

In What Language Should English Language Learners Be Tested?

GUILLERMO SOLANO-FLORES AND ELISE TRUMBULL

A fundamental notion in test validity is that the scores of a test should not be significantly influenced by factors other than the skills and knowledge that the test is intended to measure (Messick, 1989). This notion is critical to examining language issues in testing. Since language is the primary vehicle for administering tests, academic achievement measures are largely influenced by language variables, even when the constructs of interest are not directly related to language. That tests inevitably assess language skill in addition to the target knowledge domain (AERA/APA/NCME, 1999; Kiplinger *et al.*, 2000; LaCelle-Peterson and Rivera, 1994) appears not to be an overstatement.

Even when students are native, fluent speakers of the language in which a test is administered, their performance on that test is shaped by their ability to understand printed text or spoken language and to express their ideas orally or in writing (Popham, 2001). Hence, one question that must be addressed in determining in what language students should be tested is the status of their proficiency in the languages in question. For that reason, in this chapter, it is necessary to discuss language assessment—even though the book is focused on academic assessment.

Students' performance on a test is also shaped by the ability of test developers to write test items clearly, the ability of test administrators to provide test directions orally or in printed text properly, and—in the case of open-ended items—the ability of test scorers to interpret the examinees' responses properly.

CHAPTER 8

In What Language Should English Language Learners Be Tested?

GUILLERMO SOLANO-FLORES AND ELISE TRUMBULL

A fundamental notion in test validity is that the scores of a test should not be significantly influenced by factors other than the skills and knowledge that the test is intended to measure (Messick, 1989). This notion is critical to examining language issues in testing. Since language is the primary vehicle for administering tests, academic achievement measures are largely influenced by language variables, even when the constructs of interest are not directly related to language. That tests inevitably assess language skill in addition to the target knowledge domain (AERA/APA/NCME, 1999; Kiplinger *et al.*, 2000; LaCelle-Peterson and Rivera, 1994) appears not to be an overstatement.

Even when students are native, fluent speakers of the language in which a test is administered, their performance on that test is shaped by their ability to understand printed text or spoken language and to express their ideas orally or in writing (Popham, 2001). Hence, one question that must be addressed in determining in what language students should be tested is the status of their proficiency in the languages in question. For that reason, in this chapter, it is necessary to discuss language assessment—even though the book is focused on academic assessment.

Students' performance on a test is also shaped by the ability of test developers to write test items clearly, the ability of test administrators to provide test directions orally or in printed text properly, and—in the case of open-ended items—the ability of test scorers to interpret the examinees' responses properly.

When the examinees are English language learners (ELLs), the challenges posed by language as a source of invalidity are greatly magnified. For instance, even when ELLs "know" the vocabulary used in a test item, they are not likely to have the elaborated word knowledge of a native speaker of English (Garcia, 2000; Laufer, 1997), and their understanding of that test item may be highly influenced by contextual factors and their socio-cultural experiences (see Durán, 1989). As a consequence, their interpretation of test instructions or test items may be compromised.

Inclusion of ELL students in national and state assessment systems is driven by the aim of obtaining indicators of school performance based on testing as many students as possible and supported by the assumption (or hope) that effective approaches are available or can be devised for minimizing varying proficiency in English as a source of test invalidity (see Kopriva, 2000). An approach that may first come to mind is to test ELLs in their first language (L1). Another approach consists of testing them in English, their second language (L2), with certain accommodations—such as simplifying the linguistic features of items or providing students with glossaries with translations of potentially difficult words—intended to reduce or eliminate the effects of language proficiency on test performance (see reviews by Abedi *et al.*, 2000; Rivera *et al.*, 2006; Sireci *et al.*, 2003).

In this chapter, we discuss three factors that are critical to both formulating national or state testing policies on the language used to test ELLs and examining the appropriateness of such policies:

1. the soundness of the criteria used to identify students as either ELLs or English proficient;
2. the accuracy of classifications according to their level of proficiency in English; and
3. the effectiveness of testing practices in addressing the English proficiency level of ELLs.

In the first section, *Defining ELLs*, we examine the adequacy of official definitions of the ELL individual according to current knowledge of bilingualism. In the second section, *Classifying ELLs*, we examine the appropriateness of current approaches for measuring English proficiency according to current knowledge in the field of bilingual development and second language acquisition. In the third section, *Testing ELLs*, we examine the limitations of testing ELLs in L1 and in English. We also discuss how, in order to be sound, decisions in favor of one or another language need to be based on consideration of the quality of the information on the language proficiency of ELLs and the institutional capacity to properly implement testing in each language and to provide effective testing

accommodations. In the fourth section, *Future Directions*, we present alternative testing models intended to address the limitations of existing approaches to ELL testing in both L1 and English. Such testing models are based on recent research showing that the performance of ELLs is unstable across languages.

Defining ELLs

Evidence from research on language development and second language acquisition shows that the term "English language learner" refers in actuality to an individual who may exhibit any of an almost indefinite number of patterns of proficiency in L1 and English (Brisk, 2006; Romaine, 1995). Hence, it is more useful to think of an ELL as defined by a broad spectrum of language proficiencies in both L1 and L2. This research also shows that, in order to be accurate, any definition of ELL students and any strategy intended to identify ELL students should be based on the recognition that populations (either monolingual or bilingual) vary considerably due to factors inherent to language, culture, and experience, and that bilingual individuals vary tremendously in their development of L1 and L2. In this section we examine these factors. Then we discuss how the interaction of these factors produces different patterns of bilingualism among ELLs. Finally, we discuss the limitations of official definitions of ELL.

Variation Among ELLs

We identify six factors as critical to properly identifying ELLs and to making appropriate decisions concerning each ELL individual's testing: 1) relative language proficiency, 2) proficiency variation across language modes, 3) minimum bilingual development, 4) language dominance, 5) language variation, and 6) educational variation.

RELATIVE LANGUAGE PROFICIENCY

An individual's proficiency in a language is a relative, rather than an absolute, condition. Strictly speaking, total proficiency does not exist even in the native language, as even amongst the most proficient individuals there are always aspects of their linguistic skills that can be improved. Among the kinds of linguistic knowledge and skill that may continue to develop are word meanings, idiomatic expressions, complex syntax, spelling, ways of expressing ideas in writing, reading fluency, and ways of interacting socially through language (Grosjean, 1989).

VARIATION ACROSS LANGUAGE MODES

An individual's proficiency in a language varies across listening, speaking, reading, and writing modes. A person's proficiency in a language is rarely

the same across these four language modes (Wei, 2000). Moreover, different individuals have different sets of strengths and weaknesses in each mode. Whereas two given individuals could be thought of as having comparable, high levels of proficiency, one could be better at listening and speaking and the other better at reading and writing in L1 or L2 (Fishman, 1965). Although there are practical advantages of using the word "proficient" as an overall condition of an individual's command of a language, its use unintentionally implies that a person's proficiency is the same across language modes. In reality, rarely is an individual equally proficient in the four modes.

MINIMUM BILINGUAL DEVELOPMENT

In spite of their limited proficiency in English, virtually all ELLs have at least some minimum level of competency in L2. For example, they may be able to understand some words when they hear or read them, or they may be able to speak or write short phrases—rudimentary skills which result even from a limited exposure to L2 (Mackey, 1962). Virtually all ELLs can be thought of as bilinguals, albeit "incipient bilinguals" in some cases (Diebold, 1964, cited in Romaine, 1995, p. 11). Although the word "bilingual" evokes the image of someone who is equally fluent in two languages, rarely is an individual entirely bilingual (Valdés and Figueroa, 1994). Thus, terms such as "English language learner" or "limited English proficient" may lead to underestimating the proficiency of some ELLs in L2 or overestimating their proficiency in L1.

