



Assessing English-Language Learners' Achievement

Author(s): Richard P. Durán

Source: *Review of Research in Education*, Vol. 32, What Counts as Knowledge in Educational Settings: Disciplinary Knowledge, Assessment, and Curriculum (2008), pp. 292-327

Published by: American Educational Research Association

Stable URL: <https://www.jstor.org/stable/20185119>

Accessed: 25-03-2020 10:35 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

American Educational Research Association is collaborating with JSTOR to digitize, preserve and extend access to *Review of Research in Education*

Chapter 9

Assessing English-Language Learners' Achievement

RICHARD P. DURÁN

University of California, Santa Barbara

Assessment of learners' academic achievement in a second language presents important challenges to the fields of educational research and educational practice. Although these challenges legitimately concern learners' familiarity with a second language, the challenges are more complex, particularly in the contexts of large-scale assessments that are intended to hold schools accountable for what students know and can do on the basis of their performance on assessments. This chapter presents a synopsis of major trends and issues in this regard involving large-scale assessment of English-language learner students (ELLs) in the United States in core achievement areas. These students are students from non-English backgrounds who are evaluated by schools as not knowing sufficient English to benefit fully from instruction in this language and who are eligible for receipt of educational support to acquire greater English proficiency. Although the precise numbers of these students in the U.S. population cannot be determined for reasons discussed in this chapter, they have been estimated to number approximately 4.5 million and to constitute about 8% of all students in the K–12 grade range; about 80% of these students are from a Spanish-speaking background (Zehler et al., 2003). After a discussion of research trends and issues in the main body of the chapter, the concluding portions of the chapter suggest development of an alternative foundation for assessments that provide more valid information about the learning capabilities and achievement of ELLs. This section also presents an example of how one might pursue enriched assessment of ELLs in the context of classroom activities concerned with acquisition of an important class of academic English skills.

The focus is on assessment of ELLs in U.S. contexts, but the issues raised are pertinent to ELLs in broader international contexts and to learners assessed in second languages other than English. Attention to U.S. contexts, and in particular, the impact of the No Child Left Behind (NCLB) Act of 2001 on ELL assessment, highlights how in one country, the United States, policy-motivated attempts to strengthen assessment of

Review of Research in Education

February 2008, Vol. 32, pp. 292–327

DOI: 10.3102/0091732X07309372

© 2008 AERA. <http://rre.aera.net>

ELLs is tied to resolving better fundamental questions about who ELL students are, what expectations are held about their schooling achievement, how English language proficiency (ELP) affects or mitigates assessment and schooling performance, and what research base exists to answer these questions and inform new directions for linking assessment and schooling practices furthering the education of these students. Similar questions and their resolution are relevant to other international contexts, particularly in industrialized nations that face a rapid growth in their immigrant populations from non-industrialized portions of the world. Just as in the United States, these nations encounter increasingly multicultural and multilingual growth in their population and a need to improve schooling outcomes for all residents in a country (Suarez-Orozco, 2001).

With this broader set of implications as a backdrop, the chapter overviews findings related to improving assessment of ELLs in the United States at the federal, state, and school levels and the ways in which federal and state policies under the NCLB law have affected concern for inclusion of ELLs in assessments. Accordingly, the focus is on how the policies of a particular law affect assessment practices required by states and their school districts and schools and how this affects the validity of assessments administered to ELLs. Attention to large-scale assessments under NCLB is of high interest from an assessment and learning perspective in that the target of state assessments under NCLB is what students are expected to learn in an entire year of schooling in a subject matter given state standards for this expectation. This is very different from attention to how to assess isolated learning and problem-solving skills better where attention can be focused solely on assessing learning that does not bear any intended relation to other kinds of learning expected of students. In contrast, the former is about capturing a wide range of problem-solving skills and knowledge that are interrelated as components of an intended curriculum at a grade level across an entire year and, beyond that, that are capable of showing evidence of a progression in development of skills and knowledge across grades. As will be appreciated, the ensuing discussion blends concern for interpretation of policies with research and assessment issues. This back-and-forth blending of policy and research issues is deliberate. Consistent with current research on education reform and school accountability and assessment, we need to examine carefully how the theories of action surrounding educational policies are tied to research and assessment problems and issues (Baker & Linn, 2004).

Key issues drawing attention in this discussion include (a) who gets identified and assessed under the status *ELL*; (b) how ELLs are included in assessments, including with assessment accommodations and alternative assessments; (c) the performance of ELLs on assessments and associated evidence of measurement reliability and validity; and (d) the emergence of high-stakes assessments, such as high school exit assessments, and their implications for the assessment of ELL students. In addressing the foregoing questions, attention is also given to whether it is realistic to expect extraordinary congruence in ELL population definitions and assessment practices across states, and between states and the federally based national large-scale assessments. The tension between states' rights to govern assessment of students under their individual education laws and federal concerns is such that there is a question about the trumping of

federal goals for coherent assessment practices at the national level by the idiosyncratic policies of individual states. Although this may seem unsound, at first, on scientific measurement and nationally coherent policy grounds, the upshot is that the privileging of states to decide how they best include and assess ELL students may have a validity advantage. Allowing states to assert their rights on how to best identify ELL students may actually hold them more accountable, in the ideal and in the long run, for being sensitive to the particular characteristics and needs of the students they define as ELL within their jurisdictions. This is especially the case given the heterogeneity among ELLs in different regions and states in the country and the policies and resources in unique state and school district jurisdictions.

The final sections of the chapter shift attention to the inherent limits of large-scale assessments as accountability tools for ELLs as a means for directly informing a deep understanding of students' learning capabilities and performance that can be related to instructions and other kinds of intervention strategies supporting ELL schooling outcomes. This critique is not intended to imply that existing large-scale assessments of ELLs are uninformative. The existing results of such assessment clearly show that ELLs often have serious academic content learning needs. But assessments can be designed to work better for these students, if we take care to have assessments do a better job of pinpointing skill needs of students developmentally across time, better connect assessments to learning activities across time and instructional units, and better represent the social and cultural dimensions of classrooms that are related to opportunities to learn for ELL students. This final portion of the chapter starts with a concrete example for how to go about improving assessments for ELLs by redefining how assessments are conceived as tools for evaluating learning goals in the area of learning of academic English. This discussion adopts an activity theory perspective. The ensuing discussion is applicable at large to students, but focus on students whom we label as ELL in this chapter sharpens the concern for the ways that students' background, talents, and social practices need to be examined more closely with regard to how these affect the design and implementation of assessments that might guide instruction given classroom and schooling practices and community experiences.

As will be suggested, (a) students' proficiency in a subject matter cannot be captured adequately by one-dimensional constructs of academic competence, such as those operationalized by existing large-scale assessments; (b) large-scale assessments can at best provide "thin" coverage of what students know and can do given students' background; and (c) current research on the social and cultural nature of learning contexts and cognitive and interaction studies of students' interaction in learning settings suggest new ways to design assessment tools that are different from traditional tests and assessments. In closing, the chapter discusses, once again, how opening the question of what educational achievement goals should be can create opportunities for revitalizing the field of educational assessment of ELL and other students.

The issues discussed go beyond suggesting that large-scale assessments should not be expected to perform the function of local teacher-developed classroom assessments intended to guide instruction in a differentiated and formative manner and

that development of the latter represents a solution to making existing assessment approaches more valid for ELLs. Classroom instructional and learning practices, and expectations about students' learning goals, are affected by institutional and cross-institutional values, educational policies, and practices that often work at odds with each other as well as in mutual support. As will be discussed, we need to step back and examine what we value and mean by *learning* altogether and what we expect of students as evidence that they have learned, given their developing skills and knowledge and unique linguistic, social, and cultural history.

ASSESSMENT VALIDITY

Concern for assessment validity is central to the chapter. Assessment validity is the most fundamental question underlying the construction and use of assessments. The fundamental question of validity asks, Does an assessment measure what it is supposed to measure given the purposes of an assessment? The question of assessment validity centers on what inferences may be drawn from assessment performances given the objectives, design, and implementation of an assessment; see Messick (1989) for a now-classic discussion of this contemporary view of validity. Assessment validity, according to contemporary accounts, is inherently about arguments used to support inferences—in the case of this chapter, what ELL students know and can do on the basis of how they perform on assessments. These arguments, in principle, must involve both conceptual and empirical warrants. Conceptually, the rationale and documentation for the purpose of an assessment, its design and scoring, and score interpretation should offer an educational explanation of what students know and can do in an achievement area and how these conclusions are linked to empirical evidence and statistical or qualitative data interpretation of test performance and their psychometric consistency or reliability. According to validity theory, there should be a tight connection between what an assessment is expected to measure on conceptual grounds and empirical performance data that could support assessment users' claims about achievement competence targeted by an assessment.

Arguably, the historical development of modern assessment validity theory surrounding achievement and ability testing has not shown adequate sensitivity to ways that the characteristics of individuals and groups interact with the sensibility of target skill and knowledge constructs for assessment and ways that assessments seek to provide evidence for these as constructs as personal, stable attributes of individuals regardless of students' background. Although there is a long history of critiques of assessment design and practice from linguistic and ethnic minority community members citing potential bias in standardized assessments (see Sanchez, 1934, for example), only gradually has this perspective been reflected in specific investigations, as in differential item function (DIF) studies that statistically investigate whether given assessment items are significantly harder or easier for two groups of examinees matched in terms of overall performance on an assessment. Although DIF studies are of value, additional work is also emerging that questions whether the very mathematical-statistical models used to represent skills and knowledge on unidimensional, numeric (interval) scales always makes sense for every population of assessment interest (see Hambleton, Merenda, &

Spielberger, 2005). The usual way to approach these issues from an assessment validity theory perspective is to propose that assessment performance has two components: construct-relevant variance and construct-irrelevant variance. In the assessment of ELLs, the focus of research on validity and reliability of assessments in English has taken this perspective, with emphasis given to the possibility that there might be assessment English-language demands that contribute to construct-irrelevant variance rather than construct-relevant variance on assessments that are not intended to assess English language skills *per se*.

Only rarely, but with increasing frequency, do we see critiques of assessment that question the foundations for how assessments are developed on the basis of item response theory or its earlier antecedent, classical test theory. Some investigators, such as Abedi (2004), propose that the extent of linguistically based construct-irrelevant variance may vary by first-language background and that this needs to be estimated in interpreting assessment scores. Other investigators, for example, Kopriva (in press) and Solano-Flores (2006), propose that psychometric measurement models need formally to incorporate information on how cultural, demographic, and psychological and personality profiles, as well as linguistic factors, affect ELLs' assessment performance. But arguably, these critiques stop short of questioning how assessments and assessment measurement models come to be based on theories of learning and curriculum and how their form, content, and properties as evidence of achievement reflect societal and institutional forces.

