receive as part of the items, including pictures, diagrams, and text?
- Given a subset of items and directions, what changes would you make to ensure that your student population will understand what is required?
- What wording problems should the assessment avoid if possible and why?

Rubric Parameters

- What special issues associated with your student population need to be considered when the constructed response items are scored?
- Given all or a subset of the rubrics which will be used in the assessment, what changes would you make to ensure that the responses from your student population will be scored correctly?
- What issues about how items are scored should the assessment avoid if possible and why?

Non-Text Item Accessibility

- Besides wording and scoring considerations, are the requirements of the items you have reviewed accessible for students from your population? That is, do the text and non-text supports allow students to understand what they are being asked to do or know, and give them what they need to process their response?
- Generally, what guidance can you give test developers which would help ensure students from your population are able to effectively access each item?
- What item elements constrain accessibility and should be avoided if possible? Why?

- Non-text accessibility, including all text and non-text supports such as pictures, diagrams, performance activities, and access to specified tools and resources
- Conditions of administration and response, including advice from participants on strategies particularly appropriate for their populations.

Many of these are commonplace issues that educators of LEP students deal with routinely in classrooms; therefore, their insights are valuable. The advice from parents and other public stakeholders about these additional issues should be a welcome reality check, and is usually surprisingly informative.

Figures 9, 10, and 11 provide an example of how a word in one item ("fewest") can generate two meanings. In Figures 9 and 10, the students have interpreted "fewest" as the smallest value. In Figure 11, the student interpreted "fewest" to be smallest in size. To the extent that the word can be modified, for instance if it is not a mathematical term that students are supposed to know, bias review participants can help identify the problem and provide a useful solution.

3. Participants should be briefed on all steps taken to ensure accessibility throughout the development and implementation processes. These steps need to be explained in simple, straightforward language with examples if possible. This information should be included to provide an overview that participants can draw on as they are providing guidance.

4. Publishers need to allocate sufficient time for a thorough, item-by-item

Figure 9

Which of your ways uses the **fewest** number of coins. Explain why this is true.

the nickels are the fewest of the coins.

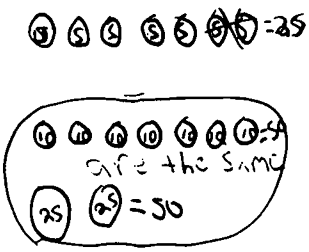


Figure 10

Which of your ways uses the **fewest** number of coins. Explain why this is true.

The Littlies nmmber is a pennies
its oniey one cent.

pennies

only

one

Figure 11

Which of your ways uses the **fewest** number of coins. Explain why this is true.

the dame Because it is the Sonkst
money you cen get.

smallest

review of materials. While it is unrealistic to assume that the panel can review all items and their associated rubrics and contextual materials — unless the test is a one-form assessment given to relatively few grades — a random subset of items/rubrics and materials should be selected for review in all subjects tested. This includes all visuals, non-text contextual materials (e.g., videos or computer-simulated materials), and examples of tools and resources the students will be able to access. In the case of a content area such as language arts, all items and associated materials might be reviewed. Specific selected items or other materials flagged as problematic by item writers might be included as well. Technical manuals should explain how items, rubrics, and materials were selected for review. This review should help stimulate focused advice from reviewers to publishers.

5. Participants should be able to review mock-ups of some of the assessments, to get an idea of layout and presentation, and assess whether these are sufficient.

Chapter 8

Scoring Constructed Response Items of LEP Students

With the advent of open-ended constructed response items with multiple possible scores, scoring had to be completed "by hand." In large-scale situations, scoring was conducted in large-volume settings where anywhere from a table of scorers to rooms of scorers evaluated piles of item responses for several hours a day. Scorers were trained to ensure score accuracy and consistency over items and scorers, and validity and reliability measures were devised to ensure technical rigor. These advances allowed a wider range of evaluation of student mastery.

As procedures and materials have become refined and standardized, concern has grown about the accuracy of LEP scores (Kopriva, 1994; Kopriva and Lowrey, 1994; Daro, 1995; Sáez, 1993). In particular, there is growing concern that answers of language minority students sometimes were scored inappropriately, due not only to the high volume and rapid manner of scoring these responses in large-scale situations, but also due to the fact that approximately 99 percent of the scorers are monolingual English-only scorers. Some suggested that sympathetic scorers tended to err by scoring LEP papers closer to the mean than they should, while others have suggested that responses heavily laden with linguistic and/or cultural issues tended to be scored lower than was correct. In either case, there is a broadly based national sentiment that attention needs to be paid to increasing the confidence we place in the scores of this population.

Test publishers differ in the ways in which they organize their reviews.

The Council of Chief State School Officers and the U.S. Department of Education (Kopriva, Sáez, and Lara, 1999) conducted a special scoring study in 1997. This study evaluated a brief training model (approximately 1 and 1/2 hours) that integrated information about LEP response-style issues into the regular training scorers receive as part of large-volume scoring of NAEP item responses. Results suggested that this LEP training succeeded in its intent: improving the accuracy of scoring done by monolingual English scorers. It also appeared to improve the accuracy of scoring responses from other "nonstandard" populations, including learning disabled students and some racial minority students.

The study yielded other helpful results. The participants encouraged heightened participation of LEP educators as scorers and in developing rubric and scorer materials, and the purposeful inclusion of LEP student work in the student response samples. In addition to training scorers, participants unanimously recommended that table leaders (those who supervise a table of approximately 10-15 scorers) receive a version of the LEP training prior to scorer training. They made this recommendation because table leaders serve important arbitration and decision-making functions during scoring.

LEP Accessibility Scorer Training Materials

Under the guidance of LEP educators from across the United States, the Council of Chief State School Officers has developed mathematics and science scoring manuals.

The manuals, *A Guide to Scoring LEP Student Responses to Open-Ended Science Items* (1999) and *A Guide to Scoring LEP Student Responses to Open-Ended Mathematics Items* (1998), outlined the features these educators determined to be the most salient for native English-speaking scorers in order to score LEP student papers more effectively. While they are designed to be used to train scorers in large-scale district or statewide achievement assessments, this information could also be used by teachers to aid them in accurately evaluating their LEP students' classroom work. The manuals are arranged so that they can and should be adapted to local training conditions, and to the needs of various large-scale and classroom endeavors.

The manuals are divided into two sections. The first section on focused cultural and linguistic issues affecting how students respond. The second section summarizes issues associated with the impact of culture and language on test development and on the conditions around which tests are administered. Since we have already provided an extensive discussion of the points made in the second section of the manuals, we focus here only on scoring information.

Clearly, several issues have a profound effect on the format and construction of written English responses by students who are English language learners. There are linguistic issues, cultural influences, stylistic preferences, and issues related to language acquisition development. Without training, the response or parts of the response are sometimes unintelligible, and often confusing, to native English speakers. These reactions are exacerbated in high-volume scoring situations.

It is common for code-switching, transposition of words, spelling, and phonetics from a students' native language to have a substantial effect on their writing.

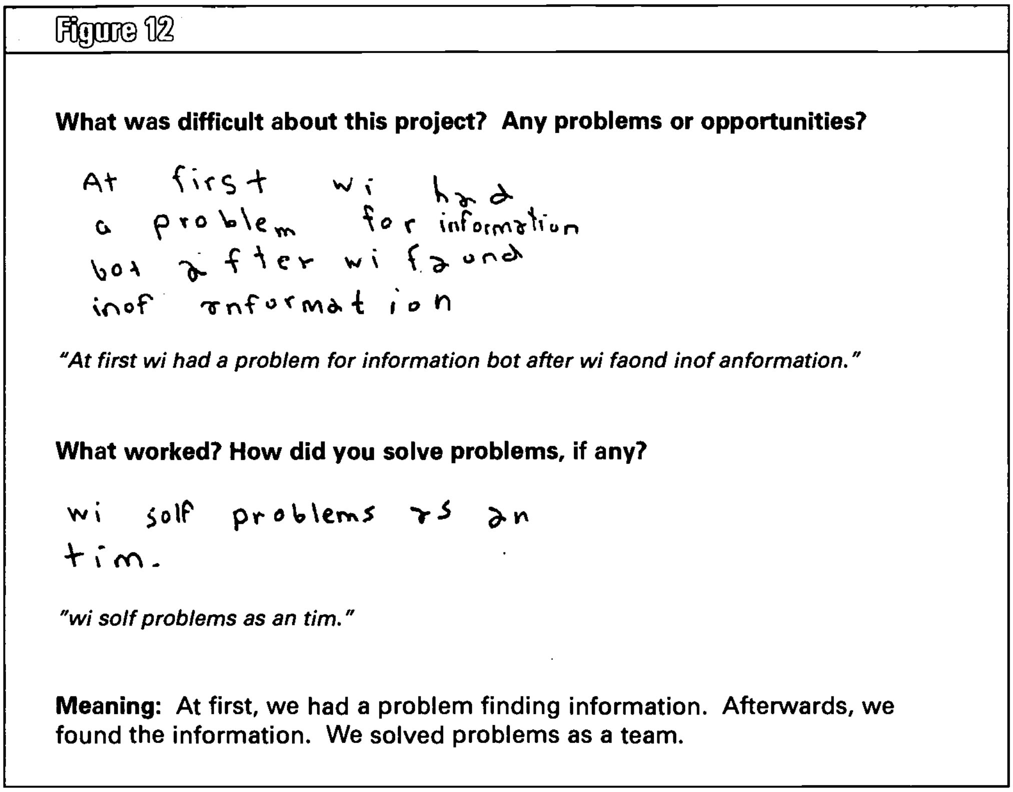
Literacy Issues

These issues include native language influences, English phonetic influences, word mergers, and omissions. It is common for code-switching, transposition of words, spelling, and phonetics from a students' native language to have a substantial effect on their writing. This occurs because LEP students often mix native-language influences with English conventions as they learn to differentiate English systems from native-language words or parts of words, native-language phonetic spelling, or spelling mechanics.

Like native English speakers, they also invent spelling of English words from what they know about phonetics in the English language. While this is developmentally appropriate, given LEP students' years of experience with English, it is often not grade-appropriate. English phonetic sections of the responses can be misunderstood by native English scorers, particularly if they are combined with other issues we have discussed in this *Guide*, or if scorers are not familiar with pervasive inventive spelling (for example, if the scorers have not taught at the elementary school level).

Word reductions, mergers (the condensing of words into one mega-word), and omissions of tense markers, articles, plurals, prepositions, or other words also confuse the writing of LEP students. The merging of words in phrases or sentences into one word is common to both native language speakers and ELLs. However, Limited English Proficient students sometimes mix first and second languages or phonetic systems in ways that are unfamiliar to native language speakers ("ghuadayamean" for "what do you mean"). Like English phonetic spelling, this practice also appears in grades developmentally appropriate given the years of English acquisition. When most native speakers become more sophisticated in their writing, the written expression of word mergers is less prevalent.

Figure 12



What was difficult about this project? Any problems or opportunities?

At first wi had
a problem for information
bot after wi faond
inof anformation

"At first wi had a problem for information bot after wi faond inof anformation."

What worked? How did you solve problems, if any?

wi solf problems as an
tim.

"wi solf problems as an tim."

Meaning: At first, we had a problem finding information. Afterwards, we found the information. We solved problems as a team.

Omissions can be attributed to many sources. There are two common reasons English language learners omit letters or words: (1) There is no equivalent convention in the students' native language, or (2) The students lack understanding of English conventions. It is very helpful for scorers to possess some rudimentary knowledge of native language conventions.

Figure 12 illustrates an LEP student's response where word reductions and omissions have occurred. In high-volume scoring this type of response can easily be underscored if it cannot be read easily.