LANGUAGE DOMINANCE

Bilingual individuals tend to be better in one language than the other. Terms such as "English-dominant" or "Spanish-dominant" are used to characterize an individual in terms of his or her preferred language. Language dominance should be interpreted as a condition that is shaped by factors such as context and personal history (Wei, 2000). For example, even an individual with a high proficiency in L2 may use L1 for tasks that involve automaticity in the use of language, such as counting or performing arithmetic computations. Or, individuals who have developed their academic language in L2 may find it extremely difficult to use L1 in writing about subject matter learned in L2. Moreover, in some cases, bilingual individuals may be better in L1 than in L2 for one language mode but better in L2 than in L1 for another language mode.¹

DIALECT VARIATION

Different communities of individuals within the same broad linguistic group may vary considerably in the implicit language rules they follow and

in the uses to which they put language. Even within populations of native speakers of the same language, ELLs may vary tremendously in their dialects of both L1 and L2.² Such variation may result from factors such as the diversity of places of origin (geography) and contact with other linguistic groups. Many people speak more than one dialect of a language or languages, and many students learn—at least comprehend—the dialect of school as well as home, when those differ. Discourse features may also differ, as in ways of organizing oral or written narrative or responding to questions (Cazden, 1999; Gee, 1996; Heath, 1983). These pragmatic or social aspects of language are probably more important than differences in vocabulary, pronunciation and syntax because they differ from what is expected in school; more important, instead of being identified as language differences, they may be wrongly identified as cognitive deficits. Such differences also have implications for test validity (Deyhle, 1987; Estrin and Nelson-Barber, 1995; Swisher and Deyhle, 1992).

Patterns of Language Proficiency

The linguistic strengths and weaknesses of each ELL student vary with language mode (oral, written), situation (more formal, less formal), role of language (thinking, comprehending, producing), and social domain (e.g., family, school) (Fishman, 1965; Romaine, 1995). Depending upon the nature of the task, the context, and the conditions, an ELL student is likely to be more comfortable or competent in one language or the other (Romaine, 1995; Valdés and Figueroa, 1994). One might better think of an individual's "dominance configuration" (Fishman, 1965), or the pattern of language proficiency represented by the interaction of all of these elements. In this section, we examine possible patterns of variation in language proficiency based solely on the language mode in question: listening, speaking, reading, or writing. Naturally, if we attempted to account for task demands, social context, and conditions surrounding the task, it would be possible to generate an endless number of patterns of far greater complexity.

Figure 8.1 shows the proficiencies in English and in Language X of eight hypothetical individuals according to language mode. For the purposes of our discussion, we make the assumption that linguistic proficiency in a language can be measured accurately. We know that in reality this is extremely difficult, as proficiency may vary considerably with context. For instance, measures of reading proficiency may have some limitations because they use reading materials that are only samples from the universe of materials that a student could possibly need to read. Moreover, it is well known that the proficiencies of linguistic minority students in both English and their native language are seldom assessed properly (De Avila, 1987;

MacSwan, 2000; Saunders and O'Brien, 2006). Despite its limitations, however, the figure is useful for illustrating the imprecision of such terms as "proficient," "monolingual," "bilingual," or ELL. If an individual's proficiencies in the four language modes were exactly the same across languages, the lengths of the bars for the four language modes would be the same. Also, if so-called "bilinguals" were equally proficient in two languages, the lengths of the bars on all language modes would be symmetrical across languages. In reality, this kind of symmetry would be a rare occurrence because even individuals who are deemed fully bilingual tend to be dominant in one language (Romaine, 1995).

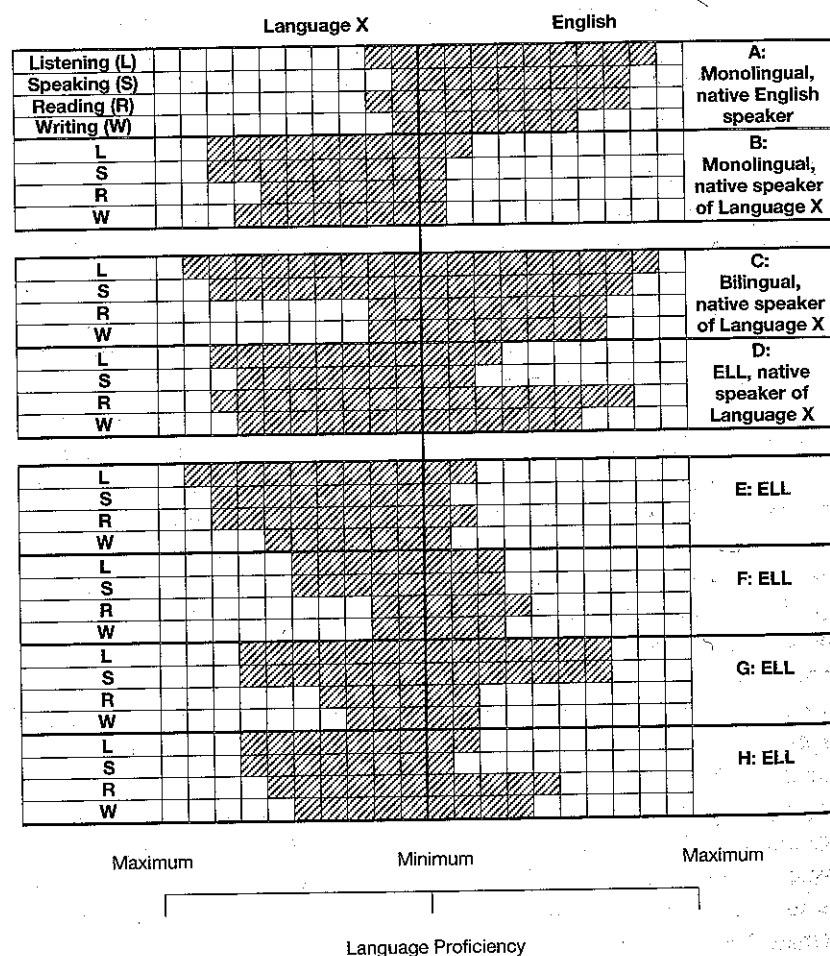


Figure 8.1 Language proficiencies of eight hypothetical persons. The length of the bars, on the left and right side of the chart represent, respectively, proficiency in L1 and L2.

The first two individuals are called "monolingual." Person A is a monolingual native speaker of English, and Person B is a monolingual native speaker of Language X. But the monolingual condition of each is not absolute. Both have developed minimal listening and reading proficiencies in their second languages (e.g., some words that they can pick up in a conversation or others that they can recognize in printed form). This may have happened incidentally; for example, from their exposure to the second language through the media or through contact with native speakers.

Persons C and D are identified as "bilingual." However, their patterns of bilingualism are different. Person C could be a native English speaker who was born in and has received formal instruction in the U.S. but has had continuous contact with speakers of Language X through conversation. In contrast, Person D could be someone who has received formal instruction in both Language X and English. Whereas Person C has acquired English through social interaction, Person D has acquired it mostly through formal instruction. As a result of these differences, Person C speaks English fluently, and her reading and writing skills in English are strong. In contrast, Person D has strong reading and writing skills but moderate listening and speaking skills in English. Some would be surprised to realize that Person C has such limited reading and writing skills in Language X after hearing her speak in the two languages. Also, from Person D's conversational skills in English, many would find it difficult to call this individual "bilingual," in spite of his having similar reading and writing skills in the two languages.

Persons E–H are students classified as ELLs. They are just examples of the many patterns of language dominance in ELLs that result from different kinds of language development in English and their native languages (see Stevens *et al.*, 2000). Of the four ELLs shown in the figure, E is probably the closest to what many people have in mind when they talk about ELLs—individuals with reasonably strong skills in their native language and poorly developed skills in the four language modes of English. The reality is that this pattern of language dominance is far from being the only one among ELLs. Indeed, E might well be a rare case of ELL.

A more complex pattern can be observed in F. He is more proficient in his first language than in English in the listening and speaking modes. However, he is more proficient in reading and writing in English because it is in this language that he has received formal instruction. His linguistic skills are moderate in both languages.

Overall, G and H are more proficient in English than F. However, there are several subtle but important differences between them in their proficiencies in both languages. First, the proficiencies of G tend to be

symmetrical across languages, with better listening and speaking skills than reading and writing skills. Second, the reading and writing skills in both languages are better for H than G. Third, whereas the listening and speaking skills in English are better for G, H is more proficient in reading and writing in English. H is more like those students who come to school “fluent” in their home language but who have developed little of what is called “academic language proficiency”³ in that language (see Cummins, 2000). It should be noted that academic language is not restricted to the written mode; there are oral forms of discourse that are highly academic, some general and some particular to the disciplines (Schleppegrell, 2004).