Moss, Girard, and Haniford (2006) show sensitivity to these points in a review of contemporary assessment validity theory, proposing a further transformation of the notion of validity so that the concept more closely examines how the target skills and knowledge of an assessment reflect fundamental assumptions about the nature of learning and performance based on the value systems and practices of institutional and policy stakeholders who advocate for and sponsor assessment systems. They suggest a hermeneutic approach where the inferences about the meaning of assessment performances lead to active, ongoing questioning of whether an assessment is really operating as intended. The bottom line here is that validation of an assessment is a process—one might add, a historical, dialectical process—wherein performance evidence from an assessment is examined over time and where the standards for assessments and assessments themselves are improved or redesigned so that they better meet their goals and usefulness to stakeholders.

The hermeneutic approach described by Moss et al. (2006) also addresses the values of assessment stakeholders and the issues of ethics and power relationships between stakeholders as issues that shape assessments and their consequences as part of this dialectical process. Questions that arise in this regard include Who decides what students are expected to know and do? What consequences are there to evidence that students are not attaining competence in target achievement areas? and Who is held accountable for improving the expected educational performance of students?

Why are the foregoing concerns of importance to this chapter? ELLs from certain backgrounds show low achievement performance on large-scale assessments. They are often students from low-income and immigrant backgrounds with parents who have

limited formal education attainment. They are also students who have shown evidence of low schooling achievement and schooling engagement based not just on test scores over time but also based on school record indicators, such as matriculation through required courses, classroom grades, attendance records, drop-out rates, and so on. Large-scale assessments consistently show that ELLs as a whole lag behind other students as a whole in their achievement. The magnitude of achievement lag has been found to be between 0.5 to 2 standard deviations in magnitude on standardized tests and assessments such as the National Assessment of Educational Progress (NAEP) that have been designed in accordance with well-established techniques for designing standardized and criterion-referenced tests common to test developers.

What does this lag in achievement test scores tell us? Is limited competence in English, the language of an assessment, the main factor underlying this gap when an assessment is administered in English? Does it tell us that ELLs know and can do less, given received notions of what an assessment is supposed to measure? Or does it also propel us dialectically to ask more deeply what it is that are we expecting ELLs (and other students) to know and be capable of doing, not just in English but in the wider range of languages and language varieties they may command? Furthermore, will we be served by advancing more comprehensive models for achievement performance and evidence of achievement performance that go beyond the constraints of traditional perspectives underlying the design of existing large-scale assessments? These are among the main questions that motivate this chapter that are tied to a deeper look at validity theory underlying assessments as evidence for what ELLs know and can do, given a federal policy such as NCLB—which putatively frames what ELL students should know and be able to do in achievement areas—and how large-scale assessments at the state level then are expected to serve as conceptually and technically sound instruments to measure students' proficiency in target subject matter areas.

Ultimately, resolution of these issues is itself a historical process tied to the evolution of how assessments will evolve over time to incorporate new research paradigms and findings linking cognitive, sociocultural, and policy-organizational research on assessment to more useful and productive accounts of how assessments themselves might guide assessments as tools to inform educational policy and practice. The aim of the concluding part of this chapter is to suggest some ways that reconstructing how we look at ELLs' (and other students') achievement as "activity" could contribute to such a transformation. If our current approaches to the design and implementation of large-scale assessments are not working well enough for ELLs, what might be a better starting point?

But to create this argument, it is important to get back to why and how we assess ELLs and other students in large-scale assessments currently tied to the policy purposes of assessments and what has been done to address the validity and accuracy or reliability of assessments for ELLs.

LARGE-SCALE ASSESSMENT OF U.S. ELLS WITH A FOCUS ON NCLB

Focus on assessment of ELL students in a U.S. context has several advantages not only because of the range of research and data available to foreground challenges but

also because of the historical evolution of U.S. federal educational policy in the form of NCLB. For states to receive federal assistance under NCLB, states, school districts, and schools are responsible for accounting for the rate at which different subgroups of students show test-score evidence of having attained competence in the subject matter areas of English language arts, mathematics, and more recently, science learning across the K–12 grades. NCLB requires that states implement learning standards systems in reading, mathematics, and science that specify what students are expected to know and be able to do across grades and that serve to indicate whether schools are doing an adequate job in instructing students. State assessment test items and assessments in English language arts, mathematics, and soon, science are constructed from test blueprints or specifications intended to measure students' attainment of standards.

Under Title I of NCLB, states are required to establish and implement adequate yearly progress (AYP) goals for students, including students from key ethnic–racial groups, special education students, and students classified as limited English proficient (or ELL). The law at present requires that by 2013–2014, all students and target subgroups in a state attain 100% proficiency in meeting standards in reading and mathematics in Grades 3 to 8, individually, on the basis of reading and math assessments and assessments administered at least once in high school. States are also required to meet state targets for high school graduation for all students. The area of science is held to similar AYP requirements but with fewer assessments required in Grades 3 to 8.

ELL students with more than 1 year of attendance in U.S. schools must be administered state assessments in English in the three NCLB target subject matter areas, though they have the option of deferring this assessment in English for up to 2 additional years in some circumstances. States have the option of administering ELL students their regular English version assessment under standard conditions or with appropriate assessment accommodations to facilitate ELLs' access to assessment content. Also, states may elect to administer ELLs modified (sometimes referred to as *alternate*) assessments that are not intended to be measure exactly the same constructs in the same way as a regular or accommodated assessment but that can be argued on analytic or empirical grounds to yield information about the same target skills and knowledge.

Accommodated assessments are versions of the regular state assessment that have modified administration conditions facilitating ELL students' (and special education students') access to the information presented on an assessment while, allegedly, not altering the meaning of an assessment performance given the specification or blueprints for items on the nonaccommodated assessment. In general (and not just for ELLs), assessment accommodations involve a change in the setting, scheduling, timing, presentation, or response mode for the standard assessment. The range of accommodations administered to ELLs include, for example, linguistic modification of test items to reduce ELP requirements, dictionaries or glossaries explaining selected construct-irrelevant terms found on assessment items, side-by-side English and non-English native language (L1) presentation of assessment items, oral translation of assessment instructions, or even the oral reading of an item as a whole in an L1 language. Other accommodations, such as

extended assessment time and assessment in small groups, also are available for both ELL and non-ELL students, depending on the state's accommodation policies.

Modified assessments are unlike accommodated assessments in that they present ELLs with an alternative assessment that is not expected (or found) to measure exactly the same skills or knowledge as the regular assessment. For example, such a modified assessment may consist of easier items than found in a regular state assessment and may not cover exactly the same range of skills as the regular assessment. But there are other possibilities, such as use of an off-the-shelf Spanish-version standardized achievement test in math in place of the regular state assessment in math. In this case, *modified assessment* refers to the fact that the standardized math achievement test is measuring a different (though arguably related) set of skills and knowledge when compared to the regular state assessment.

States are required under NCLB peer review procedures to provide psychometric and other related evidence regarding the reliability and validity of all assessments, including accommodated and modified assessments. States must provide evidence that their accommodated and modified assessments are aligned with grade-level student standards and measure the same standards. In the case of alternative (here, modified) assessments, states must make clear a procedure for how performance on a modified assessment meets the same goals as the regular state assessment.

The foregoing requirements hold for students classified by states as ELL, but Title III also requires that states implement English language development (ELD) standards and annually administered ELP tests based on these standards to evaluate ELL students' progress in attaining English proficiency.¹ States are required to establish annual measurement objectives that establish goals for ELL students' growth in English proficiency based on their ELP assessment scores and also based on ELL students' attainment of fully English-proficient status based on assessment scores. Interestingly, although NCLB Titles I and III hold that all ELL students in a state must attain proficiency in reading, mathematics, and science in target grades by 2013–2014, the same 100% proficiency rate does not apply for attainment of ELP. States have been given the flexibility to set lower target rates for ELLs' attainment of full English proficiency based on ELP test scores by 2013–2014.

A more recent requirement under Title III is that states conceptually align their ELD standards and ELP assessments with subject matter standards in reading, mathematics, and science. States are expected to present evidence to the U.S. Department of Education that the content of ELP assessment items includes coverage of language usage encountered in the three target subject matter areas. To meet this requirement, states are expanding their ELD standards and ELP assessment blueprints so that they include "academic language" and problem-solving tasks pertinent to subject matter learning in the three areas, in addition to basic social and functional English characteristic of early acquisition and learning of a second language. The federal goal for this alignment requirement, putatively, is to help schools attend to the academic language-learning needs of ELL students as they strive to support these students' attainment of proficiency in reading, mathematics, and science.

In addition to state assessments under NCLB, the NAEP is also implemented for the purpose of gauging students' mastery of content standards developed from a national perspective in a variety of subject matter areas, including but not limited to reading, mathematics, and science. The NAEP assesses students' mastery independent from states' standards. Under NCLB, states are expected to corroborate the progress of students on their state assessments in reading and mathematics (and eventually science) in light of NAEP results for their state in these subject matter areas. Although NAEP is a separate assessment from state assessments, and the assessment standards of NAEP are different (almost always more demanding than state standards), there is a federal expectation that growth in states' achievement test scores for the same grades over time will also be reflected in growth on NAEP tests for the same grades over time.

In the past 20 years, the NAEP has developed its own policy and reporting agenda concerned with disparities in the performance of ELLs relative to other groups of students, as the national population of ELLs has increased dramatically among the states and as the agency has pursued testing of Puerto Rican students in Spanish. The NAEP agenda has been deeply influenced by U.S. federal laws mandating inclusion of students regardless of background in federal data reporting. This has led the NAEP to undertake a significant body of research on testing accommodation—ways of altering assessment administration conditions—so as to ensure maximal participation of students. ELLs and students with special education status have been particular targets for NAEP accommodation studies.

The question of comparability of ELLs' scores on state and NAEP tests presents a number of issues that illustrate challenges in better understanding what sense we can make of tests supposedly administered to similar ELLs in similar subject matter areas with arguably related measurement objectives.

Attention is now turned to coverage of what research tells us about four key areas that present challenges for adequate assessment of ELLs in large-scale assessments: (a) Who is assessed under the status ELL? (b) How are ELLs included in large-scale assessments? (c) What evidence do we have about the reliability and validity of ELL assessment outcomes on large-scale assessments? and (d) What special issues are arising in the use of large-scale assessments for high school exit purposes?

Who Gets Identified and Assessed as ELL in Large-Scale Assessments?

One of the most consistent findings in current research on assessment of ELLs is that ELL status is not well defined in large-scale assessments (Abedi, 2004). The National Research Council (Koenig & Bachman, 2004), in a major synthesis report on participation of students with disabilities and ELL students in large-scale state assessments and NAEP, concluded that ELLs are not a true demographic population that can be unambiguously defined by a set of well-defined indicators that are reliably measurable in an all-or-none way. ELLs participating in state large-scale assessments are in effect a policy construction, a category of students established by individual states to satisfy their education laws to deal with a growing group of students from non-English backgrounds who show some evidence of limited familiarity with English,

patterns of low school achievement, low assessment scores in English, and propensity to drop out of school and not go on to higher education if they do complete high school. It must be admitted that linguistic and educational researchers themselves do not provide a definitive way to resolve the definition of ELLs. Valdes and Figueroa (1994), for example, present numerous typologies that can be used to characterize bilingual background students with importance to interpreting students' test performance. But what is of essence here is what states do in their own definitions with due consideration for large-scale assessments.