Cultural Influences

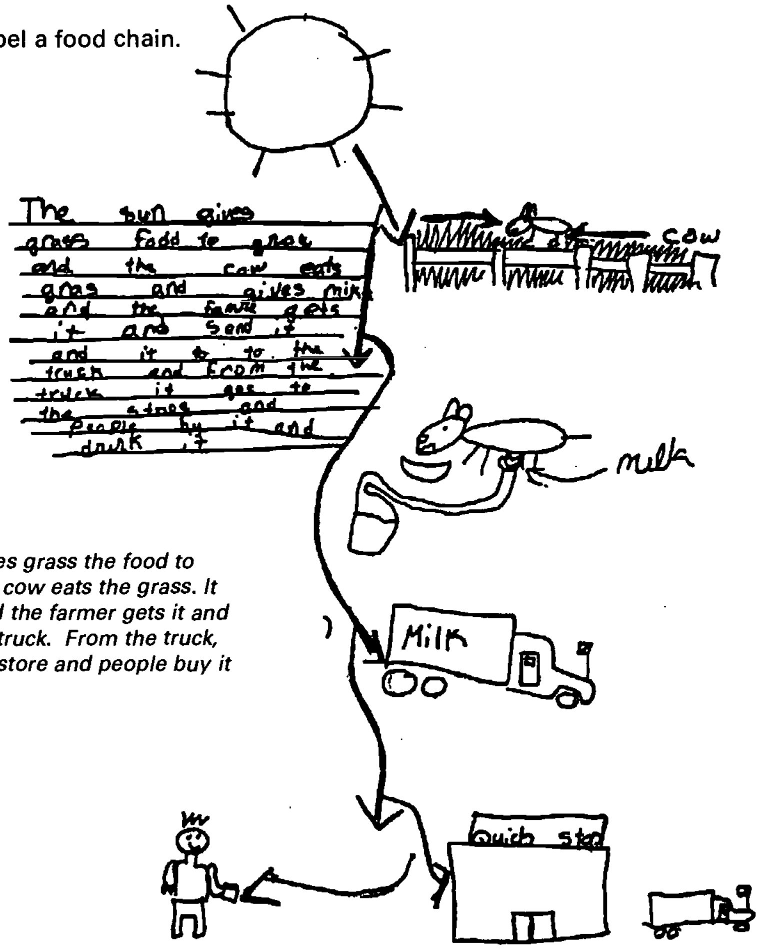
Cultural influences include:

- Mathematics symbols
- Character, marking, and accent differences
- Different scientific designations
- Different monetary and measurement systems
- Numeric differences, and
- Variations in writing and mathematics/science conventions.

The effects of cultural influences are pertinent in item and rubric development, and useful to review when responses are scored. We have noted already how confusing it can be for scorers when periods are used for commas and different words used for numbers. In addition, in some countries, the native language is written from right to

Figure 13

Draw and label a food chain.



("The sun gives grass the food to grow, and the cow eats the grass. It gives milk and the farmer gets it and send it to the truck. From the truck, it goes to the store and people buy it and drink it.")

We recommend that LEP educators be included during the development and assembly of rubric notes and general scoring materials.

left, or bottom to top. It is not uncommon to see responses from ELL students which are written in this way. In some countries, mathematics procedures are written differently. As one example, the procedure of long division often is written from the item up, rather than placing the numbers below the item as is done in the United States.

As noted earlier in the rubric writing chapter, cultural influences can often extend to different experiences or to different value systems. In Figure 13, a Punjabi student is trying to explain a food chain, building on the example of incorporating a cow like his teacher suggested. His response has been affected by the example in an unintended way; he doesn't eat beef and so he interpreted the response differently than many of his peers.

Stylistic Preferences

It is not unusual for testing rubrics to favor a limited range of response and writing styles that mirror what is taught and valued in language arts curricula in U.S. schools (for an example, see the NCTE Standards, 1998). All students can experience problems fitting into this limited range. However, some cultures encourage linguistic approaches not favored in the United States. Excessive use of an alternative style can distort student scores, if scorers are not familiar with or value different styles, or if they cannot locate easily the points the student is trying to make. The circular response, deductive, and abbreviated reasoning styles are discussed in the scoring guides.

Issues Related to Language Acquisition Development

Social communicative competency in English occurs before academic language fluency (Cummins, 1979). Immaturity in vocabulary, sentence structure, and writing formats is common and appropriate. In this section, we identify some common issues that have an effect on mathematics or science responses; these issues are associated with language development. Immature written expression may or may not be acceptable, depending on what a specific item is supposed to measure and the content standards upon which the test is built.

Language acquisition development issues include:

- Substitution of common words for precise mathematical or scientific terms and concepts (the use of "fattest" for greatest or "smallest" for fewest)
- Confusion about the meaning of words — (left as opposed to right) vs. left (remaining), whole (all of the parts) vs. a whole number (not a fraction), and sum (read as some) vs. sum (the result of adding).
- Inappropriate use of unknown words (in a problem about buying snacks from a vending machine, a student may discuss buying vending machines rather than snacks)
- Novice sentence and paragraph structures (these may suggest incorrectly that the student does not understand the answer to a complex mathematics problem)
- Non-linguistic formats to communicate answers (numbers, charts, pictures, and graphs).

Recommendations for Scoring Responses of LEP Students

1. We recommend that educators of LEP students be included during the development and assembly of rubric notes and general scoring materials.

2. It is important to include the student work of LEP and other special needs students in any and all regular scorer training packets, including training and calibration sets. Integrating these types of papers, with their special challenges, into training of scorers allows scorers to learn how to handle non-standard responses. It also sends a strong message that special needs students can and do achieve at all levels, and that their work, with their characteristic issues, is important to consider in the training and scoring process.

3. Integrating a brief LEP training into regular scoring is recommended for increasing the ability of scorers, most of whom are monolingual English only speakers, to confidently evaluate LEP responses. This training also appears to be useful for scoring responses from other "nonstandard" populations, including learning disabled students and some racial minority students.

Features of the training are discussed in detail in the CCSSO scoring guides.

They include literacy features, cultural influences, stylistic preferences, and language acquisition issues.

4. Scoring table leaders should be trained as well, because they serve important arbitration and decisionmaking functions during the scoring activity. In the case of computer scoring, the division managers serve this function and should be trained accordingly.

5. LEP educators should be included with the English speaking scorers at the scoring tables.

Chapter 9

Demonstrating the Technical Merit of Assessments

This chapter primarily focuses on evaluations of test validity which should be helpful in ensuring the validity of items and tests for LEP students, and which are designed to supplement existing procedures. A discussion about additional technical considerations related to accessibility can be found in Appendix A.

How best might assessment developers or test reviewers evaluate the validity of large-scale assessments for English language learners? Most of the validity procedures popular today in test development are not sufficient when determining whether an assessment can collect accurate information about LEP students' knowledge. In part this is because most analyses are not disaggregated by some type of LEP status (e.g., level of English proficiency has been found to be useful). Many systematic validity problems have been masked because they are "canceled out" as minority data within aggregated, heterogeneous validity samples.

Most validity designs have also relied too heavily on large scale quantitative analyses, and have not focused enough on more thorough investigations. Large quantitative studies are useful if the right questions are asked, and the right variables included in the analyses. Large-scale assessment procedures do not typically take the time to isolate validity design problems for LEP students, and analyze their impact in a systematic way.

We recommend a combination of high-quality, systematic analyses that use statistical results, student work, and student/teacher feedback. The procedures outlined here should be able to inform and improve traditional procedures, but only if validity is evaluated at each of the points described below. That is, a validity design suitable for ensuring accurate measurement for LEP students needs to document systematically validity of items, validity of rubrics, and validity of forms/tests. One set of procedures for identifying and evaluating validity at these points is discussed in this chapter. Not surprisingly, implementing the procedures outlined here should provide important validity information for all students, not just English language learners.

We recommend a combination of high-quality, systematic analyses that use statistical results, student work, and student/teacher feedback.

The evaluation methods are discussed in terms of how they would be used within the development of assessments. When off-the-shelf tests are being considered, the procedures can serve as a framework for guiding requests for additional validity information from test publishers. In addition to expecting all current validity information to be disaggregated by LEP status, for instance the level of English proficiency, agencies should be able to expect that publishers collect and document the type of evidence explained here.

General Recommendations in Determining the Technical Merit of Assessments

This chapter will focus on determining validity for LEP students. Current validity designs do not do an adequate job of ensuring validity for this population.

1. All current types of validity data need to be disaggregated by LEP status,

for instance the level of English proficiency. Aggregated data masks important validity problems for limited English proficient students.

2. Validity evidence needs to include a combination of high-quality, systematic analyses that use statistical results, student work, and student/teacher feedback. Current procedures would need to be supplemented by the types of analyses described in this chapter.

3. Validity designs suitable for ensuring accurate measurement for LEP students need to include systematic documentation at three levels: evidence of validity of items, validity of rubrics, and validity of forms/tests. The procedures outlined here should be able to inform and improve traditional procedures, but only if validity is evaluated at each of the points described below.

4. When off-the-shelf tests are being considered, the procedures discussed in this chapter serve as a framework for guiding requests for additional validity information from test publishers. In addition to expecting all current validity information to be disaggregated by LEP status, agencies should be able to expect that publishers collect and document the type of evidence explained here.

Item evaluation composes the most important portion of the test development process that actually focuses on improving test accuracy for LEP students..

Item Evaluation: A Review of Student Information and Teacher Input

Item evaluation composes the most important portion of the test development process that actually focuses on improving test accuracy for LEP students. This evaluation includes collecting and evaluating LEP student responses and student feedback. It also includes collecting input from LEP educator specialists and mainstream teachers of English language learners. Statistical data should also be included in the evaluation once the data have been collected.

Most test publishers evaluate items twice during test development. Similarly, we recommend (1) an analysis after piloting items rather early in the test development process, and (2) analyses after the large-scale field trial.

During the first review, many test publishers and testing programs conduct a small round of formal pilot-testing and/or ask item writers or reviewers to try out items on some students. The purpose of these pilots and item tryouts is to get a sense of how students respond to specific items. This review of student work helps to identify problems in item wording or presentation. In addition, test publishers sometimes collect editing suggestions from students and teachers in the field. This step provides extremely useful information about what the items measure.

However, evaluation procedures and time constraints usually do not allow a thoughtful and systematic review of the responses at this step. In many instances, special populations, such as LEP students, are not oversampled to obtain enough information about what they think the items are measuring. Even if oversampling occurs, it is not customary to separate their responses and evaluate those responses for systematic problems.

By contrast, in large-scale field tests, items are scored systematically and statistical information is computed. Other than DIF analyses, the statistical data are typically not disaggregated by group. Student work also is not evaluated at this stage. Further, the statistical data are generally the only formal indicators of how items are received by students, and provide the only information about items that are recorded with the items in data banks for use on future tests or with future clients. In fact, some of these data may be compiled as part of the large-scale field test, while some may not be completed until after the subsequent round of "real" testing.

Lack of disaggregation of findings and insensitive predictor variables in these analyses produces evidence of validity which may not be accurate for limited English proficient students. While differential item functioning techniques (DIF) flag items with different score distributions across groups, these methods do not detect certain types of bias which are particularly problematic for LEP students (this issue is discussed in more detail in Appendix A).

Therefore, current procedures leave too many items unnecessarily problematic for LEP students (and probably other students as well). In a recent study with NAEP science items (Kopriva, Saez, and Lara, 1999), one-third of the items were found to have problems unique to LEP students. Some problems were flaws in item rubrics, flaws which did not account for responses from those with very different cultural backgrounds. The rest reflected linguistic and cultural problems in the items themselves. These issues could easily have been identified and fixed during test development if a more thorough validity evaluation process had been utilized including a focus on LEP student data.

What information must be brought to the tables for evaluators to review, where should the information come from, and which procedures are particularly helpful when evaluating test items systematically for LEP students? Besides collecting quantitative data, LEP student information and teacher input should be collected and reviewed. Student information includes the student responses to both closed and open-ended items, student feedback, and a group summary from students about what items are measuring. Teacher input includes edits and reactions to specific items (as well as any feedback on forms, accommodations, and the testing process). Finally, the item statements–about what the items are intended to measure– must be present as an integral part of the evaluation process.

How should the evaluations be timed? These evaluations must be completed after the large-scale field test because, at this point, the range of responses and feedback are critical. We recommend conducting a similar systematic review after piloting or classroom tryouts as well. While the following discussions about sampling, materials to be collected, and the evaluation procedures are based on the large scale evaluation, these elements should also be present, on a smaller scale, for the piloting evaluation.

Sampling

Pilot and field tests are usually administered to a subset or sample of the total number of students from the district, state, or nation. Sometimes these trials are free-standing test administrations, and sometimes new items are piloted/field tested within one or more administrations of an existing assessment. In any case, when determining the number of LEP responses upon which to evaluate the validity of the scores, we recommend the following criteria. It is important to remember that the information gleaned from this sample can be used as many as three times for different purposes–to evaluate items and forms, to evaluate rubrics, and to evaluate accommodations.