Figure 8.1 shows that many ELLs are partially proficient in their native languages and partially proficient in English. It also gives an idea of the enormous diversity that may exist even within the same broad linguistic group of ELLs. Classifications of students as “ELL” or “LEP” (limited English proficient) are generalizations at best, as each category may actually describe a wide variety of types of ELL or bilingual students (see Aguirre-Muñoz and Baker, 1997; Casanova and Arias, 1993). As a consequence, classifications of ELLs based only on criteria such as whether English is their native language or whether they are proficient in one or two language modes in English are flawed. And this is true even if we assume that proficiency is properly assessed for those language modes. As discussed, individuals—especially ELLs—may vary considerably in their linguistic proficiencies across language modes. Moreover, in spite of their limitations in English, some ELLs may be more fluent in English than in their first language for some modes.

Additional Sources of Linguistic Differences

Individuals with similar patterns of bilingualism may vary considerably in their use of L1 and L2 due to differences in cultural background, dialect, and schooling. These factors shape the accuracy with which ELLs can be identified. Variation in cultural experiences encompasses both home culture and the culture of school. Whereas school culture is relatively homogeneous in the U.S., students’ home cultures are quite diverse. Home culture experiences include a student’s opportunities to develop proficiency with the native language of the family (including uses of language in social contexts such as school) and to absorb the values and beliefs of the family and cultural community (including those related to teaching and learning) (Heath, 1983; Pease-Alvarez and Vasquez, 1994; Trumbull *et al.*, 2004). Parallel experiences take place within the school; however, they may not be recognized as cultural in nature (Trumbull *et al.*, 2004).

ELL students may have close ties to family members in their parents’ country of origin and travel with some frequency to that country—not

unusual for Mexican-American students, for instance (Delgado-Gaitan, 1994; Solano-Flores *et al.*, 2002). They may participate in weekend language schools where they continue to learn L1 to a high level of proficiency, as do many Chinese American students (Wang, 1996). Also, they may be part of a multi-generational family in the U.S., with many of its longstanding traditions intact and opportunities to use L1 in a variety of contexts, as is the case with many American Indian students (McCarty, 2002).

Other ELLs are more thoroughly immersed in the dominant culture and have fewer opportunities to maintain and develop the language of home. Language is, of course, a fundamental cultural resource and the most important tool in carrying culture on from generation to generation (Fishman, 1989; Hinton, 1995; Wong-Fillmore, 1991). However, there is strong evidence that even when families of recent immigrants shift to English, often within a generation, they continue to pass on their traditional values—even in the ways they use English (Delgado-Gaitan, 1994; González *et al.*, 2001). So, maintenance of a family’s heritage language is by no means the only measure of persistence of cultural values and child-rearing goals, both of which affect students’ ways of using language and participating in school activities, including testing (Swisher and Deyhle, 1992; Rothstein-Fisch *et al.*, 2003).

Limitations of Official Definitions of “ELL”

Strictly speaking, sound classifications of ELLs for instructional placement decisions should be based on examining the proficiencies of each student in the four language modes in both English and L1. At the very least, the decision about the language to use to test ELLs should be made separately for each individual and should be based on examining their proficiencies in at least reading and writing—the language modes most commonly used to administer tests—in both L1 and L2.

Unfortunately, legislation concerning the identification of ELLs does not reflect the complexity of proficiency across languages and language modes discussed above. Standard definitions of “English language learner” are quite minimalist. The National Clearinghouse for English Language Acquisition defines ELLs as “students whose first language is not English and who are in the process of learning English” (National Clearinghouse for English Language Acquisition, n.d.). The federal government, along with most states and districts, uses the term “limited English proficient” (LEP) instead, which it defines as “those students who have insufficient English to succeed in English-only classrooms” (Lessow-Hurley, 1991). More often than not, the emphasis in legal definitions is on English alone, and attention is not given to the level of proficiency ELLs may have attained in L1—something that must be considered in decisions about the language

of testing (Saunders and O'Brien, 2006). Focusing on English alone in characterizing these students is to overlook other linguistic resources they may well have and may lead to wrong decisions concerning the language of testing or wrong inferences about their ability (Figueroa and Hernandez, 2000).

Another limitation of official definitions of ELLs has to do with the level of analysis used to characterize these students. Take, as an example, the No Child Left Behind Act of 2001 (NCLB), which defines an ELL as an individual

- (A) who is aged 3 through 21;
- (B) who is enrolled or preparing to enroll in an elementary school or secondary school;
- (C) (i) who was not born in the United States or whose native language is a language other than English;
- (ii) (I) who is a Native American or Alaska Native, or a native resident of the outlying areas; and
- (II) who comes from an environment where a language other than English has had a significant impact on the individual's level of English language proficiency; or
- (iii) who is migratory, whose native language is a language other than English, and who comes from an environment where a language other than English is dominant; and
- (D) whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual—
 - (i) the ability to meet the State's proficient level of achievement on (academic assessments such as reading, mathematics, and science);
 - (ii) the ability to successfully achieve in classrooms where the language of instruction is English; or
 - (iii) the opportunity to participate fully in society.

Only one of the four main components of this definition (D) addresses issues that are critical to the condition of being ELL: Components A and B set the boundaries of the demographic characteristics of the segment of the population of interest, for assessment purposes, and component C refers to attributes (e.g., place of birth, being an immigrant) that are associated with but do not necessarily determine the condition of being limited English proficient. Unfortunately, while Component D recognizes the fact that proficiency in a language may vary across language modes, it does not

actually provide a definition of ELLs but, rather, a description of some consequences of being an ELL. The possibility of failing on tests, not achieving in school, or having a limited participation in society are ambiguous criteria for defining ELLs that may be interpreted in many ways by different people.

A potential danger of the interpretation of (D) is that, in order for districts to determine English language proficiency, they may use achievement tests standardized on monolingual English speakers as a proxy for language assessment, reasoning that such tests represent the performance criteria students should meet. But these kinds of instruments were not intended for such a purpose or for use with such a population (Linquanti, 2001). Nevertheless, appropriate or not, it is these same instruments that will be used to make decisions about students' academic progress, once they have been designated "fully English proficient" and moved into English-only instruction. This will be the case, even though their performance may look more like that of ELLs than that of English-only peers for some years to come (Katz *et al.*, 2004).⁴ Unjustified judgments about students' abilities may be made on the basis of their language proficiency, and inappropriate decisions about student placement or potential may be made as a consequence (Figueroa and Hernandez, 2000).

Whereas the federal definition above may do a good job of describing the pattern of social conditions of the population of ELLs in the U.S., it does not address the fact, discussed before, that each ELL has a unique set of strengths and weaknesses in L1 and a unique set of strengths and weaknesses in L2. When applied to individuals, the definition may render flawed classifications for testing purposes. Such flawed classifications can be false positive or false negative identification cases—respectively, individuals who are identified as ELLs without actually being ELLs and individuals who are not identified as ELLs in spite of being ELLs.

Section Summary

Official definitions of ELLs provide directions for decision makers and administrators to make placement and testing decisions about students who are or may be ELLs. However, they are not sensitive to important facts about the nature of being bilingual, such as varying proficiencies bilingual individuals may have in listening, speaking, reading, and writing in L1 and the varying proficiencies in those language modes in English. In addition, official definitions of ELLs leave too much room for subjective interpretation. These inaccurate views of the characteristics of ELLs may lead to flawed decisions about who should be tested in which language and when.