All states rely on a home-language background questionnaire to identify students whose parents or caretakers report whether a language (in addition to or) other than English is spoken at home. A *yes* response to this question leads schools to further screen and assess students for the possibility that they can be classified as limited in their English proficiency and hence eligible for ELD support under state and local programs for this purpose. There is no body of research that has investigated the validity and accuracy of home-language surveys. And there is no research investigating the utility of more "textured" analysis of how one or more languages are used by whom and for different purposes in households and how this might be helpful in identifying the English language competencies and needs of students whose parents or caretakers respond *yes* to presence of a non-English language at home.

In most states a *yes* response to the presence of a non-English language at home triggers administration of an ELP assessment. Performance on this assessment is used to determine whether students can be classified initially as ELL or, alternatively, as English proficient. It is important to note, however, that most states do not classify students as ELL solely on the basis of performance on the ELP assessment. Most states allow local school jurisdictions to use additional criteria to make this decision. In California, for example, local schools are allowed to individually weigh other factors in making this decision. Additional factors that are considered typically include falling below a threshold performance level on a designated English language arts and reading test (this could be the same as the state assessment mandated under NCLB), a teacher's or reading specialist's clinical judgment regarding students' English proficiency, and parents' input regarding their child's readiness for English language instruction. Although states such as California monitor whether school districts and schools use allowable criteria such as that mentioned, a state such as California restricts its feedback to advising schools about how to improve their procedures and to advising schools of their vulnerability to litigation by others in the event they do not improve their assessment practices so as to support ELL students' learning needs.

A related set of relevant factors affecting the definition of ELLs concerns the demographic characteristics and resources of their current communities and schools. It may seem odd logically to consider that the meaning of being an ELL student is tied to the current environment surrounding the community and schooling life of an ELL student. However, it is overly simplistic to maintain a separation between student identity in this regard and the characteristics of the community and schools, because the educational meaning of ELL status is constructed, as has been noted, by local

communities and schools. As was mentioned in the foregoing, states give local communities and schools liberty to decide ELL status based on local criteria such as locally set criteria derived from English achievement assessment performance, teacher clinical judgment of students' readiness for English instruction, and parental advice. There appears to be much anecdotal evidence, for example, that local schools and school boards may favor liberally identifying, or alternatively, conservatively bestowing, ELL status on students because of pressures to maintain or obtain financial resources to provide more educational services to more students or, conversely, to avoid showing lack of resources to serve the additional educational needs of an increasing number of ELL students. There is no systematic large-scale research on this sensitive issue.

States' adoption of unique ELD standards and unique ELP assessments further leads to inconsistencies in how ELL students are defined across states. Each state undertakes an educational-political process to create its framework for ELD standards that lead to the adoption of its ELP assessment based on these standards. NCLB requires that states have ELD standards and that ELP assessment be based on these standards to measure English proficiency of students initially classified as ELL. NCLB is explicit in requiring that states create ELD standards and ELP assessments that assess skills in speaking, listening, reading, writing, and language comprehension. NCLB also requires that states design their ELD standards and ELP assessments in line with findings from research on language learning and its developmental progression, national-level efforts to define ELD standards (such as by the Teaching English as a Second Language organization), and findings on best-assessment language proficiency practices for ELL students consistent with testing standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). However, the U.S. Department of Education, under NCLB, leaves it up to states to specify their ELD standards and ELP assessments in accordance with these general requirements.

Furthermore, states are given license to establish performance standards across assessment areas and composite scores on ELP assessments to determine ELLs' movement across proficiency categories indicating increasing English proficiency and, when students have attained sufficient English proficiency, no longer to classify them as limited English proficient. NCLB does not permit states to do just anything. States are held accountable for providing logical-conceptual evidence and some modicum of basic validity and reliability evidence supporting the claim that ELD standards and ELP assessments are aligned and operating empirically as expected.

Nonetheless, given that there is no consensus in the field of language acquisition research and language proficiency on how to best create ELD standards and ELP assessments, there is very limited consistency and established agreement on who gets classified as ELL or English proficient across states (Abedi & Gandara, 2006; Durán, 2006; Koenig & Bachman, 2004).

A further issue complicating identification of ELLs is that there is great heterogeneity in the background characteristics of students so identified using existing procedures (Abedi, 2004). ELLs can be very different in terms of their non-English language and previous exposure to this language as a medium of instruction, as well as differing in

their experience in learning and using English prior to enrollment in U.S. schools. Also, ELLs vary at which age and grade they enter U.S. schools and what curricula they have been exposed to in their country of origin prior to entry into U.S. schools.

These factors have implications for understanding the learning needs and readiness for ELLs to transition to all-English instruction. These learning needs and readiness to transition to all-English instruction will be heterogeneous themselves and, furthermore, will complicate the interpretation of ELLs' scores on large-scale assessments. The same scores may have very different instructional significance for different ELLs classified as being at the ELP level.

Paradoxically, the idiosyncratic definition of ELLs as a student population among states might work against a nationally consistent way of defining this population but eventually aid individual states in developing accountability models for supporting the learning and academic progress of their unique ELL populations. As states and their local school districts grapple with the growth of their ELL populations, their heterogeneity, and local community and school resources, there is the possibility—in the ideal—that they are likely to become more concerned with a more refined way to deal with the specific learning needs of heterogeneous groupings of ELLs. For example, as will be discussed later in this chapter, there is an emerging national pattern that long-term ELL students in middle and high school (students who have been in the United States and classified as ELL for 5 years or longer) have different English-learning needs and schooling engagement and motivational characteristics as compared to more-recent-arrival ELLs who have strong educational records of achievement via schooling outside the United States in their primary language. States, school districts, and schools faced with getting students to pass high school exit examinations as a requirement for a high school diploma are faced with developing different strategies to get these different kinds of ELL students to graduate from high school, and different states have different laws and funding mechanisms that need to address these needs.

How Are ELLs Included in Large-Scale Assessments?

Despite the highly heterogeneous nature of ELLs, the criteria for deciding how ELLs are assessed in large-scale assessments relies entirely on whether they are simply classified as ELL or not. Under NCLB, students from non-English backgrounds who are determined to not be ELL are administered assessments in English in the same manner and format as students from a solely English language background. These English-proficient students from non-English backgrounds are folded into the general assessment reporting population of students at large. As a special case, under NCLB, students who were initially classified as ELL but who subsequently were reclassified as fully English proficient may be counted by states as members of the ELL reporting category for up to 2 years for AYP purposes (Frances, Kieffer, Lesaux, Rivera, & Rivera 2006). The U.S. Department of Education permitted this practice after considering how the immediate removal of former ELL students from AYP calculations would necessarily depress evidence that states were making significant progress in getting ELLs to meet AYP goals.

With this AYP reporting caveat in mind, under NCLB, states must include all ELL students in their mandated assessments in reading, mathematics, and science. States are given the option under NCLB to administer their regular English-version assessments to ELLs, to administer their English-version assessments with accommodations to ELLs, or to administer ELLs modified assessments (for the latter to be acceptable under NCLB, a strong argument is required that the modified assessment provides comparable results to the regular state assessment). In implementing these options, under Title I of NCLB, states are held accountable for providing evidence supporting the conclusion that accommodated and alternate assessments administered to ELLs measure the same range of content standards at an ELL student's grade level as the regular English version assessment. States are also required to document that the scores and proficiency-level assignments earned by ELLs have the same meaning as for non-ELLs.

State policies regarding provisions of accommodations and alternate assessments have increased dramatically since 2001, as states have implemented their state assessments responsive to requirements for student inclusion in state assessments under NCLB (Rivera & Collum, 2006). States have only sporadically conducted research examining the validity and reliability of accommodated and modified assessments for ELL students. Research has found that a school's decision to administer assessment accommodations to ELLs is based overwhelmingly on what state policies specify as a permissible accommodation (that presumably is convenient for local schools to administer) and not on a linguistic rationale intended to reduce English language load faced by ELL examinees on an assessment (Abedi & Gandara, 2006). In other words, ELLs are prone to get assessment accommodations, such as extended time or assessment in small groups, because these create less burden on schools when managing the administration of assessments. More recently, Kopriva, Emick, Hipolito-Delgado, and Cameron (2007) report experimental evidence that selection and administration of accommodations tailored to the specific language background, cultural familiarity of ELLs with U.S. schools, and schooling practice can lead to improved state test scores compared to ELLs administered no accommodations or ELLs administered only some of the recommended accommodations given their background.

What Do We Know About the Performance of ELLs on Assessments?

There is consistent research evidence that ELLs, as a whole, perform at lower levels on unaccommodated large-scale assessments administered in English. NAEP data on ELLs shows that students identified as ELL score lower on reading and math assessments at all grade levels (Mazzeo, Carlson, Voelkl, & Lutkus, 2000). Abedi and colleagues (Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, Kim, & Miyoshi, 2000) report several studies conducted in a variety of settings across the United States using assessments built from released NAEP math and science items that have shown that ELLs tend to perform better on these assessments than on assessments of English reading using released NAEP items. This research has also shown that the reliability of assessment performance is high to very high for ELLs on math and science assessments

but only moderate to moderately high on English reading assessments, suggesting that greater demand on English can make ELLs' test performance less stable.

Although states have lagged in conducting extensive research of their own on the validity of assessment accommodations administered to ELLs, there has been a growing body of such research, though it has not produced consistent findings. As the field has emerged, two key questions have come to guide research: (a) Is an assessment accommodation effective? That is, does it raise assessment scores for a target population such as ELLs above levels obtained by such examinees administered an assessment without the accommodation? and (b) Is an assessment accommodation valid in preserving the intended meaning of assessment scores as measures of target knowledge and skill areas? (Abedi, Hofstetter, & Lord, 2004; Sireci & Zenisky, 2006)

In the case of ELLs, the first question concerns whether an assessment accommodation does what it is intended to do—namely, facilitate ELLs' ability to access the information required to understand and to work an assessment item without having limited proficiency in English interfere with problem solving—in those cases where ELP is not a skill deliberately intended for assessment in a target knowledge and skill area. The second question addresses the issue of whether any advantage in assessment scores shown by ELLs administered an accommodated assessment may alter the meaning of the underlying constructs intended for measurement. A common way to investigate this is to check whether non-ELLs administered an accommodated assessment show no dramatic increase in assessment scores compared to performance shown when they are administered a nonaccommodated version of an assessment. If non-ELLs show as dramatic an increase when administered an accommodation compared to non-ELLs not administered the accommodation, this is taken as evidence that the accommodation has altered the constructs targeted for assessment. In the research literature on accommodations, a combined test of the effectiveness and validity of an accommodation for ELLs versus non-ELL students is labeled the "interaction hypotheses," referring to the use of a two-way analysis of variance for determining whether an accommodation is statistically more effective for ELLs but not for non-ELLs (Sireci, Li, & Scarpati, 2003).