Recommendations for Sampling

1. Include LEP students at all levels of English-language proficiency in the sample who will be assessed by the test under development or consideration. Federal legislation and any state or local policies that expect to use the testing results for system-wide and/or program evaluation assume that all students who receive money from the system or are taught under these programs will be part of the accountability system. Since it is impossible to use and compare scores over instruments that do not measure the same construct domains with any degree of reliability, it becomes important to use the mainstream test for as broad a set of students as possible. In order to sample students accurately with a broad range of English language proficiency, it probably will be necessary to administer the pilot/field test under accommodated conditions for some students.

2. Collect work from students from the primary language groups in the client area. For instance, California will collect work from Hispanic students (certainly from Mexico and possibly from other Latin American and/or South American groups depending on their representation in the state/districts), Chinese, Tagalog, and some Southeast Asian groups, e.g., Vietnamese and Hmong, among others. Oregon probably would collect work from Hispanic and Russian students. As a general rule, if there are high enough numbers to disaggregate the final data, client area-wide (e.g., statewide or district-wide), students from these groups should be included in the sample. Of course, broad representation from other groups is also meaningful and should occur if possible.

Since it is impossible to use and compare scores over instruments that do not measure the same construct domains with any degree of reliability, it becomes important to use the mainstream test for as broad a set of students as possible.

3. Collect LEP student work across all item score points. Sampling to ensure the full range of score points provides an inexact estimate of collecting item samples from students across the entire range of content mastery. This is important psychometrically, as it gives us an idea of how students at different mastery levels interpret the item. It is possible that a part of the item may inadvertently but systematically trigger confusion in students from particular mastery levels.

4. Determine the number of students to be sampled. First, make a matrix representing a cross between the three variables listed in recommendations 1-3 (English language proficiency, primary language groups, score points), each with their respective levels (for instance, low, mid and high proficiency levels might be identified). Table 5 provides an example of the matrix. In its most disaggregated form, each of the boxes formed by the variables is known as a cell.

The sample should contain approximately 25 students per cell, to ensure stable, generalizable results. For example, in Table 5 one cell would represent the responses of 25 Hispanic students, low in English proficiency, with an expected score of 4.

In this figure the variables and levels that categorize the cells are Primary Language (Hispanic, Chinese, Hmong), English Proficiency Level (low, mid, high) and Score Points (1, 2, 3, 4). If one assumes that there are four forms with non-overlapping item sets, this would mean that 36 cells/form x 4 forms x 25 students = 3600. That number of LEP students would need to be sampled at each grade tested. This does not take into account other demographics important for other reasons, (e.g., gender, geography, rural/urban), or the need to oversample.

5. Because of the cost and time involved in gathering adequate samples of work from LEP students, we recommend two approaches to gathering the numbers of

Table 5 ○ Sampling Design Example

	PRIMARY LANGUAGES								
	Hispanic			Chinese			Hmong		
Score Points	English Language Proficiency			English Language Proficiency			English Language Proficiency		
1									
2									
3									
4									

responses. The first approach focuses on the minimum numbers acceptable when agencies are responsible for schooling and testing small numbers of LEP students. The second approach focuses on recommendations to publishing houses.

- The minimum sample size for agencies responsible for schooling and testing small numbers of LEP students should be approximately 75 LEP responses per form (with non-overlapping item sets), per grade. The 75 should be carefully chosen to represent all cells discussed above (representation along the entire range of English proficiency, from all major language groups, and at all score points). While there are not enough responses to suggest full representation within each crossed cells, each of the variable levels should be represented by approximately 25 LEP student booklets. For example, the sampling design will call for 25 Hispanic, 25 Chinese, and 25 Hmong student responses, while concurrently trying to obtain about 20 responses at each of the four score points. These numbers are far from ideal because less than 25 responses per cell could lead to unstable results, and these results almost certainly will not be spread evenly across different variables. As an additional complication, the test developers will in all likelihood have to go out and handpick classrooms which could add significant bias to the work. Hopefully these numbers can be augmented with data from future administrations, and the results reevaluated as possible.
- Test publishing houses should collect work samples on items across clients (SEAs, LEAs), whenever possible , when they sample to document the validity for their off-the-shelf tests. Sampling the forms of off-the-shelf tests for LEP concerns can be done for clients at a lower cost and produce a greater array of student work to review if publishing houses share the collection and analysis of the responses with other states or districts. For example, Philadelphia could collect and analyze the work about some items on a selected test, whereas California could be responsible for work on other items and perhaps from more LEP primary language groups.

Publishers should build these criteria into their sampling design as new pre-built or customized tests are constructed, spreading out the costs as possible. Even custom clients should be amenable to piloting and using certain items if they know of their benefits for their LEP students (and hopefully other students who are divergent from the mainstream). Publishers, of course, could use these data not only in development, but also to build norm samples.

Collection of Student and Teacher Information

Besides completing the test items, what else can students reasonably be asked to do during the pilot/tryouts or field test trials? We recommend collecting the following student information:

- Student responses to test items
- Student edits
- Information about each student's LEP and English proficiency status, as well as accommodations used
- Student feedback about what the items measure (this information is collected from students as a group).

Teachers should be asked to provide edits, and other types of feedback to the items, test format, procedures, or other testing elements.

Teachers should be asked to provide edits, and other types of feedback to the items, test format, procedures, or other testing elements.

Recommendations for Collection of Student and Teacher Information

Student Information

1. After completing the test, students should be encouraged to edit the items, circling all words, phrases, or sentences they find confusing; they may offer suggestions for improvement as well. Student edits can be done directly on test booklets, which are collected and kept for review by the test developers, as well as on any answer sheets.

2. Students, or their teachers, should record on the student answer sheet or booklet (a) the student's primary language group, (b) what test accommodation(s) the student used, and (c) the student's level of English proficiency. Accurately assessing the student's proficiency level is very tricky. Until a standardized way of identifying and recording level of proficiency is adopted, we recommend that test developers offer the choices of low, midrange, and high proficiency, and provide an explanation that describes how each is to be identified. Local and state LEP educators can help with this in a particular client area, or test publishers may want to come up with a common set of definitions to use across clients (SEAs, LEAs).

3. Students should be asked two additional questions (in the student vernacular):

- What did students think the questions were asking them to do?
- Why did they answer as they did?

Since these questions should be asked about each test item, teachers should spend a few minutes after the test asking these questions of the students in a group. Next, they should record their answers on a sheet which is part of the testing materials, such as a pullout sheet in the back of the teacher administration booklet. If forms are spiraled within a classroom, that is, students in a classroom are given the full range of forms (in alternating fashion), students can be grouped by form after the test and a student scribe (or parent helper in the younger grades) records the responses. All recording sheets should be identified by classroom level, so that the responses can be linked to primary language groups, level of English proficiency, and range of scores in the classroom. The links are important for validity analyses purposes.

Teacher Input

4. In addition to recording student comments, teachers should be asked to edit the test booklets as well: circling what they find confusing and providing suggestions for improvement.

5. A pullout sheet in the back of the teacher administration booklet should also be completed which asks teachers to provide any other information that they think might be useful for the developers. This step is invaluable because these teachers are in a unique position to understand both what they think the test developers are trying to find through their items, and the capabilities of their students.

The pullout sheet should allow space for teachers to respond to any and all parts of the testing process, including the layout of the forms themselves, the accommodations, and other procedures. Teachers also should be encouraged to comment on the content of the items, what they think the items measure, whether the items are reasonable, and their rationale for making these decisions.

6. The teacher should record classroom-identifying information on the pullout sheet, or should describe the class in a global way, e.g., the percentage of LEP students in the classroom, primary language groups, and average level of English language proficiency of the LEP students.

Items and Related Materials

What else should reviewers expect to have in front of them in order to properly evaluate the validity of items for LEP students? In general, besides the items themselves, completed test forms, and student and teacher feedback, materials should be provided which give summary information about how the items are performing, and which define what items are intended to measure.

Recommendations of Additional Materials To Be Used in the Item Evaluation

1. Publishers should provide statistical summaries as they are available. Statistical information typically includes some type of level of difficulty index (such as p-values, or item respond theory functions), item-total correlations, frequency distributions, and perhaps differential item-functioning (DIF) results. Because accessibility issues are confounded with multidimensionality of constructs in DIF, judgment reviews of items for inclusion based on DIF should try to distinguish these sources of systematic error. These reviews may or may not have been done prior to the item evaluation. If so, these results should also be included here.

2. Copies of the item construct statements need to be collected for the item evaluation. These spell out what each item is supposed to measure. This accessibility information either has been collected as part of the item-writing process, or should be collected as part of the review of off-the-shelf tests. The evaluators need to determine if the measurement of irrelevant skills and knowledge has been minimized through the drafting of the items, and/or the conditions under which students are taking the assessments.

3. Supporting information about test accessibility needs to be available. This includes information about the development of items as it relates to accessibility, e.g., evidence of and from the bias reviews, test specifications accessibility framework, item and rubric writing procedures, administration and response accommodations allowed.

Conducting the Item Evaluations

Who participates in the evaluation, what procedures and considerations are recommended, and what types of questions and substantive perspectives should guide their reviews?

Participation in the Item Evaluations

In a typical test development item evaluation review, item writers (or item writing teams) review evidence about how the items are performing and make changes to the items accordingly. A psychometrician and staff developer familiar with practical considerations and with the technical and practical constraints, respectively, also are usually involved. Publisher project directors or their representatives and staff from the education agency, if the test is being developed for them, weigh in as well. When off-the-shelf tests are under review, participants typically include content area specialists and measurement specialists from the education institution, as well as appropriate leadership staff.

In reviewing items to determine if there is sufficient evidence of validity for LEP students, additional participation is recommended.

The procedural points addressed here are meant to provide guidance about how developers or reviewers might supplement their existing evaluation methods.

Recommendations for Participation of LEP Experts

1. Educators who work with LEP youth — as well as participants who are part of typical item evaluations — should be involved in reviewing the information from LEP students and their teachers. Based on how extensive the LEP student sample actually is, it will be important to involve educators of different primary language groups so that they can reach an accurate understanding of the performance of students with a wide range of English proficiency.

2. Educators who work with LEP youth also should be represented when final decisions are made about items. In many cases, recommendations about items and final decisions about which recommendations are accepted are two separate steps. It is important that LEP experts be involved in the latter step as well as the former.

Item Evaluation Procedures

This section discusses general considerations and outlines an approach for evaluating the validity of items. The procedural points addressed here are meant to provide guidance about how developers or reviewers might supplement their existing evaluation methods.

Recommendations for Evaluation Procedures

1. All student, teacher, and item information should be available for each item, during evaluations occurring as part of test development. It would be reasonable to expect that this information from a random sample of items should be available during the review of off-the-shelf tests.

Typically, constructed item responses and performances from small-scale pilot tests/tryouts are not scored prior to the item review, so evaluators are asked to score informally student responses while they are evaluating items for other problems. Student responses to multiple choice items are usually scored if answer documents are scannable. After large-scale field tests, items are usually scored before this evaluation review commences. While a sample of non-LEP student work may be drawn from large trials to review, all LEP information probably will be evaluated because numbers are relatively small.

2. It is recommended that the LEP student responses and associated evidence be evaluated separately from other responses initially, and then be part of a final aggregate item review at the end of the evaluation process. LEP experts, mainstream item writers and content experts, and technical and professional staff would review LEP information as a team, similar to how the reviews are currently done. But this recommendation calls for an initial disaggregated analysis of the LEP evidence that is separate from the analysis of evidence from the rest of the population. The reason for this extra step is that the validity problems typically have not been identified when items are evaluated with evidence from an aggregated sample. While it can be argued that disaggregated evidence can be evaluated within the one general step of review, we are skeptical that the proper attention will be paid to identifying the validity problems of LEP students and subsequently recommending effective solutions.

In the second step, item evaluation recommendations from the team that reviews the LEP information should be compiled with other item evaluation recommendations from those reviewing the total student sample (along with recommendations which might have come from other special populations, such as educators of students with disabilities). Decisions about final changes to items will be made at the time of the final review and be based on the composite of these recommendations.

3. The LEP evaluation review, by item, is usually done in a group that consists of three or four evaluators with complementary skills. Depending on the number of reviewers, one group may evaluate the complete set of items, or a number of these groups may review the pool of items, so any particular group will be responsible for a subset of items. Clearly, consistency across groups should be monitored.

4. The focus of the item evaluations are to identify systematic errors and legitimate problems in test items for LEP students and then to provide usable solutions.