Classifying ELLs

According to legislation derived from NCLB (No Child Left Behind Act of 2001, 2002), states are required to assess yearly their ELL students' progress toward developing proficiency in English. Unfortunately, some of the tests used by states and districts have flaws derived from the fact that the process used to develop them has not adequately addressed important issues relevant to ELLs, such as sampling; the nature of their native languages; and their proficiencies in different language modes (see Rabinowitz and Sato, 2006). In this section, we discuss schooling background as an important factor that shapes the proficiency of ELLs in English but is frequently misused as an indicator of specific individuals' proficiency. Next, we examine the challenges of defining and measuring language proficiency. Then we describe a modern view of language proficiency, which, from our perspective, characterizes the range of skills that should be measured by language proficiency testing. Finally, we discuss the limitations of tests of language proficiency that do not reflect this modern view of language.

Schooling Background

Participation in different types of programs (e.g., dual immersion, bilingual, transitional, L2 immersion) is known to have a powerful effect on the ELL student's oral proficiency in L1 and to some degree in L2 (Brisk, 2006; Saunders and O'Brien, 2006; Walqui, 2002). For this reason, students with different program experiences cannot fairly be measured against each other. It is simply unreasonable to expect ELLs who have participated in "very diverse program types [such as] . . . English-only, transitional bilingual, dual language immersion . . . with very different program goals . . . to perform at similar levels" (Huemphner, 2004, p. 379).

Programs that foster continued development of L1 also foster improved literacy in L2 as well as better achievement outcomes (Lindholm-Leary and Borsato, 2006; Riches and Genesee, 2006). Unfortunately, classifications of ELL students for purposes of making decisions about their language proficiency based on their participation in educational programs may be flawed because programs vary considerably in type and fidelity of implementation (Torres-Guzmán *et al.*, and Han, 2005), and because their effectiveness is shaped by a multitude of contextual factors (see Cummins, 1999; Greene, 1998; Krashen, 1996).

Among the many education-related factors affecting students' varying proficiency in L1 and L2 is the number of years students have spent in U.S. schools versus the number of years they have spent in schools in another country (Abella, Urrutia, and Schneyderman, 2005; Rumbaut and Cornelius, 1995); whether the schooling in either country has been continuous or interrupted (Freeman *et al.*, 2001); and the fidelity with

which a given program is implemented during the time a given student participates in it (Collier and Thomas, 2004). Curricula may emphasize different aspects of English language development (see, e.g., Crawford, 2003; Dutro and Moran, 2003; Moschkovich, 2000). Also, the performance of ELLs on academic tasks may vary according to whether they participated in competitive versus non-competitive assessment activities (Rothstein-Fisch *et al.*, 2003).

As important as program design are the social practices that students and teachers engage in as language is learned (Hawkins, 2004). Language-learning is above all a dynamic social process (Nelson, 1996; Ninio and Snow, 1988), thus it is influenced by the ways the social life of the classroom is structured and roles students are expected to take—for example in discussions, in response to questioning, in completing learning tasks (Brisk, 2006; Losey, 1995; Tabor, 1997; Tharp and Gallimore, 1991; Trumbull *et al.*, in press).

Defining and Assessing Language Proficiency

We have discussed varying patterns of language proficiency among ELL students and the need to understand an individual student's profile in order to make a decision about his or her readiness to move into assessment in English. However, it must be acknowledged that the construct of "language proficiency" is, itself, a troubled one. Two of the most fundamental problems are that a) what counts as "proficiency" is extremely variable, dependent upon the context within which one is operating (MacSwan, 2000); and b) proficiency is in a constant state of flux (Pray, 2005). The term has come to mean so many different things that some researchers avoid using it (e.g., Bachman, 1990, cited in Katz *et al.*, 2004). It follows that different language tests are based on different notions of what constitutes language proficiency, some of them on only implicit notions (Abedi, 2004). It is, thus, no surprise that different tests of oral proficiency would yield different classifications of English language learners: A student may score as "proficient" on one test and "not proficient" on another (Katz *et al.*, 2004).

Pray (2005) conducted an empirical assessment of the validity of three well-known English language proficiency tests by investigating how 40 native English-speaking fourth- and fifth-graders would score on them. All of the students were classified as "fluent English speaking" by the Language Assessment Scales—Oral (LAS—O). However, of the same group, only 85% (34) were classified as "fluent English speaking" by the IDEA Proficiency Test. And *none* of the forty participants scored in the "fluent" or "advanced fluent" English range on the Woodcock-Muñoz Language Survey. In fact, fully half of them (20) were classified as "very limited in English" by the test.

MacSwan (2000) examined the Language Assessment Scales (LAS) Español (for grades 2–12) and the Pre-LAS (for grades pre-K–1), perhaps the most widely used assessments of oral language in Spanish, and found them seriously flawed. Both assessments have a section requiring the child to answer questions about a Spanish-language story (weighted 30% and 50% of the total score for the Pre-LAS and LAS respectively). Failure to respond requires the examiner to give the child a score of “O” for the section, which will result in a classification of “non-Spanish speaker,” no matter how well the child does on the rest of the test. MacSwan uses an actual case example of five year-old “Gabriela,” who scored 14/15 on a sentence completion task on the Spanish Pre-LAS, but was nevertheless categorized as a “non-Spanish speaker.” Results from such studies as these do not give us confidence in the instruments districts routinely trust to help them make major decisions about ELLs.

Literacy and Academic Language as Part of Language Proficiency

Originally, the construct of language proficiency encompassed oral language alone; but as researchers concerned with educational outcomes tried to operationalize it relative to schooling, literacy (reading and writing) was added (e.g., Cummins, 1981). Now, the construct, as used by educators and educational researchers routinely includes literacy as well as “academic language,” which has been defined as the advanced oral and written forms and functions of a language required to achieve in school, including the vocabulary, grammar, and discourse conventions associated with the different disciplines (Chamot and O’Malley, 1994; Scarcella, 2003). It is this kind of language that requires years of learning, as distinct from conversational/social language, which may be learned by ELLs within two or three years (Hakuta *et al.*, 2000; Katz *et al.*, 2004).

There are both advantages and disadvantages to this broadening of the scope of the term “language proficiency.” Literacy and oral academic language are essential to successful academic achievement (Riches and Genesee, 2006). However, under current definitions of language proficiency, people who successfully use oral language but do not read and write would be characterized as having limited language proficiency. In addition, as academic language is held out as the pinnacle of linguistic achievement and fundamental to the ability to handle cognitive complexity (Cummins, 2000), the implication is that other language registers students may use are inferior, when they simply serve different purposes in different contexts. It is not difficult to see how current conceptions of language proficiency could lead to a deficit view of many students within our educational system and/or to judging groups of people who are non-literate or speak languages that are not written as intellectually inferior (MacSwan, 2000; MacSwan *et al.*, 2002).

Reading and writing are relatively new human skills, having begun within the last 6,000 years (Fischer, 2001). They are not universally acquired through exposure in the way oral language is but, rather, generally develop only as the result of explicit instruction. They might better be thought of as language-based academic skills rather than as components of language proficiency. In fact, literacy is not strictly linguistic: it depends upon other, non-linguistic, cognitive and perceptual skills, as well as content knowledge. Reading comprehension, itself, is confounded with one’s personal experience; it is well-known that many comprehension questions on tests can be answered correctly with reference to mainstream, middle-class experience without full reading of the text (Popham, 2001). Another apparent consequence of the emphasis on literacy is less attention in the research community to oral language. According to Saunders and O’Brien (2006), there has been only one major study in the past 20 years (Hakuta *et al.*, 2000) that has “explicitly addressed the rates of oral English language proficiency attainment” (p. 23).

Having expressed these concerns, we acknowledge that proficiency with reading, writing, and academic language is a fundamental goal of schooling and that a broad definition of language proficiency (incorporating these elements) is appropriate to the task of understanding the linguistic progress of ELLs. We must point out that commonly-used tests of English language do not assess academic language. A perspective on language proficiency is needed that maintains a focus on the complexities and variations in oral language, while incorporating the academic domains associated with language arts programs in schools.