Syntheses of research evidence on the effectiveness and validity of accommodations have not found consistent results across accommodation types (Frances et al., 2006; Koenig & Bachman, 2004; Sireci et al., 2003). Studies sometimes show evidence of effectiveness, but not validity, as with linguistic simplification of test items where it has been found to aid native English speakers more than ELLs, and in some cases, linguistic simplification has not shown evidence of effectiveness regardless of group (Abedi, 2006; Frances et al., 2006).

Two accommodation types show the most evidence of effectiveness and validity for ELLs (Frances et al., 2006). These include (a) providing ELLs with customized English dictionaries including definitions of a limited range of terms occurring in assessment math or science items and (b) providing (English or bilingual) glossaries elaborating on the meaning of select math or science terms occurring in assessment items in cases where the linguistic access to the meaning of terms was not an assessment construct target.

Extended time as an accommodation has been found to be associated with increases in assessment performance for ELLs, but this accommodation can be confounded with other factors affecting assessment performance because it is seldom administered without being coupled with another assessment accommodation, such as small-group administration or the linguistically based accommodations mentioned above, therefore making it impossible to isolate performance improvement as because of extended time alone.

Another important form of assessment accommodation that has been investigated provides ELLs with assessment items in their primary language. This can occur in the form of separate English and non-English versions of the same assessment or a dual-language assessment where items are presented in both languages on the same assessment instrument. The development and use of translation equivalent assessments is considered a form of “assessment adaptation” for cross-cultural assessment purposes (Hambleton et al., 2005). Although assessment adaptation is a more general concept referring to a range of ways to make assessments potentially more equivalent across populations, it is helpful in the context of discussing assessment in more than one language because it highlights the importance of a variety of sources of cultural issues in assessment content that interact with ways in which different languages affect the meaning and intelligibility of items represented by language on assessments (Geisinger, 1994; Hambleton, 2005; Stansfield, 2003; van der Linden, 1998).

The creation of assessment items that are considered to be translation equivalents is a complex enterprise (Stansfield, 1996; Vijver & Poortinga, 2005). Although there is no one best procedure, assessment items that are “transadapted” (Stansfield, 1996) undergo a rigorous cycle of development and checking for quality of translation. This can involve translation of an assessment item from a second language (L2) such as English to an L1 such as Chinese or Spanish, followed by back translation from an L1 to an L2, to check whether the same meanings are being communicated in the versions of an assessment item in each language. It is also possible to check whether translation of an item to a third language (L3) from an L1 and L2 version leads to recognizable equivalent items in L3. Professional translators are recommended for these purposes, and their work can lead to substitutions in wording, syntax, and idiomatic usage across languages for an assessment item for the item to be matched better with the linguistic and cultural backgrounds of examinees. Piloting and field-testing with examinees and school staff are also used to check for perception of equivalence of meaning across transadapted items. Although perceptions of quality of translation are important as indicators of the validity of translated versions assessments, they do not establish the degree to which translated assessments are psychometrically equivalent.

The general problem of establishing the psychometric equivalence of assessments in two languages is covered by Wainer (1999) in his aptly titled paper “Comparing the Incomparable.” Wainer notes that it is impossible psychometrically to equate assessments in two languages in the strongest sense—scores on two assessments have exactly the same score distributions and measurement scale properties—because there can be a confound between the ability required to solve items in one language versus the other and the ability of two groups of examinees who are assigned to be assessed in one

language versus the other. When two translated assessment items differ in difficulty, for example, when the item in L1 is solved by a smaller proportion of L1 examinees than when the item is presented in L2 to L2 examinees, what is responsible for this difference? Is it that the L1 version of the item is harder because the linguistic (and cultural) properties of the item make it harder, or is it because L1 examinees have learned less of the knowledge and skills assessed by the item?

The solution to this problem is to relax the meaning of *equivalent* by adding assumptions. As Wainer (1999) explains, one way to do this is to assume that translated items on an assessment are equivalent in difficulty across languages if their degree of difficulty within each language of assessment remains invariant across languages. That is, if we were to order the difficulty of items in each language, would we find the same difficulty order of items in the other language? If this is so, we have met a necessary, but not sufficient, condition for the equivalence of assessments across two languages. Meeting this condition is not sufficient to ensure the strict equivalence of translated assessment because additional statistical properties would need to be met. The other solution to the problem is to assume that two groups, one taking the L1 assessment and the other the L2 assessment, are of equal ability and to adjust scores statistically for any observed differences across groups.

Sireci, Patsula, and Hambleton (2005) and Sireci (2005) outline strategies of this sort to permit linking and comparison of scores on assessments that are intended to be parallel in two languages. Suffice it to say that although there is no perfect method to ensure that scores on the same intended assessment in two languages are strictly equivalent, psychometricians have worked out a variety of ways to verify that there are systematic and expected ways that scores on two assessments are related once assumptions are made. Example techniques, beyond investigating whether items show highly similar difficulty level in each language across two languages, include looking at whether items show no change in difficulty across L1 and L2 groups given their overall score in each language, whether items are measured with the same degree of reliability precision in each language, whether items cluster together in difficulty in each language on the basis of particular knowledge and skills they are designed to assess, and whether persons who are fluent in two languages perform similarly when presented items in both languages.

Before proceeding to a discussion of the cognitive issues raised by the quandary of establishing equivalence of assessments in two languages, it is helpful to add some comments on dual-language assessments of achievement. Dual-language assessments are translated assessments where an examinee is presented with assessment items in two languages simultaneously, such as in a side-by-side format, where the left portion of an assessment booklet has items presented in one language and the right side of a booklet presents the same item in a second language. No clear evidence has emerged that this accommodation leads to enhanced ELLs' performance compared to ELLs administered regular English versions of items (Sireci et al., 2003), but there is evidence also that ELLs perform no more poorly on such assessments compared to regular English assessments (Mazzeo et al., 2000). Some cognitive interview data exist suggesting that

ELLs administered dual-language assessments pick one language instead of the other and only concentrate on that language during an assessment. It also appears that preference for English on such assessments is connected to the fact that ELL students are receiving their current instruction solely in English.

Now let us turn to the problem of equivalence of items and assessments intended to assess the same skills and knowledge via assessment items found on large-scale assessments from a psychometric perspective attuned to cognitive and background issues. This discussion will serve as a bridge to the final section of this chapter suggesting the value in rethinking what counts as evidence for ELLs' learning achievement in content areas given the putative goals of large-scale assessments under policies such as NCLB.

Martiniello (2007) examined the effects of linguistic complexity on DIF for ELLs on solving math word problems. Following prior research by Abedi et al. (1998) and others, she postulated that ELL fourth graders would find Massachusetts Comprehensive Assessment System math test items with greater English language requirements harder than was the case for non-ELL peer students. Indeed, this turned out to be the case. In addition, however, she examined whether presence of visual images representing mathematical relationships necessary to solve problems mitigated the difficulty of items for ELL students. This turned out to be the case. As part of her research, Martiniello interviewed a small sample of 24 students regarding how they solved problems as they solved them. Her protocol collection method involved probing students about difficulties students encountered with interpreting linguistic terms occurring in problems, the math concepts alluded to, and ways that visual images supported performance. Her findings supported the conclusion that the presence of diagrammatic images related to critical problem information supported ELLs' performance when they encountered difficulties in understanding the English statement of problems.

Although the simple conclusion that presence of visual or other aids common to test accommodations is supported by the foregoing study, there is not always clear evidence that these aids really work as intended—a partial reason why the testing accommodation research has not always found that accommodations work as intended (see Kopriva et al., 2007). There is a growing literature regarding the cognitive functioning of persons in a first and second language that builds on cognitive psychology research that is relevant to this disparity. This chapter does not review this research in detail, but its mention is useful, nonetheless, because it suggests that the performance demands of learning tasks are an important issue that needs to be considered in understanding ELLs' learning. This research cited focuses on ELLs' performance of very specific cognitive learning and performance tasks and on how multiple representations of task information—for example, L1 and L2 representation, computer-aided figural representation, or animation—affects emerging bilinguals' accuracy and efficiency in problem solving. Three not-altogether-consistent examples of such research cited here illustrate how the theoretical notion of *cognitive load* can help explain limitations that ELL or other dual language-immersed students face when asked to perform complex learning and other cognitive tasks. According to cognitive load theory (Sweller, 1988), learning and problem solving are dependent on the capacity of working memory to keep track

of information relevant to performing the tasks at hand. That working memory capacity is limited to about seven, plus or minus two, chunks or sets of information at a time is one of the earliest findings of cognitive psychology (Miller, 1956).

Plass, Chun, Mayer, and Leutner (2003), using a cognitive load research paradigm, examined the ability of bilinguals asked to learn new vocabulary terms embedded in reading texts in a second language. They postulated that bilinguals' effectiveness in vocabulary learning in a second language is heavily dependent on the working memory capacity of learners. They tested the hypothesis that extra support for reading via computer-provided visual, written, or orally read annotations regarding word meaning in L1 would facilitate L2 vocabulary learning. Contrary to expectations, their findings indicated that visual annotations reduced vocabulary learning for second-language learners relative to other second-language learners or native-language students receiving no annotations or orally read annotations explaining the meanings of words. They concluded that the addition of supplemental visual support, rather than reducing processing load, actually overloaded the memory capacity of second-language students and that this led to a deterioration in their ability to learn. They suggested that offering students a choice in what multiple representation support they would receive would ameliorate this negative effect. Students would be able to decide on their own what form of either visual or oral support would help or hinder their information processing, taking into account their own sense of what was most effective for them.

The latter hypothesis held up in a study by Durán and Moreno (2004), who found that providing ELL elementary school students with a choice of oral support in either Spanish or English during a visually animated mathematics learning task improved students' performance compared to students not provided the extra language support. Research such as the foregoing continues, but although it is informative on cognitive grounds, and suggests strongly that cognitive and assessment tasks presented to ELL students have to be carefully understood in terms of their linguistic and information processing demands to interpret performance on these tasks, research by and large has yet to broach ways in which social and cultural understandings of context may have a concomitant effect on ELLs' performance on typical school learning tasks and assessment tasks intended to reflect school learning. Here is where attention to ELLs' background can add insights that portend such understanding.

Recall that earlier, the issue of the great heterogeneity in ELLs' background was brought up. Abedi et al. (2004), going beyond concern for the effectiveness and validity of assessment accommodations on ELLs, raise concern for understanding how background differences within ELLs might affect assessment scores. Using postassessment questionnaires to gather background information, they found that student information on time lived in the United States, language spoken at home, and self-ratings of understanding English at school, proficiency in the non-English language, and English-language status predicted NAEP science assessment performance significantly better for students not receiving an English dictionary accommodation than for students receiving the accommodation at the fourth-grade level. Although the results at the eighth-grade level did not attain statistical significance, they were in the same direction. The

evidence suggests that the dictionary accommodation reduces the importance of the selected background variables as determinates of student performance on the NAEP science assessment. The findings also suggest that students from different non-English backgrounds may benefit differentially from the availability of the dictionary accommodation. In particular, fourth-grade students from a Korean-language background seemed to benefit significantly better from this accommodation.