- The charge given to the evaluators is to determine and correct item problems, across and within: (a) score points, (b) primary language, and (c) English proficiency status.
- Responses and related information can be sorted and reviewed by one indicator at a time, or items can be evaluated wholistically across the full set of indicators at once. Either way, a running record of recommendations should be kept which

summarize problems and solutions. Trends are usually not difficult to spot.

- The amount of time needed to review the items depends on the type of item, the number of responses, the amount of related information which has been collected, and how the group participants choose to work (e.g., do they each look at all items/related information, or do they each review selected items/information and then discuss their findings as a group?).

5. The final evaluation summaries from the LEP group(s) that go to the larger, across-population, final-evaluations review should center on:

- Identifying trends across or within indicators for LEP students
- Coming to a consensus about concerns or problems
- Suggesting ways to amend the item or how to proceed from this point.

Group validity responses from the students, student and teacher edits, and other teacher feedback should either confirm hypotheses about problems based on how students responded, or call the hypotheses into question.

6. To some extent the evaluation of items differs by item type. Following is a brief discussion of considerations for forced choice (multiple choice, true false, match) and constructed response/performance items.

- The evaluation of multiple choice and other force choice items relies mostly on student edits and their validity responses, teacher feedback, student scores, p-values or other statistical indicators of difficulty, and the statements about what items are supposed to measure (the accessibility information about measurement intent).
- The evaluation of both constructed response and performance items are treated similarly here. Considerations about these item types are substantially more complex. Procedures and other issues which stem from this complexity are discussed below.

Frequently, the evaluation of these types of items begins with sorting responses (and other related information) by low, midrange, and high scores. Within these piles, responses are usually categorized as typical (correct or not) and outliers. The "typical" responses tell the evaluators what students think the item measures and what may interfere with a correct response — for example, confusing wording, irrelevant information, or format problems. Group validity responses from the students, student and teacher edits, and other teacher feedback either should confirm hypotheses about problems based on how students responded or call the hypotheses into question. Reviewers occasionally read through student and teacher feedback before analyzing student responses; next they review feedback in more detail after responses have been discussed.

As reviewers continue their work, they look for systematic concerns sorted by English language proficiency and primary language group. These concerns occur within and across responses categorized by score points, including among students who share common cultural or background experiences. For instance, it is not uncommon for students who have grown up in developing countries to respond in a similar fashion, even though they come from different language, ethnic, or racial backgrounds.

7. Measurement experts have been trained to ignore outliers. While outliers may be just that — perhaps a record of a student's wanderings independent from the question — they also may provide insight into problems only hinted at within the typical responses. As such, outliers also need to be analyzed. Sometimes reviewers will record outliner issues on a separate sheet of paper. These may not be confirmed

Table 6 ○ Questions To Guide the Accessibility Evaluation of Items

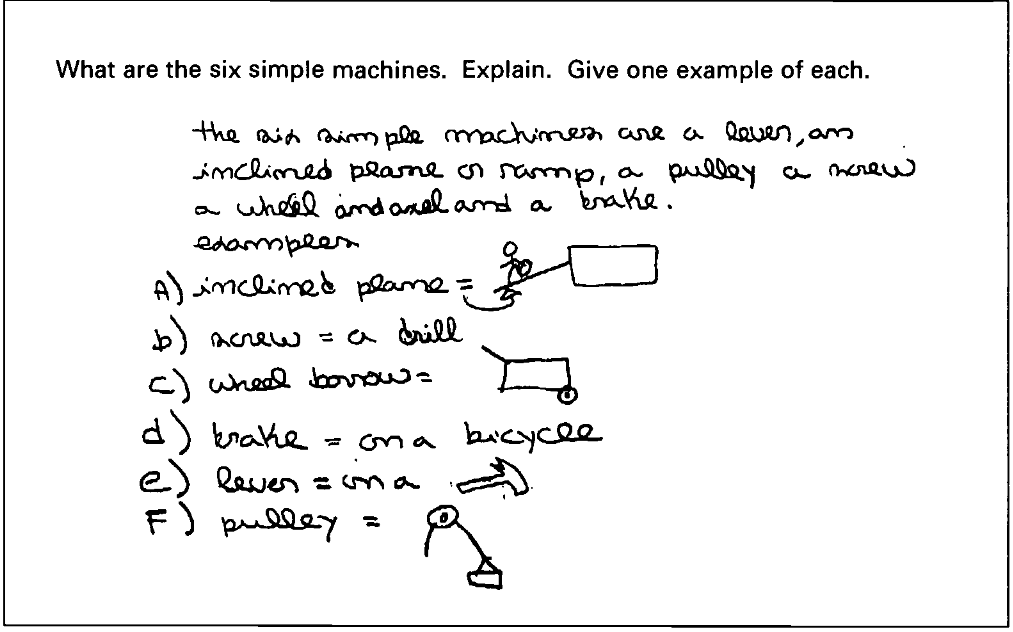
- **Are the expectations clear for what is being asked of the student?** This includes content, format (e.g., answer space), type of response, and any other limits the item writers feel is appropriate (for instance, whether they expect technical words from the content area to be used).
- **Does the item implicitly narrow down the type of response unintentionally, by the nature of what is being asked, how it is being asked, or by way of examples?**
- **Are the items free from linguistic concerns?** These concerns include the use of unnecessarily complex or unclear language and/or sentence structures, low frequency words or language, or the use of phrases that, when translated into a primary language, mean something different.
- **Are the format and presentation of items as accessible as possible?** This includes the spacing of items on the page, use of visuals as needed, and access to tools and resources when items do not measure that particular skill.
- **Is any contextual information presented around or as part of the item straightforward, linguistically clear, and presented in a clear format?**
- **Are the items free from confusion based on the diverse cultural backgrounds of students?** Do the items unintentionally assume common prior experiences, including experiences common to growing up in the U.S., or experiences more common to a particular lifestyle?
- **In evaluating the responses to the constructed response items, within and across score points, do the LEP students appear to understand what the item is measuring?** Other than the issues identified above, do there appear to be other problems in the items which lead to systematic confusion (within or across score points) about what information the item is requesting from the students or how the item is requesting the information?
- **Other than when items are measuring the capabilities of the students to read and/or write in English, are the requirements in the items free from challenges of the students based on their level of proficiency in English?**
- **Are the requirements in the items free from systematic problems based on the students' primary language group?**
- **Does item type seem to be appropriate for measuring the specified construct for these students?**
- **How do the response themes and the progression of scores of LEP students compare with the responses and scores of non-LEP students?** Are these findings appropriate or do they highlight additional problems, and if so, what are those problems?

or upgraded for possible consideration after looking at other information from students and teachers. Those which are not confirmed or discounted may might be consigned to issues to be evaluated after obtaining additional work.

8. The chapters which focus on how to make items accessible, that is the chapters on presentation options, writing accessible rubrics, and administration and response considerations, should provide a conceptual perspective which will guide the evaluation.

9. The specific kinds of questions which should guide the evaluation of items include linguistic concerns: issues in the readability of the items, as well as cultural issues, including issues related to students who come to our schools with very different experiences. The questions in Table 6 provide an approach to guiding the review.

Figure 14



What are the six simple machines. Explain. Give one example of each.

the six simple machines are a lever, an inclined plane or ramp, a pulley a screw a wheel and axel and a brake.
exampels
A) inclined plane =
b) screw = a drill
c) wheel borow =
d) brake = on a bicycle
e) lever = on a
F) pulley =

Rubric Evaluation

This evaluation includes a review of the scoring rubrics, rubric notes, and associated scoring materials. The notes and materials include examples and specify parameters of what is expected, and therefore expand the meanings of the rubrics in how each score point is to be interpreted. The accessibility evaluation of the rubrics , notes, and materials should be completed by reviewing the same information collected for the item evaluation. This review is accomplished concurrently with the item evaluation or shortly thereafter. Rubric revisions may occur independently of item changes, or proposed item revisions may affect changes or clarifications in the rubrics or vice versa.

The evaluation of rubrics, notes, and materials should occur whenever an assessment is being developed, as well as when an off-the-shelf test is being considered for adoption. Just as it is imperative that an evaluation of items be conducted when pre-built tests are reviewed, rubric evaluations are also important to ensure that the scoring guides are appropriate for a diverse population.

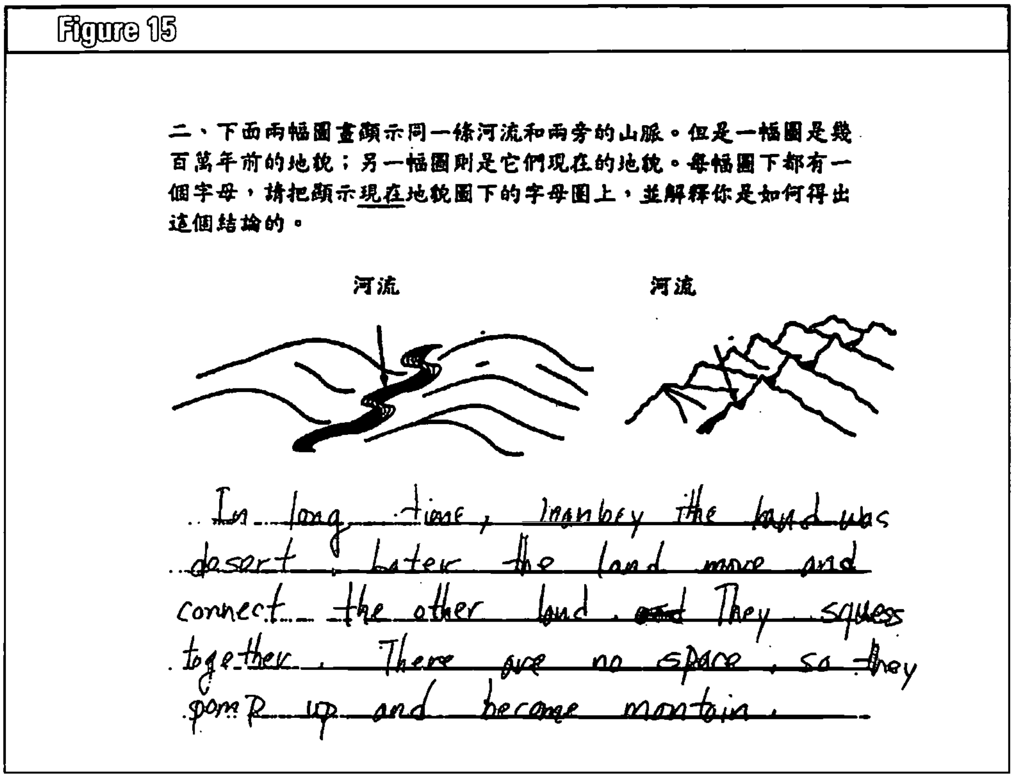
The evaluation of rubrics, notes, and materials should occur whenever an assessment is being developed, as well as when an off-the-shelf test is being considered for adoption.

Figure 14 illustrates a response which uses pictures in a central way to answer the question. It is important to emphasize the role of rubric criteria in the scoring of student work. Inclusion of illustrations (such as this figure) and other alternative response formats in scoring materials as examples of acceptable representations of content is very important.

Recommendations for an Evaluation of Rubrics

1. We recommend that rubric resources be evaluated formally to determine if they are adequate for scoring the constructed responses of limited English proficient students. We believe the systematic evaluation of rubrics, notes, and materials is necessary, because they have been a significant source of validity problems for LEP students.

Figure 15



二、下面兩幅圖畫顯示同一條河流和兩旁的山脈。但是一幅圖是幾百萬年前的地貌；另一幅圖則是它們現在的地貌。每幅圖下都有一個字母，請把顯示現在地貌圖下的字母圈上，並解釋你是如何得出這個結論的。

河流

河流

In long time, lnanbey the land was desert. Later the land move and connect the other land. ~~and~~ They squess together. There are no space, so they pomp up and become montain.

2. Resources to be evaluated include item materials, rubrics , notes, and associated materials used for scoring. The materials to be included in the evaluation include the items and item construct statements, the rubrics, rubric notes, and scoring materials such as training sets and other training material. A description of these resources can be found in the item, rubric, and scoring chapters.

3. It is reasonable to expect that test publishers will produce suitable rubrics, notes, and scoring materials for this purpose. If educational institutions find that the rubrics, notes, and materials are lacking, it is expected that publishers will work with the institution to supplement the rubric resources, so that they will meet the needs of the institution before the assessment is implemented.