A Modern View of Language Proficiency: Communicative Competence

We believe the most useful way of thinking about language proficiency is as *communicative competence*, a conceptualization that is grounded in a functional and sociolinguistic perspective (Bachman, 1990; Canale, 1983; Hymes, 1972). Historically, linguists focused on individuals’ mastery of forms and implicit rules—without particular regard for the social purposes and contexts of language use. In recent decades, there has been increasing recognition that language must be understood in context, as it is actually used. This approach has been widely embraced by applied linguists, speech/language professionals, and educational researchers (Cazden, 1999; Cummins, 1980; Gee, 1996; Heath, 1983, 1986). In fact, a communicative competence perspective captures both a focus on proficiency with rules and forms *and* proficiency with language use in context. What counts as competence is relative to the cultural community one is participating in (Romaine, 1995). So, what is considered communicative competence at home may not be considered so at school and vice versa.

Figure 8.2 below shows three common ways of thinking about the components of language proficiency. Teachers tend to think in terms of the elements of the language arts curriculum, and educational standards are framed in those terms, as we discussed earlier. Traditional linguists have characterized language in terms of the following components: phonology (the sound system), morphology (the word-forming system, including word roots, prefixes, and endings), syntax (the implicit rules of sentence formation—what we usually think of as “grammar”), and semantics (meaning conveyed by words and relations among them). Pragmatics, which concerns language use in social contexts is now included in many taxonomies. As mentioned earlier, reading and writing are not part of the traditional linguistic domain.

Communicative competence entails “grammatical competence” (where “grammatical” refers to the forms and rules of all systems—phonological, morphological, etc.), as well as three other components. “Pragmatic competence” refers to knowledge of the implicit rules of language use in social contexts, which could be thought of as specifying who can say what to whom, when, and in what manner (Hymes, 1972). It entails knowing how and when to speak formally or informally, for instance, or which language or dialect to use. Of course, pragmatic rules or expectations vary from one speech community to another as well as according to social setting. “Discourse competence” refers to knowing how to carry on conversations and compose and interpret spoken or written texts such as arguments, explanations, or narratives. Discourse competence is heavily

Language Arts View				
Listening	Speaking	Reading	Writing	
Traditional Linguistic View				
Phonology	Morphology	Syntax	Semantics	Pragmatics
Communicative Competence View				
Grammatical competence	Pragmatic competence	Discourse competence	Strategic competence	
↓				
phonology morphology syntax semantics				

Figure 8.2 Three views of language proficiency. Adapted from Trumbull *et al.*, 2005, p. 13.

entailed in using and processing the academic language of the classroom (Adger *et al.*, 2002).

“Strategic competence” means proficiency in communicating one’s message, making corrections or changes as needed, varying one’s volume or tone, or clarifying what has been said. Strategies such as these are a metacognitive aspect of language use (Bachman, 2002). Strategic skills are relevant not only to social conversation, but also to academic communication. Anticipating the kinds of clarifications needed in a formal discussion or essay are more academic aspects of strategic competence.

The communicative competence approach is the most comprehensive and useful approach for evaluating a student’s language proficiency. To operate successfully in a second language, a learner needs to be proficient in all of the domains specified within the communicative competence framework; and to understand a student’s school functioning via a second language, a teacher needs to be able to evaluate the student’s success in all of these domains (Wong-Fillmore and Snow, 2000). Taking a test, itself, calls upon proficiency in all of these aspects of language: using phonological and morphological knowledge to identify, spell, and derive words; understanding vocabulary and navigating syntax to comprehend instructions and test items; grasping pragmatics of language use (How formal should the writing be? What kinds of questions is one permitted to ask the teacher or another student?); understanding discourse patterns that crop up repeatedly on tests; and trouble-shooting *vis-à-vis* potential miscommunication, by proof-reading, putting oneself in the place of the person judging the answer, and the like.

One acknowledged problem with assessing students from a communicative competence standpoint is the difficulty of operationalizing all of the constructs associated with it and the amount of time it would take to get a full picture of a student’s language proficiency (Saunders and O’Brien, 2006).

Limitations of Mandated English Language Development Tests

Language testing, no matter what form it takes, is limited in that no standardized test or process can ever fully capture a language user’s competence—something that makes it difficult to operationalize language proficiency levels (Katz *et al.*, 2004, p. 14). The dominant theoretical perspective in the field of language assessment, however, is that assessment that provides authentic contexts for eliciting learners’ language use—one that at least integrates form and meaning in as natural a way as possible—holds the best hope of getting an approximation of a language learner’s language proficiency (e.g., Bachman, 2002; Oller, 1992).⁵ This approach mirrors dominant second-language instruction theory that stresses language

use and learning in context (Cummins, 2001; Krashen, 1982, 1991), though it has been tempered with recommendations to include some explicit skills instruction (Dutro and Moran, 2003; Wong-Fillmore and Snow, 2000).

Unfortunately, existing approaches for measuring language proficiency do not appear to reflect recent conceptualizations of the range of skills within the language domain (Kopriva *et al.*, 2004). Moreover, the effectiveness with which tests can measure English development in ELLs may be limited by the fact that no standards are available that specify how language development is supposed to take place in ELLs. As a consequence, the development of many tests of English language is heavily influenced by theoretical frameworks that underlie English language arts standards rather than frameworks that capture more accurately the complexities of the linguistic competence to be developed in a new language. For example, California's "English Language Development" (ELD) standards (California Department of Education, 2002) are framed in the same terms as the state's English Language Arts (ELA) standards. This is probably the reason why the California English Language Development Test (CELDT), which was developed from those standards, has proven to be only moderately predictive of a student's linguistic readiness to handle the content on achievement tests, particularly at higher grade levels (Katz *et al.*, 2004).

The standards proposed by the national professional organization, Teachers of English to Speakers of Other Languages (TESOL) (Gottlieb *et al.*, 2006) are similar to California's in that they focus primarily on students' ability to use language in relation to the academic domains (mathematics, language arts, etc). The TESOL standards are educational/academic standards, related to but not explicitly tied to a full developmental characterization of language proficiency in terms a language specialist might propose. Moreover, all of these standards (ELA, ELD, TESOL) are expressed in terms of broad grade-level clusters (e.g., K-2, 3-5, 6-8), guided generally by research on monolingual English speakers but without a basis in empirical research "documenting the rate at which the average [native English-speaking] student is able to acquire [meet] specified content standards" (Shin, 2004, p. 254).

At the time of this chapter's writing, some states are updating their practices concerning language development testing for ELLs. (see Forte, 2007; National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs, 2006). However, there is still a strong need for a more refined timetable, with many yearly benchmarks of progress toward ELD standards—related not only to grade level but to language proficiency level, as assessed by valid and reliable instruments (Shin, 2004).

Whereas a communicative competence model of language proficiency may be theoretically more defensible, it must be acknowledged that efforts to create language assessments based on a communicative competence model have had mixed success as predictors of which students are likely to have difficulty with achievement tests in L2. It appears that both tests that assess language competence in terms of discrete aspects of language (e.g., syntax, vocabulary) and those that assess communicative competence yield some information about students' readiness for testing in their second language (see citations in Katz *et al.*, 2004). One question that remains is whether a well-developed test of communicative competence might help us understand why it is that many students who do poorly on standardized tests are actually doing quite well in their classrooms (e.g., Katz *et al.*, 2004). Perhaps they are able to negotiate academic language in contexts where they can seek clarifications from teachers and peers (using strategic competence) and use social language skills to scaffold their academic language skills.

Underlying the limitations cited above is the fact that learning English as a second language (ESL) is often confused with learning English as a foreign language (EFL). Of course, there are obvious parallels between these ways of learning English. For example, all newcomers to a language have to learn its systems of pronunciation, grammar, vocabulary, and socially acceptable ways of using the new language. However, research identifies important differences between individuals who are learning English as a foreign language and individuals who are learning English as a second language—and whom some refer to respectively as "circumstantial bilinguals" and "elective bilinguals" (Valdés and Figueroa, 1994, p. 13). The former are doing so voluntarily, not as a social necessity (e.g., González *et al.*, 2001); they are often adults or secondary students taking an elective course. By contrast, ELLs are learning a second, though not foreign language; they are using English as a medium for learning mathematics, science, social studies, and the like while they learn the language; and they are attempting to catch up with a moving target—the continuously developing language proficiency of their native English-speaking peers (Shin, 2004). Moreover, they are still in the process of developing their first language, are thrust into a situation in which they *have* to learn English in order to survive academically and socially, and often do not have access to a strong program that builds both languages and supports students' academic success.