Solano-Flores (2006) and Solano-Flores and Nelson-Barber (2001) suggest that student, schooling, and language background variables interact in systematic patterns that affect cognitive performance on NAEP and specially developed standards-based science items in English for ELL students. The key idea is that variation in assessment performance on particular items in English for ELLs is not solely because of student ability factors and an unexplainable noise-error factor but also systematic error of measurement that can be explained by an interaction between these factors and student background factors. Consistent with Lee's (2005) advocacy for greater sensitivity to cultural influence on science assessment performance, they argue that there is an underlying notion of assessment cultural validity that needs to be considered when explaining performance on science items (and items in other content areas) by ELLs. Solano-Flores and Nelson-Barber (2001) and Solano-Flores and Trumbull (2003) report generalizability theory statistical studies and cognitive lab studies of ELLs with different language, schooling, and national origin backgrounds on specific NAEP assessment items. The generalizability studies found that some ELL student groups from different language backgrounds show more variability in performing on specific assessment items than others. The cognitive lab studies isolated specific comprehension issues encountered by students from different backgrounds in understanding required information in assessment items. Solano-Flores (2006) points to evidence that ELLs from different background might require different numbers of assessment items for assessment reliability to be high enough to support the validity of assessments. However, Solano-Flores and Nelson-Barber are careful to point out that the underlying question of assessment validity needs to take into account more specifically the cognitive and linguistic backgrounds of examinees and how the interpretation of assessment item information is constructed by examinees.

High School Exit Examinations and ELLs

In 2005–2006, there were 25 states implementing high school exit examinations as requirements for receiving the high school diploma (Center for Education Policy [CEP], 2006). These states comprise 65% of the nation's students and 76% of the nation's ethnic minority students. Although not computed formally, this would comprise between 50% to 80% of all ELLs who are thus required to pass high school exit examinations to receive a high school diploma. Use of high school exit examinations for many states is also tied to meeting NCLB requirements for AYP goals in reading, mathematics, and science, but several states, such as New York with its Regents Examination system, include additional areas of subject matter assessment as part of their high school examination system.

ELLs lag considerably across the country in passing high school exit examinations. The CEP (2006) reports that among 22 states providing breakouts of pass rates for ELLs in 2005 on their first try at passing, between 17% and 60% fewer ELLs passed the reading and English arts high school exit exams compared to students as a whole in their respective state (median = 35%). The corresponding gap in mathematics pass rates was noticeably lower, with the gap within a state ranging from 41% to 4% (median = 20%). These differences are consistent with the hypothesis that ELLs will have more difficulty with assessments that require a greater reliance on English language assessments.

There is a lack of research comparing the background and previous achievement records of ELLs who pass or fail their state's high school examination on their first attempt at passing the examination. There is some reason to believe that those failing to pass are heterogeneous in terms of their prior achievement and history in U.S. schools. Specifically, there is evidence (Callahan, 2005) that a significant number of these students represent "long-term ELLs." These are students who have been classified as ELLs for more than 5 years and failed to be reclassified as English proficient by the time they reach the 10th grade. Callahan (2005), in her research on achievement of ELLs in California high schools, suggests that academic tracking of these students is occurring and that these students show a historical pattern of cumulative low school achievement prior to high school entry. She also cites research by Raudenbush, Rowan, and Cheong (1993) reporting that low academic tracking of students is associated with less demanding language and discourse practices in classrooms as well as less demanding academic content. These possibilities are ones deserving greater research attention, as they suggest that the socialization of students to schooling and the opportunity for ELLs to learn and develop an identity as successful students are critical issues tied to both communication and learning of content.

RECONSTRUCTING ELL STUDENTS' ACHIEVEMENT AS ACTIVITY AND AN EXAMPLE OF IMPROVING ASSESSMENT OF ACQUISITION OF ACADEMIC ENGLISH

Classroom Learning Activity

The notion of cultural validity of assessments in the context of the goals of large-scale state assessments under NCLB and on high school exit examinations to measure what ELL students know and can do based on learning standards is an awesome task, given the cumulative issues that have been cited. Both language and culture are deeply implicated in the validity of assessments, and they are also intertwined with ELL students' educational and experiential histories and opportunity to learn what is expected in classrooms under existing educational policies. What can be done to create assessments that go beyond the limitations of existing large-scale assessments as indicators of what ELL students know and can do? It is important to realize that resolving these issues by refining the design and implementation of existing large-scale assessments will get the field only so far.

Baker (2007), looking toward the future, suggested that we need additional tools to complement existing large-scale state assessments to create assessments and bodies of evidence that reflect students' development and learning accomplishments in a manner more sensitive to public policy and the emerging goals of schools in the new millennium. She advocated that current psychometric methods need to be extended to better map curriculum goals and students' growth in subject matter competence and its application to new learning over time. She suggested that educators explore a wide variety of assessment tools and bodies of evidence of student accomplishments that schools and the public at large will value as important learning outcomes of enduring significance for students. She mentioned, in this regard, that we need to consider tools for the accumulation of evidence of achievement growth sensitive to students' background and unique propensities to learn different skills and knowledge in different ways. Also at issue are students' unique pathways through schooling, given these characteristics. Baker cited the distortions to instruction that occur, given schools' and teachers' accountability to raise scores of students on large-scale assessments under state and federal schooling accountability policies. Although these perspectives are relevant to U.S. students at large, they have special relevance for second-language learner and immigrant students in international contexts, given second-language learners' diversity, language needs, and learning needs, as previously cited.

One way through this thicket is to start by reconceptualizing what we can mean by *classroom achievement* itself. Although there is no one way to resolve this issue, researchers in the tradition of cultural historical activity theory (CHAT; Cole, 1996) provide valuable insights that imply we ought to look more closely at classroom interaction itself as the locus of assessment and learning. From a CHAT perspective, human development and competence emerge through socialization processes that support individuals' acquisition of skills and knowledge that allow persons to develop and exercise identities as members of social groups and participants in social domains and social institutions. Fundamental to CHAT and related sociocultural approaches (see, e.g., Gee, 2007), human social interaction is the primary route for learning to take place. This concern, of course, also includes students' functioning and interaction in classrooms as learning sites.^{2,3}

Scribner (1979), in her research on the cultural foundations of cognitive and linguistic skills, called attention to the notion that members of cultures and social groups acquire and develop "genres" for thinking, problem solving, and language use attuned to their everyday living circumstances. Building on Vygotsky (1978) and Leontiev (1981), Cole (1996) and Wertsch (1998) call attention to the fundamental principle that human action is tied to how individuals come to perceive and interpret the situations they construe and how they use these construals to project and guide their purposive action. Situations are interpreted in terms of components involving projections about who, where, when, what (goals), why, and how. These are notions that in contemporary cognitive science undergird social and cognitive action described in terms of scripts (Nelson, 1996) or cultural mental models of situations and action (Holland & Quinn, 1987). What do these ideas suggest about conceptualizing the development of learning agency among ELL students, assessment, and opportunity to learn in the classroom?

ELL students need to acquire identities intimately tied to their agency as learners in the context of classroom cultures and genres for thinking, problem solving, and language use that fit the demands of classroom learning goals. Scarcella (2003) and Bailey (2006) provide elaborate accounts of the range of academic English skills ELL students need to acquire, and Adamson (1993) elaborates how cognitive schema and script theory and Vygotskian theory improve our understanding of how ELLs come to be socialized to participate in classroom learning tasks and to meet communicative competence requirements of these tasks. Consistent with the accounts of Nelson (1996), Holland and Quinn (1987), and Cole (1996), ELL students need to acquire mental and cultural models and scripts for how to act out being competent participants in classroom learning activities, and this includes acquiring competence in using language and language structures to participate in learning activities for communicative purposes.

Taking things a bit further from a Vygotskian perspective, Tharp and Gallimore (1988) assert that bona fide instruction can occur only when students are assisted in performing and progressively internalizing knowledge and skills they have not previously mastered. According to this "strong" definition of instruction, true instructional learning does not occur unless students actually come to know and do things that they previously have not mastered through participation in instructional activity. For learning to occur via instruction, students and teachers or teacher aides must jointly construct understandings of what the goals of learning activity are about and of how to evaluate competence in learning progression. Interaction and communication are central to this process, and this must involve participants' assessment of their intentions and common construals of activity and social and self-regulation of performance to respond to assessment. As effective instructional activity proceeds, students go from not being able to perform tasks and apply knowledge to being able to do so with support from more capable others. This support may be in the form of direct advice on how to perform tasks and apply knowledge, to modeling of competence by more capable others, and to reciprocal interaction where a more capable other extends scaffolding via interaction to guide a learner through next steps of competent performance within their "zone of proximal" development. As student learners acquire competence, they begin to internalize this socially shared understanding of competence so that it may be displayed with less overt support from more capable others, though always, competent performance remains a social act requiring that learners recognize how competent performance is realized through shared social perceptions of action in the classroom and its ongoing range of cultural practices.

The importance of the foregoing resonates with the seminal qualitative research findings of Wong Fillmore (1976), who found that ELL children's social interaction with native English speakers provided the ELL children with exposure to native, fluent exercise of English discourse and speech act forms that the ELL children could then approximate in their interaction with the native English-speaking children. Wong Fillmore's research showed that this strategy, when accompanied by the reciprocal response of competent English speakers, acknowledging the functional intent of ELL students' utterances, helped ELL children acquire increased English competence through repeated refinement of the intended linguistic form.

So as we think about how to develop a stronger foundation for ELL assessment, it would seem very helpful to worry about how our assessment designs and assessment targets might benefit from the foregoing perspective.

Assessing ELL Academic English Competency: An Example

In what follows, an account is given of how ELL student assessment might be reconceptualized toward this end, with an emphasis on acquiring one very important class of academic English skills: being able to recognize questions and knowing how to answer them appropriately. The question-recognition and answering contexts under consideration involve reading and understanding a text and answering a question about what was read. This is but one example of the great range of academic language skills in English that ELL students are required to master, but its consideration has much to offer in that it moves the issue of assessment and its connection to learning in a concrete manner. Rather than addressing how assessments might be improved to inform the learning of academic English in the abstract, the example allows us to be concrete about ways that a specific model for a particular class of functional English usage can be coupled with assessment in a manner tied to the improved use of assessment for the purpose of evaluating students' learning.

Questions, like other linguistic forms, do not exist in everyday reality in isolation. They arise and are used as communicative discourse forms within situations and activities where they assist participants in negotiating meanings and getting practical business done. This is certainly the case in classrooms beginning in the earliest grades, where they form a foundational genre for classroom interaction and learning. Students must develop the ability to recognize their occurrence and how they fit into academic tasks as part of classroom cultural practices. And beyond this, students must acquire the ability to go about answering them competently in written as well as oral form.