4. It is crucial that educators familiar with other cultures review and evaluate the rubrics and rubric notes. In many cases, the item is appropriately accessible, but the responses have been too constrained for certain types of LEP students. The perspective of these experts is important if response bias is to be reduced.

Figure 15 illustrates a common phenomena in LEP student responses. It is not unusual for students to read an item in one language and respond in another. Rubrics and scorers must be prepared for this.

5. In the analysis of the rubrics, notes, and materials, the focus is to ensure that:

- the rubric resources are consistent with the expectations and accessibility of the items under review
- specific, identified, construct elements are what is being scored in the item responses, and not skills which are irrelevant to any particular construct element
- the same meaning is applied consistently through each of the score points.

Table 7 ○ Questions To Guide the Accessibility Evaluation of Rubric Resources

- **What types of response forms are allowed by the rubrics? Do they allow for (a) oral response, (b) non-text paper and pencil responses, and (c) non-paper-and-pencil performances of mastery?**
- **Do the rubrics allow LEP students, with their attendant linguistic challenges in English and divergent cultural experiences, access to the full range of score points?** This should be true whenever the specific linguistic or cultural constructs are not what the item measures.
- **Do rubrics allow for a variety of different student responses?** The more ways rubrics allow students to respond, the better. This is consistent with what we know in general about the diverse ways students learn and demonstrate knowledge. Non- standard response strengths are often more visible than usual when students are not proficient in reading or writing English, and/or when their cultures value response modes other than the types valued by the U.S. or allowed in the assessment.
- **What is the percentage of items and rubrics on a test form or in an item bank that satisfies the conditions in Questions 1 and 2?** It is fashionable today for tests to say they are accessible, and often one or two constructed response items/rubrics will allow diverse responses with some thought given to cultural issues. However, a majority of items/rubrics should be designed within conditions that *require* accessibility. Of course, accommodations affect this accessibility, over and above the structure of the items and rubrics. Rubrics and their notes should be built and evaluated to be accessible whenever linguistic and cultural issues are irrelevant to the construct being measured. When that occurs, the accessibility of a test or form as a whole can be evaluated in light of accommodations, as well as in light of whether the items and rubrics have been found to be satisfactory based on the points discussed here and in the preceding sections.

The chapters on writing rubrics and scoring responses provide the conceptual framework to guide this evaluation. These chapters discuss the issues in more detail and should be used as a resource during the evaluation. Table 7 outlines the types of questions to be addressed during the accessibility evaluation of rubrics, notes, and scoring materials.

Form and Test Evaluation

Throughout this *Guide,* we have maintained that significant evaluation of technical merit must occur, by item, more often and at a deeper level of scrutiny than is typically done, in order to end up with an assessment that measures what it is supposed to measure — for all students and over time. This evidence of technical merit must include inspection of direct student evidence and expansion of opportunities for expert judgment (including input from teachers, specialists, and students).

Traditional methods to determine validity have relied almost exclusively on interpreting construct or domain validity at the form or test level, and on formally evaluating technical quality through summary statistics limited to large samples, and usually across items. These methodologies have relied only on indirect means of evaluation, such as comparing a new test to an established test, assuming that the established test is valid.

We realize that, since most items measure only parts or elements of constructs, technical evaluations at the item-aggregate level are important. This section focuses on explaining procedures for evaluating the accessibility of forms and tests. We will not discuss the range of analyses which are typically at this level, because this information is readily available elsewhere. Rather, this chapter will discuss the evaluation of how items are allocated to forms and the relationship of forms to tests.

The discussion which follows is appropriate whether tests are being developed or tests are being reviewed for selection.

The Allocations of Items to Forms and Forms to Tests

The purpose of this allocation is to determine the accessibility for LEP students within and across test forms. The allocation of items to forms and forms to assessments, per content area and per grade, involves a series of decisions about coverage of content/process and also some understanding about depth and its coverage. It also demands practical decisions that are constrained by issues such as cost, as well as student, teacher, and district time. As we pointed out earlier in this *Guide*, allocation, in terms of accessibility, focuses on whether the coverage decisions remain constant for all students who take the test. For example, are the same science constructs and their elements actually measured in the forms and tests for all students? Or are the measurements confounded by such issues as linguistic or cultural problems to the extent that they are systematic and pervasive enough to distort the summary estimations of the mastery of LEP students?

The first part of the evaluation of forms and tests is to determine the level of accessibility in the allocations of items and forms.

Recommendations for the Accessibility Allocation of Forms and Tests

1. The accessibility of items should be a central consideration when items are allocated to test forms. As a test form is finalized, items are typically allocated to forms based on content coverage and other coverage considerations. Coverage is usually defined in the test specifications, which were written to guide the development of tests. In the chapter on test specifications, we recommended that adequate accessibility coverage be explicitly expected, in addition to content, process, and depth coverage.

2. An accessibility framework for determining allocation, such as the Accessibility Framework for LEP Students, should be completed at the item level for each form and for each form of a test. A discussion of the Accessibility Framework can be found in the test specifications chapter. The framework not only guides the development of items, but can serve as a method to determine if forms containing finalized items are adequately accessible for students.

3. Accessibility of multi-form tests, per content area and per grade, is seen as adequate coverage over matrices, consistent with the reporting expectations of school-level information as well as individual-level information. Determining what allocations to review for accessibility depends on how scores will be reported.

For instance, when a student gets a score at the content area level, e.g., mathematics, then accessibility for students will be focused on the form taken by the student. If a student gets subtest scores, then adequate accessibility should be expected for each reported subtest score within the test form administered to the student. In both these cases, the accessibility allocation will focus on the forms. In the first case, the entire form will be evaluated for adequate accessibility; in the second case, the set of items comprising the each subtest score should be reviewed for accessibility.

Sometimes tests are designed so that items are matrixed over students, and school-level rather than individual-level scores are reported. In this case, often subtest scores are reported on composite sets of items that are given to different students. Accessibility, then, should be determined based on the types of scores that will be reported, and on the items that make up the subtest scores. In this case, the items will appear on different forms and the focus of the allocation will be the items, over forms, rather than on the forms themselves.

Evaluation of Forms and Tests

The evaluation of form and test accessibility should determine whether there has been sufficient accessibility coverage, for any and all sets of items upon which results are reported. The Test Specifications Accessibility Framework for LEP Students specifies that access is accomplished by reducing the reliance on certain modes that are confounded with the limitations of LEP students.

The evaluation of form and test accessibility should determine whether there has been sufficient accessibility coverage, for any and all sets of items upon which results are reported.

Recommendations for Accessibility Evaluation of Forms and Tests

1. There should be maximum accessibility of items for LEP students. Regardless of the types of scores that are reported, it is expected that the items which comprise the scores will be adequately accessible for limited English proficient students. If not, then it is unclear if scores represent their mastery of particular subject matter, or their literacy challenges and different cultural experiences.

2. Most items and conditions should allow for multiple-access opportunities over a variety of different strengths. We define accessibility as a function of access to the item requirements and access to the ability to demonstrate mastery. For sufficient coverage to occur, students must be able to access what items are asking, and be able to demonstrate what they know. This can only occur if students can use their skills, rather than be blocked by their limits.

3. We recommend that at least 75-80 percent of the items should be accessible to LEP students. This is an “educated guess” of the percentage of accessible items which are necessary to accurately reflect what students know. Research needs to be done to determine the lowest thresholds in this situation. Certainly, if an assessment has few items, then a greater percentage would be required.

4. Evaluation recommendations should be made at the level of items and/or forms. That is, if test-level accessibility appears to be sufficient, then the form or test is approved to be used as is. If accessibility is not sufficient, then problematic items should be identified for replacement or modification in order to increase accessibility. Administration or response conditions which affect the entire form or item blocks also may need to be expanded. Once these recommendations are followed, the revised forms will be ready to use.

Additional Validity Evaluations

There are two types of additional information that help validate assessments for LEP students:

- An independent validation of the items based on a collection and analysis of more detailed information from a smaller sample of students, and
- An analysis of classroom teaching/learning opportunities, their relationship to subsequent scores, and their connection to the quality of student responses.

These analyses certainly go beyond the scope of data collections described here, and are recommended as supplementary analyses which should be conducted early in the implementation of the assessment.

Independent Validation

To help ensure that items measure what is intended, additional data from a small subgroup of students should be collected concurrent to the pilot, field test, or implementation of the actual assessment. These data can be collected by asking students to use the "think aloud" procedure. This procedure asks students to talk about what they think is required by the items, and the thought processes and approaches they go through in answering the items. The interviewer records what the students say. Such data also can be collected by using a "stimulated recall" procedure where students go back over their work shortly after the assessment administration is completed and explain to the interviewer what they were thinking when reading the item and selecting/formulating a response. Interviewers also may record student grades or teacher judgment of mastery in the subject which is being tested, and collect further work samples if possible. Obviously, this data collection exercise can be made quite rich and complex as time and funds allow. The reason a data collection such as this is important is that there are extreme shortages in information about what items actually measure for the LEP population. Studies of this type would enhance what item reviewers can surmise by looking at student responses and related information described throughout this chapter.

Ongoing Access to Learning and Performance Opportunities

The second type of investigation focuses on the relationship between what students learn in class and how they respond on the assessment. It involves analyzing the relationship between their test responses and what is being taught and how as well as the experiences the students have had in class with the types of questions and available accommodations which are used in the assessment. It is beyond the scope of this *Guide* to delineate the types of data collection which these investigations should involve. Survey forms or interview protocols can easily capture if students are using the types of accommodations on an ongoing basis in the classrooms, and a review of their tests can highlight the types of assessment opportunities which are currently available to them in particular classes. Evaluating pedagogical opportunities and curriculum content is more complex. However, all of these types of analyses are important, especially if program/school or individual stakes are attached to these assessments.

Chapter 10

Accessibility Considerations When Reporting and Using Test Results

In this chapter, we briefly address selected issues associated with the reporting and use of assessment results, including the reporting requirements spelled out in recent federal legislation. We also discuss issues related to reporting test results for LEP students.

Reporting Results

Recent federal legislation requires that results of assessments must be disaggregated for special populations, including LEP students (Elementary and Secondary Education Act, 1994). The same legislation emphasize that all students should be included in the same mainstream testing system. For example, Section 306 of Title III requires state improvement plans to include "a process for developing and implementing valid, nondiscriminatory and reliable state assessments aligned to content standards." These federal mandates state that school-level results, including district or county results if they are disseminated, be reported by categories of students (gender, LEP, poverty, and special education status) as well as in the aggregate. These disaggregation policies are intended to ensure that stakeholders can ascertain how well different groups of students perform under the same school administration policies and within the same (or parallel) school programs.

Recent federal legislation requires that results of assessments must be disaggregated for special populations.

In addition, the *1999 Standards for Educational and Psychological Testing* repeatedly underscore that all reported results must be technically sound. This means that validity data should be available by group (e.g., level of English proficiency) to support the reported results, if credible research suggests that the inferences of the scores might differ over groups.

While some might attribute disparate results to a range of background factors, recent reform efforts suggest that many of these factors can be mitigated or overcome by effective pedagogy and systemic support across the educational delivery system (Slavin and Fashola, 1999). Therefore, the differential impact of instruction is an indicator of curricular or instructional weakness that should be addressed through instructional policy mechanisms.

Of course, in order to make judgments about the causality of differential impact, several elements must be in place. To determine that the disparity in educational access is a primary explanation, it must be very clear that the assessments are properly aligned to the curriculum standards upon which the instructional system is based. If alignment is not strong, than results could be attributed to the gap between what is valued by the educational system and what is being tested. It is also imperative that the assessment be valid for all students who take the tests, which includes evidence of validity for any and all scores (such as subtest scores) for all students included in the test. If the assessments are in alignment with the standards, but there has not been adequate demonstration of test alignment for all students (including language minority students), then disparity could be a function of problems in the assessment instruments.

The impact of excluding LEP students from mainstream assessments is substantial. It is easy to ignore their needs if fiscal and pedagogical priorities depend on the assessment results of only a subsection of the student population. The same is true if English language learners are included, but their performance is not evaluated separately. It is impossible to determine how well or poorly LEP students are learning, particularly in programs where a range of students are being served.

Some believe that if aligned, reliable, and valid results are collected for the full range of students taking the assessments, then it should be possible to both aggregate and disaggregate the results. This applies to results where accommodation options have been employed in test format, administration, and/or response conditions. This would adhere particularly in accountability systems when assessment results are used to evaluate the success of programs when heterogeneous student populations are being served.