Section Summary

Tests of English language development provide for classification of students according to broad levels of English proficiency. However, they are not refined enough to capture gradations of language development. Moreover, because they are aligned with English language arts standards, they are not

designed to yield a full picture of the communicative competence required for proficiency in a second language. In addition, performance on even these tests—designed to assess the language skills needed in school—may not correlate highly with performance on achievement tests in language arts. Due to these limitations, existing tests of English development may produce invalid measures of English proficiency or fail to measure important aspects of language proficiency that are relevant to learning in an English-dominant school context or to demonstrating knowledge on tests administered in English.

Testing ELLs

As discussed in the previous sections, ELLs even within the same broad linguistic group vary tremendously on their linguistic proficiencies across the four language modes in both L1 and English. This heterogeneity is accentuated by important factors usually not considered in the definition of ELL, such as dialect differences, history of formal education in L1, history of participation in bilingual programs, and fidelity of implementation of those bilingual programs. In addition, recall of experiences is influenced by the language in which an experience was first encoded (Altarriba, 2003; Marian and Neisser, 2000). If the student is asked to use the language associated with the original experience, he or she is more likely to recall the experience. What this means is that in order to get at an ELL student's knowledge, one needs to test in the language in which the original learning experience was encoded.

Because of this heterogeneity, ideally a decision about the language used should be made individually for each student, based on his or her proficiencies in both L1 and English. Moreover, the form in which achievement tests are administered to each ELL student should take into account his or her weaknesses and should take advantage of his or her strengths in each of the four language modes within each language.

We know, however, that such an individualized form of testing for ELLs is impossible to implement because of its high costs. Even if cost were not an issue, this approach would fail because of the fact that accurate information on the proficiencies of ELLs may not be available. As discussed before, even if scores from tests of English development are available, these are suspect because the instruments are limited in their effectiveness to render accurate classifications of ELLs.

Given these circumstances, there is a different set of limitations associated with testing ELLs in either L1 or English. In this section, we discuss the limitations of testing ELLs in English and the limitations of testing ELLs in L1. Then we discuss the fact that ELL performance on tests is unstable across languages.

Limitations of Testing ELLs in English

It would be ideal if ELL students were not tested in English until their English proficiency approached that of their native-English-speaking peers (Katz *et al.*, 2004). Until that point, testing in English is not likely to reveal what ELL students know. They may well be performing at a higher level in the classroom than they can show on a test, which requires them to comprehend language out of context, cope with the language of tests (the "test register"), and read and respond in a timely fashion (Katz *et al.*, 2004). The requirement of the No-Child Left Behind Act of 2001 that an ELL student be tested in English after three years flies in the face of considerable research showing that ELLs require from four to seven years on average to attain full proficiency in English, including the linguistic forms and functions mentioned above (Hakuta, 2001; Ramirez *et al.*, 1991). Thus, this effort to legislate educational accountability is driving practices that threaten the validity of academic achievement measures (Abedi, 2004).

Critical to testing ELLs in English is the use of testing accommodations intended to reduce score variance due to students' limited proficiency in this language. In Chapter 10, Koran and Kopriva provide examples of algorithms and computer software for processing multiple student background variables that can increase the ability of assessment systems to make appropriate accommodation assignments. Unfortunately, these exciting technological innovations are not paralleled by conceptual developments on the nature of the accommodations. To date, there is no theoretical framework that educators, administrators, or even assessment systems can rely on to design defensible testing accommodations for ELLs—a reflection of the fact that testing practices for ELLs have not incorporated current knowledge from the fields of bilingualism and second language acquisition (e.g., see Figueroa and Hernandez, 2000; Hakuta and Beatty, 2000; Solano-Flores and Trumbull, 2003).

Many accommodations used in ELL testing are borrowed from the sets of accommodations used for students with disabilities or are based on misinformed views of language. For example, Ferrara *et al.* (2004) reviewed surveys conducted by different agencies and programs (including the National Center on Education Outcomes and the Council of Chief State School Officers) on ELL testing practices and found that some of the actions reported by some states as accommodations for ELLs included the use of enhanced lighting conditions or larger font size. Such practices expose a deep misunderstanding of the nature of ELLs' needs (Figueroa and Hernandez, 2000) and cannot be expected to yield valid measures of academic achievement.

Whereas not all practices concerning accommodations are deplorable, the accommodations used by major assessment systems in the country fail

to fully address student proficiency in the four language modes. For example, as Table 8.1 shows, of the nine testing accommodations permitted by the National Assessment of Educational Progress (2005), four (see Accommodations 1–4) do not appear to be directly related to language and, therefore, their ability to reduce the effect of language proficiency on the performance of ELL students may be seriously limited.⁶ Although they may enhance the conditions in which ELL students are tested and they may even contribute to reducing the performance gap between ELLs and non-ELLs, they do not operate on factors related to language proficiency; and their impact on student performance is not specific to ELLs. The other five accommodations address language, but only partially. Three of them (Accommodations 5–7) are effective only for ELLs who have well developed literacy in L1. The other two (Accommodations 8–9) are effective only for students with specific combinations of differences in their listening and speaking proficiencies across English and L1. Since information on ELLs'

TABLE 8.1 Testing accommodations for ELLs permitted by NAEP

Accommodation	Related to language	Condition assumed in ELL student
1. Small group	No	—
2. One-on-one (tested individually)	No	—
3. Extended time	No	—
4. Preferential seating	No	—
5. Bilingual version of test (Spanish/English) ^a	Yes	Student is able to read in L1
6. Bilingual word lists or glossaries (Spanish/English) ^b	Yes	Student is able to read in the L1 and is familiar with the academic language in L1
7. Bilingual dictionary without definitions ^c	Yes	Student is able to read in L1
8. Passages, other stimulus materials, or test questions read aloud in English or presented by audiotape ^c	Yes	Student is more proficient listening in English than reading in English
9. Passages, other stimulus material, or test questions translated aloud into native language or presented by audiotape ^d	Yes	Student's comprehension of academic language spoken in L1 is better than student's reading comprehension of academic language in English

Source: NAEP (2005).

^a Only for mathematics and science.

^b Only for science.

^c Except for reading.

^d Only for Spanish/English, bilingual mathematics and science.

reading and writing skills in both English and L1 is not always collected or available, it is not difficult to see that many accommodations for ELLs have only a slim chance of contributing to the production of more valid scores. In other words, blanket measures not tailored to individual differences are likely to mask their effects on certain students.

Limitations of Testing ELLs in L1

While testing ELLs in English is limited in its capacity to produce valid academic achievement scores for ELLs, it cannot be assumed that testing ELLs in their first language is necessarily a better option, even in cases in which that is their preferred language (August and Hakuta, 1997). Students who have been placed in English-only programs will have been learning content-area vocabulary in English and may not have parallel vocabulary knowledge in L1. The standard wisdom is that such students should be tested in English (e.g., Abedi *et al.*, 1998; Kujawa *et al.*, 2001). However, as mentioned, these same students cannot be expected to have vocabularies in English equivalent to their native English-speaking peers (Baker, 2000; García, 1991). Their teachers may overestimate their English proficiency because of their conversational fluency (Kujawa *et al.*, 2001). Moreover, even ELL students who have been classified as English-language proficient are often unable to demonstrate their content-area skills on English-language achievement tests (Linquanti, 2001) and may perform better on a test in their home language (Abella *et al.*, 2005, p. 141).

Even if it is assumed that limited academic language proficiency in L1 is not an issue in the testing of ELLs in that language, there are serious threats to the validity of tests in L1 that result from the process of their development. Usually, tests administered in L1 are translations of tests originally developed in English. Tests in English have been developed through a more or less careful process in which their wording is refined and tried out with students to identify and correct for misinterpretation of the test items. In contrast, tests administered in L1 are usually translated from English under very tight timelines that limit the possibilities for refining the wording of items (Solano-Flores, 2007b; Stansfield, 2003). Moreover, several mistakes commonly made in the selection of translators jeopardize the validity of translated tests. Those mistakes include judging the translation skills of individuals on the basis of their perceived speaking skills in the target language or assuming that individuals are good translators just because they are bilingual or because they are good writers in the target language (see Hambleton and Patsula, 1999). These are very serious issues, if we consider that even when sufficient time and resources are allocated for test translation, test items may end up measuring different constructs (Van de Vijver and Poortinga, 2005).