Durán and Szymanski (1997) reported a study using formal pre- and postassessment of third-grade ELL students' ability to benefit from question-answering instruction. An important feature of this research, with import for new forms of assessment, is that the study also involved analysis of discourse interaction among students and their teacher during question-answering instruction. The research used a CHAT and Vygotskian perspective to analyze when and how students actually worried collectively about the specific linguistic and semantic properties of questions and how they pursued answering them in writing through their interaction.

From a Vygotskian perspective, the challenge of the research was to start with the research issue: How well can individual students working alone answer written questions about meaning conveyed in text passages prior to start of a specific lesson sequence intended to improve question-answering skills? This was the role of the pre-assessment. A second question was How well can individual students working alone answer questions following participation in the instructional intervention? Evidence of increased independent performance would be consistent with the inference that students had learned from instruction—though a stronger research design would have required a comparison group with no intervening targeted instruction on question

answering to control for the possibility that repeated exposure itself to question-answering demands on the posttest improved performance just because students were already familiar with question-answering demands from the pretest.

However, the most critical component of the Vygotskian and activity theory perspective in the Durán and Szymanski (1997) study was its analysis of how the teacher's instructional interaction with students, and students' interaction with each other and the teacher, made visible students' acquisition and use of the teacher's cultural model or schema for how to answer questions. The analysis of this interaction also revealed evidence of formative assessment. As the students interacted, they evaluated their knowledge of the teacher's model of how to analyze and answer questions, and they regulated their understanding of this model through offering feedback to each other so as to better apply the model. Consistent with a Vygotskian account, the talk or explanation of how to answer questions was first shared as interaction between the "expert" teacher and "novice" students. The teacher explained how to identify different "*wh* question" linguistic markers—*who*, *what*, *when*, *why*, and *how*—and that they occurred at the beginning of a question sentence or clause. (Consider, e.g., "What did Jim buy at the store?") She also modeled and explained that the question always pertained to a semantic subject (e.g., "Jim") explicitly mentioned in the question and that the question marker specified the kind of semantic information requested about the subject (e.g., *what* pertained to an object bought by Jim at the store). She showed students how to keep track of the *wh* question marker and subject by underlining pertinent parts of the written question. She next modeled how students might search a portion of a target text for the answer to the question—again underlining possible relevant sections. She also showed how once the relevant information was found in the story, students could next begin planning and writing their answer. This, prototypically, was modeled by the teacher as first "echoing" the subject (e.g., "Jim") at the start of the written answer, followed by restating information already given in the question with an appropriate change in verb tense (as in "Jim bought") and then finishing the answer-sentence with the new information completing the semantic requirement for an answer (e.g., "a shirt"). The teacher in this process also explained to students the difference between answering a written versus spoken answer to a question. She explained that in face-to-face talk, people typically just give the answer without all the other information made explicit in the written academic form of a question answer (e.g., by just uttering, "A shirt").

The Durán and Szymanski (1997) study found evidence of gains on ELL students' ability to answer the same *wh* question forms from pretest to posttest. They also showed evidence that the students were able to emulate the teacher's way of talking about question answering, and steps in question answering, through their social regulation of problem solving to answer questions in small groups. The interaction analyses revealed that students reminded each other of the steps that the teacher had shown them for answering different types of *wh* question forms and that they assessed how well they were carrying out these steps through their interactions with each other and how they refined their problem solving so as to reflect better the teacher's model and explicit steps for answering questions.

This example of research is suggestive of how one might improve assessments of ELL students by having a clear specification of what competencies students are expected to acquire as a result of specific instructional sequences and experiences. It further illustrates the value and importance of having available a specific framework for instruction that can help analyze how students might practice and improve their learning competence through steps that they undertake in performing a complex learning task. The specificity cited does not imply that there is one best or only way to teach question answering. It does imply, however, that having a clear conception of a target skill domain and how instruction in that domain is designed to occur helps and that evidence from formal pre- and postassessments and assessments rendered in instructional interaction can be interpreted to ground an understanding of important learning.

How does an approach to instruction and assessment such as the preceding link with the goals of educational standards—informed instruction for ELLs? First, it is important to note that such links are possible. Statements of learning standards in English language arts can be connected to instructional practices addressing acquisition of those standards by students. Formal pre- and postassessments tied to standards can be used to evaluate whether students have acquired knowledge and skills. Note, though, that large-scale assessments are not capable of such refined diagnosis. Large-scale assessments used by states in an area such as English language arts are “status” instruments and are administered only once per year and intended only to be gross measures of what students know and can do in a subject matter area. These assessments are not intended sensitively to measure growth in students’ learning across time.

In addition, it is also important to understand that statements of learning standards in an area such as English language arts or any other academic area cannot adequately treat how learning is supported and how it might use assessments to guide instruction. The notion of “classroom formative assessment” has been explored in recent years to address this issue (Black & William, 1998). This notion of assessment addresses how teachers can use day-to-day assessment information to guide instruction. Whereas this work has concerned use of formal classroom assessments for this purpose, others have extended the notion of formative assessment so that it is considered as potentially occurring within learning interaction itself (Torrance & Pryor, 1998), consistent with the analyses described by Durán and Szymanski (1997).

Erickson (2007) refers to this interaction-embedded form of assessment as “proximal formative assessment.” Proximal formative assessment centers attention to the ways in which instruction is supported through social interaction. Direct instruction of knowledge and skills will not suffice. Students cannot demonstrate acquisition of new knowledge and skills if they are exposed merely to what they are expected to know and do via a “broadcast” model, where teachers explain and model what needs to be learned without students’ having the opportunity to practice use of knowledge and skills with feedback from a teacher or other students regarding the suitability of performance. Instructional contexts are socially constructed, started, maintained, and temporarily terminated by participants through their actions and through their interaction. Assessment is provided via interaction and via ability to understand how to take up the implications

of assessment by revisions of subsequent action by learners (see Durán, 1994, for how this take-up is tied to Tharp and Gallimore's [1988] notions of assisted performance). Erickson also raises the issue that we ought not to view learning and assessment supported by interaction as bounded by immediate instructional incidents. He states that the danger is that "what is learned" by students is treated as an entity that comes to exist after instruction has taken place, and thus, can be measured *as a whole thing of the past*" (p. 190). As an alternative, he proposes that learning be considered as a continuously constructed entity in the ongoing course of classroom life and its activities.

Consistent with Erickson (2007), Durán and Szymanski (1997) and Putney, Green, Dixon, Durán, and Yeager (2000) discuss the notion of consequential progression as a concept capturing how present understandings of a context and activity have a historical reality and the potential to shape and support future interpretations of context and activity. ELL students' learning of new knowledge and skills is not historically isolated and understood by considering just one occasion of instruction that seemed to benefit learning. New learning is based on prior learning, and new learning has further consequences for additional learning. Better understanding how formal assessments and assessments in interaction guide the consequences of prior learning on new learning is complex to consider and lies at the heart of enhancing assessment for ELLs' and other students' school learning. For example, ELL students' acquisition of academic English skills in an area such as question answering is not an isolated learning accomplishment. Such knowledge and skill acquisition is coupled with learning of other academic English skills and becomes part of a repertoire of communication skills that enables learning across domains of social and academic experience. The intent of existing standardized English proficiency tests and state English language arts large-scale assessments is to address evidence of what has been learned, but the design of such assessments is incapable of representing or addressing how learning progresses as a real cultural and social interactional process.

As discussed at the start of this chapter, the most fundamental question about assessment is about what valid inferences may be drawn about competence in a target content domain, given assessment performance information. Large-scale assessments have a severe limitation in this regard. Their results can at best provide a long-distance, coarse, static understanding of what ELLs know and can do, and these inferences are severely affected by ELLs' diverse backgrounds and by the limitations in the design and sensitivity of assessments for this population. That stated, this does not imply that such assessments are not without value, and yes, there are improvements in the ways that such assessments can be supplemented by more targeted formal assessments that reveal cognitive and linguistic growth and by proximal formative assessments that look closely at how interaction in classroom learning activity supports learning from the lived perspectives of ELL students as participants with teachers in creating classroom culture.

CONCLUSION AND FUTURE DIRECTIONS

Looking to the future, beyond refinements of the sort suggested above that make instruction and assessment more tightly linked to ELLs' acquisition of specific

important skill types, it is helpful to step back and consider a snapshot critique of large-scale assessment. Students' proficiency in a subject matter can be captured in only a limited manner by one-dimensional constructs of academic competence, such as those operationalized by existing large-scale assessments using item response theory. Although there is research on multidimensional versions of item response theory, there are no versions of such assessments used for accountability purposes by states, and it is not clear how such models would have application to highly complex skill and knowledge domains (though see Mislevy, 1997). The heterogeneity among ELLs, as cited, for example, in the work of Abedi (2004), Martiniello (2007), and Solano-Flores (2006), leads to evidence that assessments administered many ELLs are measuring more than an intended skill and knowledge area. The unaccounted-for error variance in assessment scores is not random error; the research cited indicates that it includes systematic variance in performance that results from an interaction between the knowledge and skills required to solve particular assessment items and the background and schooling characteristics of ELLs. There is information in these interactions that may have instructional value for students, particularly if instruction can be made culturally and linguistically responsive to the backgrounds of ELL students.

A second limitation of large-scale assessments is that, at best, they can provide only a "thin" coverage of what students know and can do. Such assessments are built to sample discrete skills and knowledge specified in subject matter and corresponding assessment blueprint frameworks. Constructive critics of large-scale assessment, such as Popham (2006), suggest that given the reality that large-scale assessments can present only so many items that cover only so many standards in a subject area, it is wiser to have assessments target fewer assessment standards and to do so in a manner that reflects more systematically how standards might be interrelated. The current call under NCLB for growth models exacerbates this issue, in that state assessment systems lack strong developmental progression models based on cognitive and curriculum design theories for representing how students advance in expertise across years in subject matter domains. These issues are important for understanding ELLs' performance on large-scale assessments within and across years. ELL students enter U.S. schools at varying ages, with varying fluency in English, and with different background and schooling experiences in their primary language. Inferences drawn from large-scale assessment results for ELLs may be invalid because the results do not indicate simple presence as opposed to absence of skills. Existing assessment accommodations for ELLs may not be powerful enough to level the assessment playing field for these students. This is more than an issue of reliability of assessment performance; it can also reflect a mismatch between what ELLs are expected to know and do and systematic variation in their opportunity to learn these skills and knowledge. Not understanding this match and mismatch results in lost information on how schools might better serve these students.

A third limitation of existing large-scale assessments is that they do not reflect current research on the social and cultural nature of learning contexts or cognitive and interaction studies of students' interaction in learning settings that suggest new ways to

design evidence of achievement that are different from traditional assessments. This critique is very much in evidence in Baker's (2007) American Educational Research Association presidential address, where she focused centrally on cognitive science research on students' development of subject matter expertise and suggestions on the need for 21st-century students to develop complex skills and knowledge sets attuned to developments in technology applications to daily life, social and institutional demands for increased flexibility and adaptability of employees and leaders, and the increasing importance of communication skills across institutional and global communities. Certainly, here, ELLs' fluency in a non-English language becomes a potential asset for learning and expertise development rather than a deficit to be replaced by English and knowing how to apply skills and knowledge solely in English-speaking contexts.