Use of Assessment Results

The results of academic achievement tests for K-12 students are used in a variety of ways. Large-scale assessments are typically standardized. Results may be reported at the individual student level or at the district or school level (reports from the school level could be broken out by teacher, grade, school, or program). The scores may be used, in whole or in part, to make high-stakes decisions such as placement, promotion, or graduation at the individual level. They also could be used to distribute rewards (and sanctions) or set fiscal or pedagogical priorities at the group (district or school) level. They can be used in various low-stakes situations, such as a formative evaluation indicator of individual progress or program success. However, some argue that no situations are low-stakes — all tests carry implications for use that ultimately affect students.

The responsibility of standards-based evaluation is to make sure the full range of the standards are assessed.

The assessments we discuss are built or selected in response to educational curricular constructs identified in the state or district content standards. The responsibility of standards-based evaluation is to make sure the full range of the standards are assessed.

Recommendations for using test results for LEP students in high stakes decisions will be briefly outlined.

Recommendations for Using Tests in High Stakes Decisions at the District/School Level

1. Limited English Proficient students must be included in these large-scale tests and their presence must influence the tests in all ways. If LEP students are excluded from large-scale tests based on an educational or psychometric justification, then comparable information must be collected about these students' academic progress. Aggregate test results often are the only information used to make high-stakes decisions. These results should show how well students are doing in school.

2. Decisions should be based on information about achievement relative to the range of knowledge and skills in the content standards. This includes more aspects of achievement than can be determined from only forced-choice and short-answer items. While standards differ to some extent across states and districts, all standards value both breadth of content and depth of knowledge about the content.

Priorities in content curriculum standards must be reflected accurately in the assessments by item types and weights of responses. These evaluate not only content breadth in low-level problem-solving situations, but depth of knowledge in more complex problem-solving milieus.

3. If significant differences are found by schools or programs, or if there is disparate impact in the disaggregated data by LEP status, racial/ethnic background, or poverty level, a determination of its cause should be a central element in making high-stakes school or program-level decisions.

This type of investigation should provide evidence about how and where the curriculum and/or instruction implementation is out of alignment with the standards. This can be done only if assessments collect enough information that is consistent with the range of goals and priorities identified in the standards. If the schools or programs are not held responsible for the range of content and depth identified as valuable in the standards, the assessment actually encourages the narrowing of curriculum from the intent of the standards to the reality of the instrument being used to evaluate the standards. Unfortunately, research has shown that minority populations, including LEP students, will be some of those most affected by the abandonment of the standards in the everyday practice of their schooling.

Recommendations for Using Tests To Make High-Stakes Decisions at the Individual Level

1. If access to educational programs may be denied certain individuals based on test results, two questions need to be addressed:

- Are students allowed to provide evidence apart from the test that reflects their mastery relevant to the entire spectrum of the standards?
- Is this evidence used as a key component when high-stakes decisions are made?

2. The test should provide a reasonable indicator of achievement for LEP students, as well as for other students. Additionally, evidence from other multiple measures should be collected and used along with the assessment results to support sound and accurate decisions for students. Unfortunately, it is not unusual for educators and test publishers to assume that it will be the other multiple measures evidence (besides the test) that will provide the evidence for "those kids." This can be read as: "We don't need to fix the tests to include all of your population, even though we encourage you to have an independent indicator besides teacher judgment (i.e., standardized tests) for most of your population." Somehow, the other measures, by themselves, magically seem adequate enough when evaluating special needs populations which include LEP students.

3. Additional achievement evidence includes classroom information, and possibly some types of embedded items or portfolio results. However, to be valid, any evidence must be assembled systematically to reflect alignment with the full range of knowledge and skills identified in the standards. The evidence should be collected as per the guidance and criteria developed by those who establish alignment evaluation procedures. These procedures and the information they yield should be able to withstand technical and legal scrutiny; they should produce valid and stable inferences about the performance of all students.

4. Significant legal questions remain:

- What happens if students are not able to provide evidence of sufficient mastery to be granted access (as defined by the nature of the high stakes decision), but have not been provided ample opportunity in their schooling to develop these knowledge and skills?
- How should these failures be handled?
- What outcomes can be expected?

These legal questions extend far beyond the parameters of this *Guide*. However, this *Guide* should provide sound guidance to ensure that valid information from tests and from other types of evidence is collected for English language learners. Literacy levels and/or culturally-based challenges of these students do not need to be impediments. Instead, evaluations responsive to the guidance we have outlined can show conclusively that these students can and are meeting the highest standards of academic achievement.

Appendix A

Technical Discussion of Issues Related to Accessibility

In this Appendix, we highlight selected issues and considerations that pertain to the statistical analysis of bias, comparability issues, linking across tests, and some additional thoughts related to construct validity and equity.

Using Differential Item Functioning (DIF) Results

What is DIF?

In 1984, the court settlement between the Golden Rule Insurance Company and the Illinois Department of Insurance/Educational Testing Service specified that a raw difference of 0.15 or more in an item's p-values (level of difficulty of an item), favoring White over African American applicants, was evidence that the item is biased. The case focused on considerations for inclusion of items in the Illinois insurance licensing examinations. This agreement determined that biased items, as defined by the above criteria, should not normally be included in the test (original cite).

Is the difference due to irrelevant characteristics, such as racial/ethnic group membership, or is the difference due to content central to the construct?

While this finding has initial appeal, a sounder distinction lies in the identification of items where population subgroups respond differently within comparable test score levels. That is, given a Caucasian group and a Hispanic group of students, do students who receive a similar test score on a particular test respond differently to an individual item from the test, and does this difference appear to be systematic by group? This is the central question which differential item-functioning (DIF) statistics have been designed to address.

A second related question also needs to be considered: Is the difference due to irrelevant characteristics, such as racial/ethnic group membership, or is the difference due to content central to the construct? If the difference is central to the construct, the variance could be due to unequal familiarity with the particular concept being evaluated (in other words, an unequal educational access issue). It could also signal multidimensionality in the construct wherein groups behave differently on different dimensions (a valid and construct central explanation, but one which needs to be understood and clarified by test developers). Shepard (1982) emphasized that "the assumption of unidimensionality underlying all of the (DIF) methods is not merely a statistical prerequisite, but it is central to the way in which item bias is defined" (p. 25).

Differential item functioning procedures are designed to identify differences in item responses due to construct-irrelevant variance, with the understanding that these differences will be subsequently minimized by reviewers once the problematic items have been identified. Two steps in the DIF procedures are required to address the questions of systematic differences and relevant or irrelevant causation. The first step is a statistical step, of which three primary methodologies appear to be used most often: the Mantel-Haenszel, the standardization technique, and the item response theory (IRT) procedure. Readers are directed to Dorans and Holland (1993) and Thissen, Steinberg, and Wainer (1993) for an explanation of these three techniques.

While these descriptions focus on multiple choice data, the basic methodologies also have been adapted to be used with graded scoring scales (e.g., score responses of open-ended items) (cite). Each of the statistical techniques results in an estimate of differential item functioning between a focal group (e.g., African American or women), and a reference group (e.g., Whites or men). The higher the DIF result, the more unequal the functioning between groups for that item (after taking into consideration the variance due to the criterion test score).

The second step is judgmental — attributing medium to large DIF results to a source. This is a necessary and extremely important aspect of the DIF procedures, but one where unfortunately there is still a great deal that is not known. Currently, there are not many replicable, systematic links between high DIF and specific causal judgements, which has led to far fewer general understandings than expected. This, in turn, has stunted the ability of test developers to develop and implement substantive improvements in test construction in an attempt to minimize unintended consequences for groups (Linn, 1993).

The following discussion highlights three issues related to DIF and accessibility of tests for English language learners that might affect the use and consideration of these results for LEP students: 1) the question of bias, 2) pervasive bias, and 3) factors which explain differential functioning.

Psychometricians realize that these procedures do not prove or disprove bias.

The Question of Bias

Psychometricians realize that these procedures do not prove or disprove bias. Instead, these statistical techniques detect items that function differently for different groups. It is also important that "this difference indicates that the identified item does not appear to measure the same construct as the total test" (Dorans and Holland, 1993, p. 61). As stated above, a high DIF could indicate bias, unequal educational access, or a violation of unidimensionality.

Readers need to remember:

- High DIF doesn't necessarily mean bias, and low DIF does not preclude it.
- While item content can present problems because of unfamiliarity or multidimensionality, bias problems might include one or more of the issues identified in this *Guide*.

Unfortunately, these problems typically are not addressed, even in bias reviews which usually have a rather narrow charge. Particularly at issue are content problems in terms of cultural access, whether they are due to prior experiences or differential perceptions, and also possible values related to construct-central or contextual issues, such as events and/or objects. These problems may also include test construction or implementation process issues which could be pervasive, or could be item-specific. For example, a specific item may have a confusing visual or lack important access to tools or resources.

Pervasive Bias

As noted at the outset of this section, a key constraint of DIF is that the items do not detect pervasive bias, or bias over items (Linn, 1993). Differential item functioning statistical methods assume that the test from which the questionable items are flagged is valid and unbiased for all of the focal groups. Several authors in Holland and Wainer (1993) explain the pros and cons of using the test as the criterion to detect differential functioning of items, but nothing better has been identified to replace it.

Obviously, the perspective in this *Guide* is that many of the assessments in current use are not valid for many students, including LEP students. While the items may attempt to measure the constructs which have been identified as important and central, the "hows" of the test's presentation (including language, contextual issues and test format), and the test's administration constrain the ways in which students are allowed to demonstrate their knowledge. These confounding processes substantively and significantly block student access to the constructs the tests are attempting to evaluate. Not infrequently, these barriers occur over a large portion of the items.

The barriers may be processes which affect the entire test, such as requiring a written response from students as the only acceptable response, or they may occur for one or five items such as those which share a contextual passage. However, bias is often compounded over the type of problem. That is, students who have difficulty with the complexity of English text on tests will be affected by such problems as amount of text, complex sentence and language structure, administration constraints, lack of visuals, and possibly access to tools which they can use to compensate for this limitation. Some researchers have also suggested that aggregate issues tend to affect student access exponentially (Farr and Trumbull, 1997). Students simply shut down because of fatigue, disgust, and/or confusion.

However, there is a place for DIF methodologies. The strength of the DIF techniques is that they address bias item-by-item, which is consistent with our approach in this *Guide*. If tests are made more valid, then these procedures will identify items which continue to deviate for some reason beyond the types of pervasive problems which the improved tests have alleviated or minimized. In fact, the DIF analyses can be used routinely to confirm if the recommendations discussed in this guide are effective.

Factors Which Explain Differential Functioning

Linn (1993) suggested that researchers may learn more about factors which contribute to DIF when they examine other types of information beyond or in conjunction with the typical group memberships DIF addresses. These factors could help educators design better learning situations, e.g., if items that focus on unequal access to educational opportunities show systematic functioning over types of school programs, educators can take action to improve those programs. They also might allow DIF methodologists to fine-tune their statistical procedures to allow more causal types of determinations to be built into the statistical analyses.

As noted previously, Linn (1993) reflected that "far fewer general principals about test construction have been derived as the result of DIF analyses than most researchers expected. The majority of items with large DIF values seem to defy explanation of the kind that can lead to sounder test development practice. More often than not judges have a rather poor record of predicting which items will or will not be flagged" (p. 358). Without additional help of the type offered in this *Guide,* substantially reducing bias problems or problems which inflate systematic irrelevant variance for subgroups has remained difficult. Some types of factors, like school programs, are important for policymakers and researchers to know, but these factors do not affect the development of the assessment. Some types of construct irrelevant issues do, such as linguistic and cultural considerations. Recommendations in the *Guide* can serve as a framework to define questions judges can ask in order to determine causes of high DIF.

In addition to having a positive effect on test construction, the issues we discuss in this *Guide* might be used to guide future refinements in the DIF statistical procedures. Perhaps a second set of statistical analyses can be done before the statistical findings are given to judges. The second set can narrow down the determinants by taking a meta-approach over items and looking for systematic concerns or trying to sort items differently. Many of the issues surrounding the fair testing of LEP students share common ground with concerns of other populations, such as students with disabilities. Some are more intransient for LEP students who come from cultures very different than mainstream U.S. culture. However, these concerns might be shared by some students who live in "non-mainstream" settings in the U.S. and do not have language minority status. Statistical data that uncovers these issues could be very helpful to judges, and also would improve the ability to use DIF effectively.