Procedures for the simultaneous development of tests in two languages have the potential of eliminating the threats to validity that result from translating tests (Rogers *et al.*, 2003; Solano-Flores, 2007a; Solano-Flores *et al.*, 2002). These procedures are based on the notion that meaning cannot be transferred totally across languages (Greenfield, 1997) and the fact that translated items may function differently across the populations tested in the original language and in the target language (Sireci and Allalouf, 2003). This differential functioning can be extremely sensitive to even subtle variations in the precision of translations (Ercikan, 2002).

The simultaneous development approach is based on the notion that an equal treatment of languages throughout the entire process of test development is most likely to preserve meaning across languages. In this approach, tests are developed in two languages concurrently and pilot-tested with an equal number of students from the two target linguistic groups (monolingual and ELLs who are native speakers of a given language); the two language versions go through the same number of review-revision iterations; also, any modification made in one language version as a result of pilot testing has to be reflected in the other language version.

It would not be reasonable to attempt to develop tests simultaneously in English and in multiple languages. However, it is feasible to develop tests simultaneously in English and in Spanish, the native language of the vast majority of ELLs in the U.S. The simultaneous development approach does not appear to increase test development costs (see Solano-Flores *et al.*, 2002). However, it requires from assessment systems the willingness to modify their test development practices in a way that the minority language is included in the test development process from the very first draft versions of a test, rather than as an afterthought, when few significant improvements can be made.

In the absence of proper mechanisms for examining ELL students' schooling background and their exposure to academic language in L1 and without rigorous procedures for test translation, testing ELLs in L1 may not produce valid measures of academic achievement.

Major Challenges in the Selection of Language for Testing ELLs

It appears from the discussion above that no perfect solution exists when it comes to deciding what language should be used to test ELLs. Even if student classification and implementation were not a problem in the testing of ELLs, an issue yet to be properly addressed is a fact well known among linguists—that bilingual individuals do not function as two monolinguals in one (Grosjean, 1989). Since bilingual individuals tend to have different sets of strengths and weaknesses in each language, their ability to address the cognitive challenges posed by an item is not replicated across languages

(Genesee, 1994; Hakuta *et al.*, 1987; Stevens *et al.*, 2000). As a consequence, there is something about an ELL's academic achievement that is not captured by testing him or her in L1 but also something about the same student's academic achievement that is not captured by testing him or her in English.

This limitation of testing ELLs in one language is expressed as a considerable amount of error when ELLs are given the same set of items in English and in L1. By using generalizability (G) theory (Cronbach *et al.*, 1972; Shavelson and Webb, 1991), a psychometric theory of measurement error, it is possible to estimate score variation due to student—the object of measurement—and several facets (sources of measurement error). In a series of investigations performed with native speakers of different languages, we have estimated the score variation due to the main and interaction effect of student, rater, item, and the language of testing. Table 8.2 presents results for native Haitian-Creole speakers in the same school district (Solano-Flores and Li, 2006), who were given the same set of grade 4 mathematics items sampled from the NAEP 1996 public release (National Assessment of Educational Progress, 1997) in English and in their first language. The table shows the estimates of the amount of variance (and the percentages of that variance) due to 15 terms that result from the main

TABLE 8.2 Generalizability study for ELL students native speakers of Haitian-Creole tested in L1 and in English: s x r x i x c random model. Fourth grade, mathematics

Source of Variability	<i>n</i>	Estimated Variance Components	% of Score Variation
student (s)	49	.0299	20
rater (r)	4	.0000 ^a	0
item (i)	10	.0097	6
language (l)	2	.0074	5
sr		.0000 ^a	0
si		.0164	11
sl		.0100	7
ri		.0004	0
rl		.0001	0
il		.0016	1
sri		.0000 ^a	0
srl		.0000 ^a	0
sil		.0589	39
ril		.0000 ^a	0
srile		.0166	11

Source: Adapted from Solano-Flores and Li (2006).

^aSmall negative variance of negligible magnitude set to zero, following Brennan's (1992) approach.

effect and interaction effect of student, rater, item, and language. The last term (σ^2_{ϵ}) is the error term, which cannot be interpreted because the effects of student, rater, item, and language are confounded with error, ϵ , due to other, unknown sources of score variation. The most important source of measurement error in the testing of ELLs is the interaction of student, item and language (σ^2_{ϵ}), which accounts for 39% of the score variation.⁷

These results indicate that, to great extent, a given student may perform well or poorly on a given item, depending on the language in which the item is administered. Also, these results indicate that, in addition to its intrinsic cognitive and content knowledge demands, each item poses a particular set of linguistic demands in English and a particular set of linguistic demands in L1. At the same time, as we have discussed, in addition to his or her academic strengths and weaknesses, each ELL has a particular set of strengths and weaknesses in English and a particular set of strengths and weaknesses in L1.⁸

Section Summary

The validity of scores from both tests administered in L1 and in English is threatened by the fact that the official definitions of ELLs and the tests used to measure English language development may produce erroneous information on the preparedness of students to be taught or tested in one or another language. Effective testing of ELLs in L1 is limited by the fact that it does not take into account, among other things, the formal instructional experiences that these students have had in English. On the other hand, effective testing of ELLs in English is limited by the lack of an appropriate theoretical framework that allows making sound testing accommodation decisions and the lack of mechanisms and resources that ensure fidelity in the implementation of testing accommodations. Moreover, effective testing of ELLs in either L1 or English is limited by the fact that the performance of ELL students varies considerably across test items and the language in which those items are administered. As a consequence, even under conditions in which ELL classification and fidelity of implementation are not a problem, testing ELLs only in L1 or only in English yields only a partial picture of their academic achievement.

Future Directions

Unfortunately, there is neither a simple nor an optimistic answer to the question, "In which language should ELLs be tested?" Asking the question would not be necessary if testing practices were sensitive to the characteristics that are critical to the condition of being bilingual. As a

consequence, sound decisions on the use of one or another language to test ELLs require deep examination of both the quality of the information on the students' linguistic proficiencies and the resources needed for properly implementing testing in each language. Equally unfortunately, the goals of the No Child Left Behind Act—to make districts accountable for the academic success of every student—cannot be met if the quality of assessment information about ELLs is problematic and if schools are not well equipped to properly implement tests in L1 or to provide proper accommodations when they give ELLs tests in English.

Throughout our discussion, it has become evident that, to a great extent, the process of deciding about the language used to test ELLs has to address the uncertainty that results from flawed classifications of ELLs, limited effectiveness of testing accommodations, and limited resources needed to properly implement those few testing accommodations that have been proven to be effective in reducing the effect of language proficiency on test performance. Thus, we conclude this chapter by sharing results from research that examines which language should be used to test ELLs from the perspective of score dependability. Ultimately, this research is intended to allow development of testing models that are resilient enough to respond to the fact that ELL populations are tremendously heterogeneous and that address factors critical to the condition of being bilingual.

Local Specificity of Testing Models

In the context of our studies on the instability of the performance of ELLs across items administered in L1 and in English (see above), we have performed decision (D) studies with the purpose of determining the minimum number of items needed to obtain dependable scores with tests in which all the items are administered in English, all the items are administered in L1, and some items are administered in English and some in L1. Key to examining the effectiveness of those testing models are two types of G coefficients: ρ , a relative-decisions, norm-referenced coefficient that refers to the relative standing of students, and ϕ , an absolute-decisions, criterion-referenced coefficient that refers to the absolute level of student performance (e.g., the number of items correct) (see Shavelson and Webb, 1991).