The real-world and instructional learning environments are much too complex to be represented in depth by assessment items presented in an isolated manner from authentic social and cultural settings for learning, where skills and knowledge are deeply interrelated and integrative in nature. Learning standards themselves suffer the same limitation in that they are embodied in stand-alone statements about skills and knowledge that break down what is to be learned in isolation from the actual processes and sociocultural practices that constitute participation in schools. Sociocultural and activity theory research, evidence-centered design of assessments, and postmodern views of assessment suggest possibilities; see, for example, Moss et al. (2006), Mislavy (1997), and Mislavy and Huang (2006). The focus in these accounts is on the validity of assessments. What do we "really" want evidence of and how do we evaluate and use this evidence? Rhetorically, one might ask, do we really want just to know that students have acquired discrete bits of knowledge and skills, or do we really expect more from students, teachers, and schools? Gee (2007) puts it quite concisely:

To fairly and truly judge what a person can do, you need to know how the talent (skill, knowledge) you are assessing is situated in—placed within—the lived social practices of the person as well as his or her interpretation of those practices. (p. 364)

You in the foregoing can be seen to refer to all stakeholders concerned with educational outcomes, be they parents, teachers, employers, policymakers, and so on. And importantly, *you* can also refer to students' own inferences about their competencies and their peers' inferences of their distributed competence.

Given these challenges, where might we go in improving assessments of ELL students? Baker (2007) provides some suggestions that would seem quite appropriate. She introduced the notion of *qualification*—a "trusted, certified accomplishment" in school, but also possibly outside of classrooms, by a student that would augment accountability assessment performance as indicators of a broader range of student achievement, beyond learning of specific skills. A qualification would involve choices made by students about preferred achievement areas, goals, and a school-sanctioned crediting system for students regarding these accomplishments. A crediting system would need to certify what would count as significant learning experiences, organization of work and effort, required judgment and certification of expertise by others in

an achievement area, and validation of such a system based on inferences of what a qualification means. Instead of a test score, a student's qualification would represent a set of significant accomplishments in a complex achievement area that would include but go beyond subject matter expertise, such as application of knowledge and skills to a public service project, complex artistic performance, or some other complex domain of applied knowledge and skills.

It will be very interesting to consider how Baker's call for a qualification system of assessment could be made relevant to ELL assessment in U.S. schools and elsewhere. The immigrant and transnational experience of ELLs presents these students with many linguistic, cultural, and social challenges that, when overcome, represent the resiliency and adaptability of humans and human intelligence at large. This is a new frontier for ELL assessment. The contemporary notion of setting learning and assessment goals through "backward mapping" seems relevant here (Wiggins & McTighe, 1998). In backward mapping, a curriculum, instruction, and accompanying classroom assessments are designed from the top down, starting with a clarification and detailed statement of educational goals, followed by specification of instructional practices and assessments that could be used to achieve goals. Viewed as a dialectical process, what has become popularly known as backward mapping can be represented from a CHAT perspective as being framed by four questions (Durán, Escobar, & Wakin, 1997):

- What is achievement?
- What activities give rise to achievement?
- What evidence is there of achievement in activities?
- What are the socioeducational consequences of the foregoing?

These four questions are interlocking. They form a conceptual schema not unlike the schema underlying the paradigm of critical pedagogy, which sets forth an unending dialectical cycle of investigating how to conceptualize important educational and social problems, how to devise and implement strategies to solve problems, and how to evaluate the success of problem solving and then go on to better reframe problems and strategies and implementation of strategies (Wink, 2005). Deeply understanding these questions and how to approach answering them from a CHAT perspective commensurate with views on how to establish new understandings of what achievement means for ELL and other students is nontrivial. Durán and Szymanski (1997) postulate that students' moment-to-moment interpretation of learning activity as social and cultural practice and as manifestation of identity and agency are central. The notion of consequential progression applies. How might past and current learning experiences add up to form trajectories for personal development and new forms of learning? What activities and evidence provide insights and data into this historical as well as immediately situated development? How does our thinking about education and its goals get transformed by considering enhanced notions of achievement akin to those described by Baker (2007) and others? Within the boundaries of an academic year, it has been possible to use ethnography to trace students' progression through

learning sequences that provide interactional evidence of how young elementary school ELL students form ongoing identities as early learners of social science (Putney et al., 2000) or natural science (Reveles, Kelly, & Durán, 2007). How do we augment such research to track both qualitative and quantitative evidence of students' longer range development of identities that can be meaningful parts of students' lives and can show evidence of Baker's endorsements? These are important next questions.

In closing, it is helpful to return to the work of Moss et al. (2006). Along with Rochex (2006), these investigators remind us of the deep interplay that exists between the expectations of educational policymakers at different levels of government, and even across nations, and assessments administered to students. There is no escaping that national and major jurisdictional levels of educational governance motivate and set the expectations of local educational practitioners, teachers, parents, and students themselves regarding the meaning and implications of assessments for schooling accountability purposes and for student high-stakes purposes such as high school graduation (McDonnell, 2004). The reality is that very few educational stakeholders are able to comprehend the technical characteristics of assessments well, if at all. These stakeholders also, by and large, have only a limited grasp of the meaning and implications of terms such as *English-language learner* and *English proficient*, and the complexity of this limited understanding is also caught in scientific and social science debates of the meaning of such terms, as well as in policy debates.

Much work remains to be done to better ground the field of assessment of ELLs. As part of this process, in closing, it is important to understand that the findings and issues cited in this chapter are of equal relevance to all students. Focus on ELLs helps us understand better how important historical, cultural, and linguistic background differences have to be taken into account in interpreting the results of assessments and the design of new assessment strategies aligned to the characteristics of students. Also at issue is how to locate and treat educational practitioners as full partners in pursuing these matters, and it is good to see at the close of 2007 that in the United States, federally funded efforts are under way to create consortia among states to support teachers' development of formative assessments based on general backward-mapping strategies that require discourse and dialogue among teachers, parents, policymakers, and students regarding what is desired and valued as learning and what evidence can count for learning (Cech, 2007). This will certainly contribute to understanding better how locally developed assessments grounded in authentic learning activity can complement information currently provided by large-scale state achievement assessments, though much work will be required. All of the concerns discussed here also deserve further exploration in different national contexts and comparatively across countries in light of variations in nations' education policies and conception of desired schooling outcomes (see Rochex, 2006).

NOTES

¹The No Child Left Behind stipulation that states develop English language development (ELD) standards requires states to set explicit expectations of about the basic English (reading, writing, listening, and speaking) skills required of ELLs learning English throughout the K–12 grade

span. Typically, these ELD standards are arranged by specific grades or in clusters of adjacent grades, so that they are appropriate to the English language requirements of curricula encountered by students.

English language proficiency (ELP) tests that are administered to English language learner (ELL) students are built to sample the requirements of ELD standards at a given school grade or grade band appropriate to a student. They should not be confused with tests of English language arts administered annually to all eligible students as part of large-scale assessments. The former are to gauge ELLs' mastery of English as a second language and readiness to receive instruction in English. The latter are to assess students' command of English as found and taught in regular English language arts classes at a target grade level. In reality, ELD and English language arts (ELA) competencies are expected to blend for ELLs. As ELLs become more competent in English, they acquire foundational skills for meeting ELA standards.

²As mentioned, the discussion regarding cultural historical activity theory (CHAT) as a useful perspective from which to reconceptualize classroom and schooling achievement is not intended to suggest that there are no other perspectives based on alternative statements of theory, research, and practice that address similar concerns with diverse populations of students. For example, Banks and Banks (2001) provide comprehensive coverage of such topics in their *Handbook of Research on Multicultural Education* from multiple disciplinary perspectives.

Much of this relevant theory, research, and practitioner guidance not described explicitly as under the aegis of CHAT often is described as founded on sociocultural, sociolinguistic, or constructivist approaches to learning, instruction, and classroom communication. Early work of this sort emanated, for example, from the writings and research compilations of investigators such as Au (1980); Cazden, John, and Hymes (1972); Cook-Gumperz (1986); Erickson and Mohatt (1982); Erikson and Shultz (1981); Green (1983); Green and Wallat (1981); Heath (1983); and Trueba, Guthrie, and Au (1981). Works such as those mentioned were among the first investigations to address how analysis of discourse and interaction laid bare how perception of interactional context and participation in face-to-face interaction was centrally dependent on the cultural, social, and linguistic resources and background of communicants in schooling and other settings. Investigators such as Saravia-Shore and Arvizu (1992), in their edited volume *Cross-Cultural Literacy*, were among the first to extend concern for ethnographic and sociolinguistic study of classroom communication so that it included examination of how schooling policies, and in particular, resistance to bilingual education policies, affected everyday classroom interaction.

Wells (1999), in his volume *Dialogic Inquiry*, was among the first sociolinguistic and discourse analyst researchers to link Vygotskian theory (and issues tied to CHAT) to classroom research. He was also influential in bringing increased attention to the research of functional linguists such as Halliday and British, European, and Australian researchers on what later came to be known as the "new literacies" research paradigm (see, e.g., Gee, 2004; and Barton, Hamilton, and Ivanic, 2000). The new literacies paradigm has proven quite valuable in suggesting that the functions and linguistic form of spoken and written texts arise from socialization to cultural practices and instrumental goals associated with texts.

It is not possible in the present chapter to survey the extensive literature that followed these earlier works that has put increased attention on how sociolinguistic, ethnographic, and new literacies research has addressed how learning in schooling and other contexts is driven by interactive processes. That stated, readers are referred to volumes by Bloome, Carter, Christian, Otto, and Shuart-Faris (2005) and Schleppegrell and Colombi (2002) for examples of recent research in this area. The latter volume is of particular relevance to this chapter because it address how second-language learners (as well as first-language learners) acquire competence in controlling the linguistic and discourse features of texts tied to language as used in academic settings. Freeman and Freeman (2002) do an excellent job of pursuing these issues from the perspective of teachers and focus on the special challenges of addressing the academic literacy needs of long-term high school ELL students.

It is important to note that all of the research cited above has involved primary use of qualitative research methods and virtually no use of quantitative test data or use of experimental design techniques. One upshot of this is that research of this sort has been excluded from consideration in a recent major synthesis of findings in a chapter on findings of sociocultural research on literacy development of second-language learners by August and Shanahan (2006). This omission is logical on the basis of empiricist criteria used in this volume to define what counts as defensible research findings. The criteria are appropriate only for studies that permit comparison groups and use quantitative methods. However, these criteria for valid findings do not hold for researchers in the traditions of sociocultural, sociolinguistic, ethnographic, and CHAT perspectives. See Erickson (2007) and Moss, Girard, and Haniford (2006) for a further explication of this matter.