Comparability Issues and Accessibility

Currently, a greater range of students are expected to take the same large-scale tests taken by mainstream students.

Systems for determining comparability within and across tests have been identified and used for years (Linn and Haertal, 1995; National Research Council, 1999). Comparability issues recently have been intensified because of legislative and policy mandates which specify that broader populations of students (including a greater range of LEP students and students with disabilities) are expected to be included in the administration of and reporting from mainstream assessments. This legislative push occurs at a time when troubling discrepancies between traditional theoretical underpinnings in psychometrics and recent gains in cognitive learning theory are evident. Traditional comparability solutions must be clarified and possibly updated in the face of these issues and competing validity, generalizability, and accessibility needs.

Traditionally, standardization has been employed in academic testing because of the belief that providing the same conditions, including the same or "parallel" content, and the same format, administration, and response procedures for all students was the best way to assure dependable, equable results within and across the students. Students who had special challenges or who were outside mainstream educational programs were often waived out of the tests and not included in norming or test development samples. Furthermore, as Nancy Cole (1993) pointed out, prior to the 1970s, "the test community's technical orientation was color-blind. Its procedures did not take racial-ethnic identity into account in any explicit, intentional way, and the community's implicit definition of fairness was essentially to be color-blind — this at a time when the world around it was increasingly color conscious" (pg. 26). Since that time, the development of differential item functioning procedures and the addition of bias reviews in test development have formed the bulk of how academic measurement has updated its response to ensure that standardized testing is fair and comparable.

Currently, a greater range of students are expected to take the same large-scale tests taken by mainstream students. How does one make these tests equitable, when certain challenges of access cannot be ignored? Were the traditional tests actually equitable before, given that we now know that all students bring different strengths and challenges to the work of accessing items, processing information, solving problems, and articulating responses?

The ability to equate scores and generalize across students and subject area domains or constructs is certainly essential in large-scale testing. However, we should be troubled by the assumption that we arrive at "equable" information in large part because we impose uniform conditions and materials. Some work has been done which expands traditional notions of standardization in testing. The use of rubrics in large-scale testing has demonstrated the viability of standardizing constructs and evaluation processes, rather than specific responses, as a way of achieving comparable results. Computer-adaptive testing illustrates that students can take different combinations of items and receive parallel scores.

Perhaps it is time to remember the goals of testing. The standards movement of recent years has specified what students are expected to learn. Our paramount goal should be to understand what students know and can do in the context of the standards or other specified domains. We also want to understand student mastery in a way that gives us reliable information so that we can compare many students to those standards and/or so that we can aggregate and compare over students. Our particular challenge is to not lose sight of the goal in our quest to generalize over our populations of test-takers.

If we hold constant the notion that we expect to get a reasonably accurate reading of student mastery for each and every student who takes a given assessment, we should be able to rethink how and what to standardize in order to achieve high-quality results in large-scale testing. Perhaps some processes or materials should remain exactly the same, and perhaps some might include a limited set of options from which to choose, such as rubrics or computer-adaptive testing. When comparability first was conceptualized, we looked at many technical aspects in virtual isolation from others. Now that we have used standards-based standardized assessments for a while, we believe that our understanding of the best solutions of many of the technical issues are actually determined through a series of tradeoffs.

For instance, we know that changes as simple as altering the order of items within a test affect score results. However, item order differences have been deemed less important than several competing concerns, including a need to matrix different items over forms to provide more reliable group information about standards or constructs.

Perhaps we can develop a solution which recognizes diverse access, processing, and response demonstration needs, as well as generalizability needs. We should expect both validity and reliability for all who take the test in order to be accountable to the public. In so doing, we must move forward to find the next generation of comparability solutions.

This *Guide* should provide useful suggestions from which to build a flexible testing system. The Council of Chief State School Officers and the U. S. Department of Education sponsored series of meetings to discuss how to achieve meaningful standardization and comparability in large-scale assessments for all students (Martin, Lara, and Kopriva, 1998). The purpose of these meetings was to identify a research agenda which will form an empirical base from which to address comparability issues within the competing needs of standardization and accessibility.

Linking Across Tests

A number of researchers consistently explain the difficulties of linking results across tests so that scores from different assessments on a common scale may be compared. Dilemmas such as different content domains, different amounts of coverage of the domains at the various levels of process and depth, and different test construction procedures and norming samples, present large hurdles to surmount (Sheldon, 1997).

Sometimes it is imperative to administer different assessments. As one example, Kentucky has pioneered the use of the alternative assessment for those students whose curriculum is functional rather than academic (Dings, 1997). While some advocate that using primary language assessments solves many issues for LEP students, these assessments are only useful when students are fully literate in their home language and when their schooling in the subject to be tested has occurred in that language (Solano-Flores, 1998). However, room should be made in any testing system to accommodate those students who have been schooled and are literate in their home language.

As a general rule, some testing experts conclude that approximately 98 percent of the student population can be evaluated in a *flexible* testing system, where content domains and levels of coverage are roughly parallel (Reidy, 1998). What constitutes "flexible" is the issue. Nonetheless, it is clear that what distinguishes the 98 percent from the two percent seems, for the most part, to be an issue of differing domains or standards. Therefore, students who are in a functional rather than academic curriculum should take a different assessment. But students who are expected to meet the same academic standards as mainstream students, even though they come to school with diverse challenges or experiences, should be accommodated within the same mainstream system.

As a general rule, some testing experts conclude that approximately 98 percent of the student population can be evaluated in a flexible testing system, where content domains and levels of coverage are roughly parallel

One primary implication is that primary language assessments should be built either from the same test domains or standards specifications as the ones written for the mainstream assessment, or they should pass rigorous reviews to determine a high degree of alignment with both the academic standards and test specifications of the mainstream assessment. They should be seen as part of the flexible system, not adjunct to it.

Construct Validity and Equity

The *1999 Standards for Educational and Psychological Testing* (Joint Committee of APA, AERA, NCME) advocates evaluating validity through streams of evidence. These streams seek to demonstrate the viability of measuring content- and use-specific constructs. Instead of referring to different types of validity (such as content, criterion, or predictive validity), the score inferences which are constrained within and defined by the evidence suggest a degree of construct-validity confidence.

Typically, this confidence is determined through careful test construction and implementation procedures, and through post hoc studies. Certainly it is important to ensure careful development and implementation procedures, but to date, no guidelines or analytic framework have been developed which systematically determine what it means to be careful and how careful is careful enough. Procedures have been routinely adopted and refined, but there is no guidance that tells us which of these procedures are sufficient and necessary to ensure construct validity.

Construct validity guidelines should have a framework to determine the measurement intent of items and item-composites, or tests. This should include information about what different item types measure and under what conditions. These guidelines also should include an evaluation of the accessibility of each item for diverse populations, both in terms of requirement access and response access. Construct validity guidelines should insist on multiple points of evidence during the development and implementation phases, aggregated in some manner to determine a level of validity. In addition, some type of system might be outlined which requires some steps and which allows for choice at other junctures.

Virtually no formal analytic work has been done to understand how to create large- scale mainstream assessments from the ground up that accurately reflect what LEP students know and can do. We need to define a research agenda that systematically analyzes the relationships between student responses and item/test scores, and between types of items under specific conditions. Undertaking studies that seek to understand what elements in the tests provide barriers for specific students, for what reasons, and what can be done to alleviate these barriers, will provide important construct-validity evidence.

Appendix B

Highlights of Key Research on Assessment of LEP Students

Much of the empirical work on improving large-scale assessments for LEP students focuses on test accommodations. Most of these studies have noted that the items and tests for LEP students have pervasive validity problems. Researchers cite language misunderstandings, inaccurate assumptions that pervade items and rubrics, and accessibility problems which are not eliminated under current test development procedures.

Malcolm (1991) encourages a different approach to assessments: Rather than trying to avoid bias, she suggests that assessments support equity from the ground up. Specifically, she asks what conditions have to be incorporated to construct an equity-supported assessment, and makes the following suggestions:

- Rules about what is to be known must be clear;
- Resources which allow students to access what is required and demonstrate what they know must be available;
- Ways of demonstrating knowledge must be accessible, multiple, and varied;
- Concepts valued by different groups must be reflected and respected within the items and carried through into test administration and analysis.

Rather than trying to avoid bias, she suggests that assessments support equity from the ground up.

Linda Darling-Hammond (1994) cautions that there are many forms of bias, which include:

- How items are built and chosen for inclusion in the assessment;
- How and what responses are deemed appropriate;
- What content and context is deemed important;
- How weighting on tests is accomplished to achieve alignment.

There is considerable discussion, but limited empirical work, related to other test considerations, including:

- Reducing the literacy load of items;
- Evaluating items and rubrics more extensively for considerations related to LEP students;
- Cultural effects of testing on LEP students; and
- Handling context and literacy concerns in an appropriate manner.

For instance, several experts, including Saville-Troike (1991) and Shaw (1997), focus on the need to "plain language" items and formats so that unnecessary vocabulary and structural complexity can be reduced or eliminated. This issue has been researched by Abedi (1995, 1997) and Hanson, Kopriva, Brooks, and Saez (1996). Hampton (1996), having noted similar problems in a national language arts test, began using a technical writer to reduce unnecessary language complexity in the items.

Malcolm (1991), Shannon (1996), and Wiley, Kopriva, and Shannon (1997) discuss broader approaches to evaluating items which include a systematic investigation of student work for diverse populations. This includes "acceptable" outlier responses which would be captured and, where appropriate, included in training examples for teachers to illustrate the range of appropriate responses (Malcolm, 1991, p. 397). Shaw (1997) encourages extending ways in which students are allowed to respond by specifically stating acceptable modes of written response within items, (e.g., charts and drawings), and by evaluating scoring rubrics to determine whether students are given proper credit for non-standard modes of response.

Several authors have emphasized the pervasive effects of culture and diverse experiences on items, particularly when students are from countries or regions where values or experiences differ significantly from those of the majority culture in the United States. For example, Leap (1988) and Rothmen (1993) discuss cultural values found in native American communities which affect how students interpret and respond to test items. Heath (1986) and Farr and Trumbull (1997) discuss similar issues with rural African American and foreign-born populations.

Some writers have suggested additional approaches. Gordon (1992) encourages the use of item choice to enhance validity, and Farr and Trumbull recommend that item choice should include "tailoring items" so that they measure the same construct but are presented in different contexts which reflect common experiences of different cultural or linguistic groups. Shaw (1997) presents some key recommendations for improving performance assessments for LEP students:

Several authors have emphasized the pervasive effects of culture and diverse experiences on items,

- Using team-building activities for groups;
- Using effective visuals;
- Providing information in the same format throughout the test;
- Including 20 minutes of teacher orientation for every 2 hours of assessment administration;
- "Plain languaging" items;
- Explaining non-central vocabulary; and
- Minimizing the numbers of items at any one sitting.

Accommodations and Item Development: Considerations for Equity

To date, most of the accommodations literature has focused on accommodations provided during test administration. Little research has been done on student response options, primarily because of issues of cost. Some work has focused on modes of response on paper-and-pencil tests that extend beyond written responses per se (e.g., using diagrams, pictures, or performances). Other investigations have begun to address how items might be presented to LEP students so that items will be more accessible to them.

Throughout their book, Farr and Trumbull (1997) suggest several excellent assessment administration procedures, including some that could prove useful in large-scale situations. As one example, they discuss the need to be direct, clear, and specific about how to use additional time, as do Hiebert and Calfee (1990). These researchers emphasize that while all students should be given specific directions about how best to use their time during any assessment, this specificity is especially critical to LEP students.

Farr and Trumball emphasize the need to clarify vocabulary, context, and item requirements. Test items often are unintentionally linked to mainstream American experiences, vocabulary, and discourse practices. While the manner in which items are presented on a test is partly an item-development issue, it also affects how tests are administered. Key decisions which need to be made during the development process include:

- Using dictionaries or glossaries during the test-taking process;
- Putting items in a context that is more familiar to the students;
- Reading contextual passages aloud;
- Reading or re-reading directions or the complete text of items aloud, in English or in the students' first language; and
- Simplifying the English that is not central to the content being evaluated.

Some researchers are investigating these issues. Kopriva (1994) and Kopriva and Lowrey (1994) researched oral and written administration of mathematics tests in English and in Spanish for students from a number of home countries and whose proficiency in English spanned a wide range. Abedi and other researchers at CRESST (the National Center for Research on Evaluation, Standards, and Student Testing, 1998) currently are involved in a set of studies focused on the validity of accommodations and modifications in assessments. Hafner (1998) is engaged in a study that explores the use of extended time and the use of dictionaries for LEP students.