³Also, comprehensive approaches to reconceptualization of classroom and schooling achievement are becoming more and more prevalent from a CHAT perspective. For example, an excellent resource in this regard are papers and reports published electronically by the Center for Research on Education, Diversity, and Excellence (CREDE) at the Internet site <http://crede.berkeley.edu/index.html>. The CREDE Web site is particularly powerful in its attention to ways that teacher training, staff development, and instructional practice promote effective instruction, given students' cultural and linguistic backgrounds.

REFERENCES

- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14.
- Abedi, J. (2006, August). *The use of accommodations for English language learners: Latest research and current practice*. Presentation at the LEP Partnership Meeting, Washington, DC.
- Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practices*, 25(4), 36–46.
- Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74, 1–28.
- Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP*. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Adamson, H. D. (1993). *Academic competence: Theory and classroom practice. Preparing ESL students for content courses*. New York: Longman.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Au, K. H. (1980). Participation structures in a reading lesson with Hawaiian children: Analysis of a culturally appropriate instructional event. *Anthropology and Education Quarterly*, 11, 91–115.
- August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners*. Mahwah, NJ: Lawrence Erlbaum.
- Bailey, A. (2006). *The language demands of school: Putting academic English to the test*. New Haven, CT: Yale University Press.
- Baker, E. (2007, April). *The end(s) of testing*. Presidential address given at the annual meeting of the American Educational Research Association. Retrieved September 1, 2007, from <http://www.softconference.com/Media/WMP/270409/s40.htm>

- Baker, E., & Linn, R. L. (2004). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47–72). New York: Teachers College Press.
- Banks, J., & Banks, C. (Eds.). (2001). *Handbook of research on multicultural education*. San Francisco: Jossey-Bass.
- Barton, D., Hamilton, M., & Ivanic, R. (Eds.). (2000). *Situated literacies: Reading and writing in context*. New York: Routledge.
- Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Bloome, D., Carter, S. P., Christian, B. M., Otto, S., & Shuart-Faris, N. (2005). *Discourse analysis and the study of classroom language and literacy events: A microethnographic perspective*. Mahwah, NJ: Lawrence Erlbaum.
- Callahan, R. M. (2005). Tracking and high school English learners: Limited opportunity to learn. *American Educational Research Journal*, 42, 305–328.
- Cazden, C., John, V. P., & Hymes, D. (Eds.). (1972). *Functions of language in the classroom*. New York: Teachers College Press.
- Cech, S. (2007, August 15). 10-state pilot preparing teachers to develop tests. *Education Week*, p. 10.
- Center for Education Policy. (2006). *State high school exit exams: A challenging year*. Washington, DC: Author.
- Cole, M. (1996). *Cultural psychology*. Cambridge, MA: Harvard University Press.
- Cook-Gumperz, J. (Ed.). (1986). *The social construction of literacy*. New York: Cambridge University Press.
- Durán, R. P. (1994). Cooperative learning for language minority students. In R. A. DeVillar, C. J. Faltis, & J. Cummins (Eds.), *Cultural diversity in schools: From rhetoric to practice* (pp. 145–159). Albany: State University of New York Press.
- Durán, R. P. (2006). *State implementation of NCLB policies and interpretation of the NAEP performance of English language learners*. Palo Alto, CA: American Institutes for Research, NAEP Validity Studies.
- Durán, R. P., Escobar, F., & Wakin, M. (1997). Improving classroom instruction for Latino elementary school students: Aiming for college. In M. Yepes-Baraya (Ed.), *1996 ETS invitational conference on Latino education issues* (pp. 39–53). Princeton, NJ: Educational Testing Service.
- Durán, R. P., & Moreno, R. (2004). Do multiple representations need explanation? The role of verbal guidance and individual differences in multimedia mathematics learning. *Journal of Educational Psychology*, 96(3), 492–503.
- Durán, R. P., & Szymanski, M. (1997). *Assessment of transfer in a bilingual cooperative learning curriculum* (CSE Technical Report 450). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Erickson, F. (2007). Some thoughts on “proximal” formative assessment of student learning. In P. Moss (Ed.), *Evidence in decision making: Yearbook of the National Society for the Study of Education* (Vol. 106, pp. 186–216). Malden, MA: Blackwell.
- Erickson, F., & Mohatt, G. (1982). Cultural organization of participation structures in two classrooms of Indian students. In G. Spindler (Ed.), *Doing the ethnography of schooling. Educational anthropology in action* (pp. 132–174). New York: Holt, Rinehart & Winston.
- Erickson, F., & Shultz, J. (1981). When is a context? Some issues and methods in the analysis of social competence. In J. Green & C. Wallat (Eds.), *Ethnography and language in educational settings* (pp. 147–160). Norwood, NJ: Ablex.
- Frances, D. J., Kieffer, M., Lesaux, N., Rivera, H., & Rivera, M. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. Houston, TX: University of Houston, Center on Instruction, Texas Institute for Measurement, Evaluation, and Statistics.

- Freeman, Y., & Freeman, D. (with Mercuri, S.) (2002). *Closing the achievement gap. How to reach limited-formal-schooling and long-term English learners*. Portsmouth, NH: Heinemann.
- Gee, J. P. (2004). *Situated language and learning. A critique of traditional schooling*. New York: Routledge.
- Gee, J. P. (2007). Reflections on assessment from a sociocultural perspective. In P. Moss (Ed.), *Evidence in decision making: Yearbook of the National Society for the Study of Education* (Vol. 106, pp. 362–375). Malden, MA: Blackwell.
- Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7*, 121–140.
- Green, J. L. (1983). Research on teaching as a linguistic process: A state of the art. In E. W. Gordon (Ed.), *Review of Research in Education, 10*, 151–252.
- Green, J. L., & Wallat, C. (Eds.). (1981). *Ethnography and language in educational settings*. Norwood, NJ: Ablex.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.
- Heath, S. (1983). *Ways with words: Language, life, and work in communities and classrooms*. New York: Cambridge University Press.
- Holland, D., & Quinn, N. (Eds.). (1987). *Cultural modes in language and thought*. Cambridge, UK: Cambridge University Press.
- Koenig, J., & Bachman, L. (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment*. Washington, DC: National Academy Press.
- Kopriva, R. (in press). *Improving testing for English language learners*. London: Routledge.
- Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20.
- Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*, 491–530.
- Leontiev, A. N. (1981). *Problems of the development of the mind* (M. Kopylova, Trans.). Moscow: Progress.
- Martiniello, M. (2007). *Linguistic complexity and differential item functioning (DIF) for English language learners (ELL) in math word problems*. Unpublished doctoral thesis, Harvard University, Cambridge, MA.
- Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES 2000–473). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.
- Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Miller, G. J. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
- Mislevy, R. (1997). Postmodern test theory. In A. Lesgold, M. Feuer, & A. Black (Eds.), *Transitions in work and learning. Implications for assessment* (pp. 180–199). Washington, DC: National Academy Press.

- Mislevy, R. J., & Huang, C. W. (2006). *Measurement models as narrative structures*. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing and Center for the Study of Evaluation.
- Moss, P. A., Girard, B. J., & Haniford, L. C. In J. Green & A. Luke (Eds.), (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.
- Nelson, K. (1996). *Language in cognitive development: The emergence of the mediated mind*. Cambridge, UK: Cambridge University Press.
- Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior, 19*, 221–243.
- Popham, J. (2006). *What are some general observations about 21st century skills and assessment?* Presentation given at the annual conference of the Council of Chief State School Officers, San Francisco.
- Putney, L., Green, J., Dixon, C., Durán, R., & Yeager, B. (2000). Consequential progressions: Exploring collective-individual development in a bilingual classroom. In C. Lee & P. Smagorinsky (Eds.), *Vygotskian perspectives on literacy research* (pp. 86–126). New York: Cambridge University Press.
- Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal, 30*, 523–553.
- Revels, J., Kelly, G., & Durán, R. P. (2007). A sociocultural perspective on mediated activity in third grade science. *Cultural Studies of Science Education, 1*, 467–495.
- Rivera, C., & Collum, E. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Lawrence Erlbaum.
- Rochex, J.-Y. (2006). Social, methodological, and theoretical issues regarding assessments: Lessons from a secondary analysis of PISA 2000 literacy tests. *Review of Research in Education, 30*, 163–212.
- Sanchez, G. (1934). Bilingualism and mental measures: A word of caution. *Journal of Applied Psychology, 18*, 765–772.
- Saravia-Shore, M., & Arvizu, S. (Eds.). (1992). *Cross-cultural literacy: Ethnography of communication in multiethnic classrooms*. New York: Garland.
- Scarcella, R. (2003). *Academic English: A conceptual framework*. Santa Barbara: University of California, Santa Barbara, Linguistic Minority Research Institute.
- Schleppegrell, M., & Colombi, M. (2002). *Developing advanced literacy in first and second languages: Meaning with power*. Mahwah, NJ: Lawrence Erlbaum.
- Scribner, S. (1979). Modes of thinking and ways of speaking: Culture and logic reconsidered. In R. Freedle (Ed.), *New directions in discourse processing* (pp. 223–243). Norwood, NJ: Ablex.
- Sireci, S. G. (2005). Unlabeling the disabled: A psychometric perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3–12.
- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). Amherst: University of Massachusetts-Amherst, School of Education.
- Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaption process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–116). Mahwah, NJ: Lawrence Erlbaum.
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–348). Mahwah, NJ: Lawrence Erlbaum.
- Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record, 108*(11), 2354–2379.

- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
- Stansfield, C. W. (1996). Content assessment in the native language. *Practical Assessment, Research and Evaluation*, 5(9). Retrieved June 3, 2007, from <http://PAREonline.net/getvn.asp?v=5&n=9>
- Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 20(2), 189–207.
- Suarez-Orozco, M. (2001). Globalization, immigration, and education: The research agenda. *Harvard Education Review*, 71(3), 345–365.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Tharp, R., & Gallimore, R. (1988). *Rousing minds to life*. Cambridge, MA: Harvard University Press.
- Torrance, H., & Pryor, J. (1998). *Investigating formative assessment*. Philadelphia: Open University Press.
- Trueba, H., Guthrie, G., & Au, K. (Eds.). (1981). *Culture and the bilingual classroom: Studies in classroom ethnography*. Rowley, MA: Newbury House.
- Valdes, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.
- van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195–211.
- Vijver, F. J. R., van de, & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–64). Mahwah, NJ: Lawrence Erlbaum.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.
- Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and Practice*, 18(4), 10–16.
- Wells, G. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education*. New York: Cambridge University Press.
- Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.
- Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Wink, J. (2005). *Critical pedagogy: Notes from the real world* (3rd ed.). New York: Allyn & Bacon.
- Wong Fillmore, L. (1976). *The second time around: Cognitive and social strategies in second language acquisition*. Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.
- Zehler, A. M., Fleischman, H. F., Hopstock, P. J., Stephenson, T. G., Pendzick, M. L., & Sapru, S. (2003). *Descriptive study of services to LEP students and LEP students with disabilities: Volume I. Research report. Final report submitted to U.S. Department of Education, Office of English Language Acquisition*. Arlington, VA: Development Associates.