Allowing students to respond in written form and in their home language, has been informally and formally investigated (for instance, see Daro, 1996; Solano-Flores, 1998; Kopriva and Lowrey, 1994; Abedi, 1998; CCSSO, 1998). While this accommodation seems straightforward, the findings have been similar to Shaw's (1997) experience: If students are literate (at grade level) in their home language, this accommodation is appropriate and can be scored by bilingual experts. Otherwise, written response in the student's home language presents some major issues. Often students responding in written Spanish lack written proficiency in the language. Shaw's study revealed that student responses had to be translated into "comprehensible Spanish" to be scored by Spanish scorers. If the responses are in English, scorer training is imperative — and this might adhere also for scorers evaluating work that appears to be in the students' first language.

The use of oral student responses are discussed throughout Farr and Trumball. While this might be a reasonable accommodation in large-scale assessment, the issues of cost and time need to be addressed. The state of Delaware is investigating the use of voice recognition technology — where students speak into computers and their responses are printed (Kopriva, Brooks, and Saez, 1996). This approach may offset many of these concerns.

Scoring Considerations for LEP Students

In scoring the open-ended responses in the Kopriva and Lowrey (1994) study, researchers and bilingual scorers noted that the responses of LEP students were often non-standard and not considered in the regular training scorers received from the test publisher. For those reasons, responses could be easily misread by scorers in large-volume testing situations.

The training that scorers receive typically does not include information particular to LEP students' responses. In addition, scorers are almost exclusively monolingual English speakers who are expected to score a large number of papers in a short time. Daro (1994) noticed informally that LEP responses tended to be scored "towards the mean." He suggested that the scorers appeared afraid to under- or over-read the responses from LEP students, and recommended more structured guidance to aid scorers in properly evaluating student work.

The Council of Chief State School Officers (CCSSO) and the U.S. Department of Education convened a study with NAEP items which investigated the legitimacy and usefulness of training scorers in how to score responses from English language learners (Kopriva, 1997). The content of their training was similar to what Shaw found to be important (1997). That is, scorers should be familiar with both linguistic and cultural issues that relate to particular language groups taking the test. They also need to be well-versed in language development issues and transitional expressions of English literacy. Kopriva found that a brief training made a difference in scoring LEP papers and types of non-standard responses. Delandshere and Petrosky (1994) emphasized the broad variation in responses, especially within groups whose experiences were different from the majority U.S. culture. Wiley (1992) and Shepard (1991), among other measurement experts, stress the importance of allowing multiple paths to successful performance for all students. This appears to be particularly important for some population groups, including LEP students.

References

Abedi, J. et al. (May 1995). *Language Background as a Variable in NAEP Mathematics Performance: NAEP TRP 3D–Language Background Study.* University of California, National Center for Research on Evaluation, Standards, and Student Testing, Technical Review Panel for Assessing the Validity of the National Assessment of Educational Progress.

August D., and Hakuta, K. (Eds.) (1997). *Improving schooling for language minority students.* Washington, DC: National Academy of Science.

Berman P. et al. (1995). *School reform & student diversity: case studies of exemplary practices for LEP students.* Washington, DC: National Clearinghouse on Bilingual Education.

Cole, Nancy (1993). History and development of DIF. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning,* Hillsdale, NJ. Lawrence Earlbaum, (pp. 25-30).

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

Cummins, J. (1979). *Cognitive /academic language proficiency, linguistic interdependence, the optimum age question and some other matters.* Working papers on bilingualism, 19:197-205.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review,* 64, (1).

Daro, P. (*Personal communication,* May, 1994.)

DeAvila, E. (*Personal communication,* September, 1997.)

Delandshere, G. and Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment. *Educational Researcher,* 23, (5), 11-18.

Dings, J. (1997). Paper presentation to Title 1 Comprehensive Assessment System Project, CCSSO, Washington, DC.

Dorans, N.J. and Holland P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning,* (pp. 35-66).

Farr B.P. and Trumbull, E. (1997). *Assessment alternatives for diverse classrooms.* Norwood, MA: Christopher-Gordo Publishers, Inc.

Gordon, E.W. (1992). *Implications of diversity in human characteristics for authentic assessment.* CSE Technical Report 341. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Haertel, E. H. and Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. Mislevy, and I. Bejar (Eds.) *Test theory for a new generation of tests.* Hillsdale, NJ: Earlbaum, pp. 30-59.

Hafner, A. (1999). Assessment accommodation that provide valid inferences for LEP students. Paper presentation at CCSSO's large-scale conference, Salt Lake City, UT.

Hampton, S. (*Personal communication,* April, 1996.)

Hansche, L. (1998). *Handbook for the development of performance standards: meeting the requirements of Title I.* Washington, DC: U.S. Department of Education.

Hanson M., Hodgkins, M., Kopriva, R., Brooks, M., and Saéz, S. (1996). *Which options work the best? a qualitative inquiry into accommodations for learning disabled students.* Technical Report, Delaware Department of Instruction. Dover, DE.

Heath, S.B. (1986). *Beyond language: social and cultural factor in schooling language minority students.*

Heubert, J. and Hauser, (1999). *High stakes: testing for tracking, promotion and graduation.* Washington DC: National Academy of Science.

Hiebert, E. and Calfee, R. (1992). Assessment of literacy: from standardized tests to performances and portfolios. In S. Samuels and A. Farstrup (Eds.), *What research has to say about reading instruction (2nd edition).* Newark, DE: International Reading Association, pp. 165-174.

Hodgkins, M. (1997). *Expanded bias reviews: recommendations from the field.* Technical Report, Inclusive Comprehensive System. Dover, DE: Delaware Department of Public Instruction.

Joint Committee (1999). *Standards for Educational and Psychological Testing.* Committee sponsored by American Psychological Association, American Educational Researchers Association, and the National Council for Measurement in Education, Washington DC.

Kopriva, R.J. and others (1999). *A conceptual framework for the valid and comparable measurement of all students.* Unpublished Technical Paper, Washington DC.: Council of Chief State School Officers.

Kopriva, R.J., (1997). *Making large-scale assessment accessible for LEP students.* Unpublished report. Washington, DC.: Council of Chief State School Officers.

Kopriva, R.J. (1994). *Validity issues in performance assessment for low, mid, and high achieving ESL and English only elementary students.* Technical Report. Sacramento, CA: California Department of Education, California Learning Assessment System Unit.

Kopriva, R.J., Brooks, M. and Saéz, S. (1996). *Inclusive comprehensive assessment system: annual report to OERI.* Dover, DE: Delaware Department of Public Instruction,

Kopriva, R.J. and Lara, J. (1998). Scoring English language learners' papers more accurately. *Science education reform for all.* Y.S. George and V.V. Van Horne (Eds.), American Association for the Advancement of Science, 77- 82.

Kopriva, R.J. and Lowrey, K. (1994a). *Investigation of language sensitive modifications in a pilot study of CLAS, the California Learning Assessment System.* Technical Report. Sacramento, CA: California Department of Education, California Learning Assessment System Unit.

Kopriva, R.J. and Lowrey, K. (1994b). *Development of CLAS for spanish speakers: interim report.* Sacramento, CA: California Department of Education, California Learning Assessment System.

Kopriva, R.J. and Saéz, S.M. (1997). *Guide to scoring LEP student responses to open-ended mathematics items.* Washington DC: Council of Chief State School Officers.

Kopriva, R.J. and Sexton, U.M. (1999). *Guide to scoring LEP student responses to open-ended science items.* Washington DC: Council of Chief State School Officers.

Kopriva, R.J., Wiley, D.E., and Schmidt, W. (1997). *Building a construct validity technology for standards based assessments.* Research proposal submitted to OERI, TIMMS Center. East Lansing, MI: Michigan State University.

Koretz, D.M., Bertenthal, M.W., and Green, B.F. (1999). *Embedding questions: the pursuit of a common measure in uncommon tests.* Washington, DC, National Academy of Science.

Lara, J. and August, D. (1996). *Systemic reform and limited English proficient students.* Washington, D.C.: Council of Chief State School Officers.

Leap, W. (1988). *Assumptions and strategies guiding mathematics problem solving by Ute Indian students.* In Cocking, R. and Mestre, R. (Eds.), Linguistic and cultural influences on mathematics, pp. 85-107.

LeCelle-Peterson, L. and Rivera, C (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review,* 641(1), 55-75.

Linn, R. L. (1993). The use of differential item functioning statistics: a discussion of current practice and future implications. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning,* Hillsdale, NJ: Lawrence Earlbaum, pp. 349-364.

Lipski, J.M. (1985). *Linguistics aspects of Spanish-English language switching.* Tempe, AZ: Center for Latin American Studies.

Malcolm, S. (1991). Equity and excellence through authentic science assessment. In E. Kulm and S. Malcolm (Eds.), *Science assessment in the service of reform*, pps. 313-330. Washington, DC: American Association for the Advancement of Science.

Olson, J.F. and Goldstein, A.A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments.* Washington, DC: National Center for Education Statistics.

Reidy, E. (*Personal communication*, April, 1998.)

Resnick, L. (1997). Symposium on new standards implementation of standards-referenced assessment. Presentation at National Council for Measurement in Education Annual Meeting, Chicago, IL.

Rivera, C. (*Personal communication*, April, 1998.)

Rothman, R. (1993). ACT unveils new assessment, planning system. *Education Week*, XII, 24, 1, 17.

Saéz, S.M. (1994). *Testing LEP students in Delaware.* Presentation to the LEP Consortium Project, Council of Chief State School Officers, San Francisco, CA.

Saville-Troike (1991). *Teaching and testing for academic achievement: the role of language development.* National Council of Bilingual Education, No. 4. Arlington, VA: Center for Applied Linguistics.

Shannon, A. (1996). *Developing mathematics items: new procedures.* Technical Report. Pittsubrgh, PA: New Standards Project.

Shaw, J. Reflections on performance assessment of English language learners. In B. Farr and E. Trumbull, (Eds.) *Assessment alternatives for diverse classrooms.* Norwood, MA: Christopher-Gordon Publishers Inc., pp. 334-342.

Shepard, L.A. (1982). Definitions of bias. In R. Berk (ed.), *Handbook of methods for detecting test bias*, pp. 9-30.

Slavin, R.E. and Fashola, O.S. (1998). *Show me the evidence!* Thousand Oaks, CA: Corwin Press, Inc.

Solano-Flores, W. (1998). *Using shells: making performances assessments in science accessible for English language learners.* Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Thissen, D., Steinberg, L, Wainer, H. (1993). Detection of differential item functioning using the parameters of the item response models. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Earlbaum, pp. 67-114.

Thurlow, M.L. (1998). *Snapshot of large scale testing accommodations across the U.S. for students with disabilities.* Paper presented at the American Educational Researchers Association Annual Meeting, San Diego, CA.

Thurlow, M.L., Elliott, J.L. and Ysseldyke, J.E. (1998). *Testing students with disabilities.* Thousand Oaks, CA: Corwin Press, Inc.

Wiley, D.E., Kopriva, R.J. and Shannon, A. (1997). *Standards-based validation of performance assessments.* Technical report. Pittsburgh, PA: New Standards Project.

Wong-Fillmore, L. (1994). The role and function of formulaic speech in conversation. In Grimshaw and D. Jennes, (Eds.), *What's going on here? complementary studies of professional talk.* New York: Ablex Publishers, pp. 60-93.

Council of Chief State School Officers
One Massachusetts Avenue, NW
Suite 700
Washington, DC 20001-1431
Telephone: 202-408-5505 • Fax: 202-408-8072
www.ccsso.org

CHIEF STATE SCHOOL OFFICERS · COUNCIL ·
CCSSO

FL026759

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Ensuring Accuracy in Testing for English Language Learners	
Author(s): Rebecca Kopriva	
Corporate Source: Council of Chief State School Officers	Publication Date: 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Cynthia G. Brown	Printed Name/Position/Title: Cynthia G. Brown, Director	
Organization/Address: Resource Center on Educational Equity Council of Chief State School Officers One Massachusetts Ave. NW - Washington, DC 20001	Telephone: (202) 336-7007 E-Mail Address: cindyb@ccsso.org	FAX: (202) 408-8072 Date:

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com