We (Solano-Flores and Li, 2006; in preparation) have found that the most effective testing models for ELLs—the ones that give the highest G coefficients with the minimum number of items—are those in which students are tested only in one language. However, as Table 8.3 illustrates, we have also observed that even within the same broad linguistic group (i.e., ELLs who are native speakers of the same language), the G coefficients obtained may vary considerably across groups of students from different localities (schools or school districts). For example, we have observed that

TABLE 8.3 Relative and absolute decisions coefficients obtained for ELL Spanish speakers from two geographical regions tested in Spanish. Fourth grade, mathematics

	Site V	Site W
ρ^2	.21	.14
ϕ	.20	.13

Source: Solano-Flores and Li (in preparation).

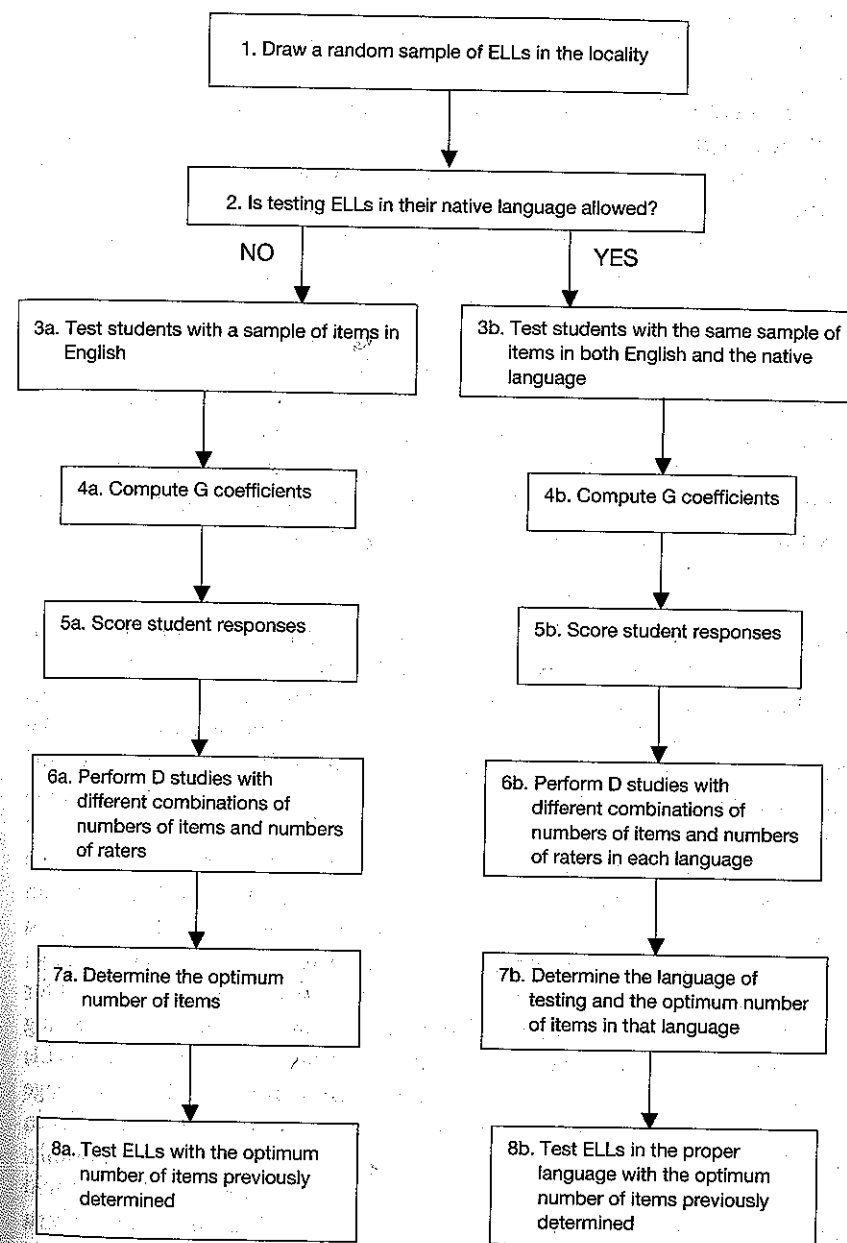
the minimum number of items needed to obtain dependable scores if students are tested in L1 and the minimum number of items needed to obtain dependable scores if students are tested in English may be different for students from communities where the same L1 is spoken.

A consequence derived from this finding is what could be called, the *principle of local specificity of testing models*: the number of items needed to test ELLs in either L1 or English cannot be generalized across broad linguistic groups (all ELLs) or across localities within the same linguistic group. Therefore, in order to ensure score dependability, any decision about the language in which ELLs students should be tested should be made separately for each locality.

Optimum Number of Items as Testing Accommodation

The finding that a considerable amount of score variation in the testing of ELLs is due to the interaction of student, item, and the language of testing supports the notion that, in addition to being samples of the knowledge domain that a test is intended to measure, test items are unintended samples of the language of testing (Solano-Flores, 2006). Thus, a form of testing accommodation can be devised in which the language to be used to test a particular group of ELLs is determined by giving students from that group the same set of items in two languages and then performing a series of D studies to determine the minimum number of items needed to obtain dependable scores when they are tested in L1 and the minimum number of items needed to obtain dependable scores when they are tested in English. With this form of accommodation, different groups of ELLs may be found to need to be tested with different numbers of items in different languages. Also, different number of items may be needed to produce dependable scores for ELL students and non-ELL students.

Figure 8.3 shows the procedure for implementing optimum number of items as a form of testing accommodation for ELLs. In addition to testing ELLs in English, as is now mandatory, the procedure considers the possibility of testing ELLs in L1. As a result of the No Child Left Behind legislation, the national trend is to test ELLs in English after a short

**Figure 8.3** Procedure for using optimum number of items as a form of testing accommodation for ELLs.

time of instruction. However, there may be states or school districts for which testing ELLs in L1—although with some restrictions—may still be an option.

Consistent with the notion that each group of ELLs is unique and each content area poses a specific set of linguistic demands, this procedure should be performed separately for each locality, for each broad linguistic group in the locality, and for each content area.

Some of the terms and directions used in the figure deserve elaboration. First, the term “locality” (see Box 1 in Figure 8.3) refers to either a school or a school district. Whereas, ideally, this form of accommodation should be performed at the school level, that is not always possible because a good number of schools in the country have small enrollments of ELLs.

Second, the sample of items used to test ELLs in the preparatory stage should reflect the types of items (e.g., open-ended, multiple choice) to be used in the operational stage of the test. Therefore, the sample of items should be provided by the assessment system (e.g., they can be assembled with items from a pool of released items from previous test administrations).

Third, when testing ELLs in L1 is an option (Box 2), it is important to keep in mind that translations made by translators in a “standard” version of a language may not be as sensitive to differences in language use across localities and the ways in which curriculum is enacted as are translations made by teachers from the students’ localities (Solano-Flores *et al.*, 2005; Solano-Flores, 2006). Ideally, the translations of test items into the students’ native language should be made by translation teams that include teachers from the same localities.

Fourth, testing students “with the same items in both English and the native language” (Box 3b) means giving each item in the two languages on different occasions, not providing side-by-side formats in which the two language versions of the same items are administered simultaneously (e.g., Sireci and Khaliq, 2002). Appropriate actions must be taken to control for the effect of learning, fatigue, and repetition. For example on each testing occasion, students are given some items in English and some items in L1; on the second occasion, they are given the complementary language versions of the items administered on the first occasion. Also, the order in which the items are sequenced on each occasion is determined randomly.

Fifth, the importance of rater as a source of measurement error should not be underestimated (Boxes 5a and 5b). Even if ELLs are tested only in English (Box 5a), the proficiency of raters in the students’ native language may have an important effect on the accuracy with which they interpret student responses to open-ended items. In our research (see Solano-Flores and Li, 2006), we have found a negligible amount of score variation due to rater. However, we believe that this small effect was due to the fact that the

raters we used were teachers from the same school districts as the students and were native speakers of the students’ native language. In another study in which groups of teachers developed and scored a literacy test for students who were newly transitioned to English-only instruction, one of us documented how knowledge of students’ first language (Spanish) aided interpretations of responses written in English (Beaumont *et al.*, 2002; Trumbull and Koelsch, in press). Unfortunately, reports on the testing of ELLs provide only superficial information on the language proficiency of the raters. To what extent the linguistic skills of the raters in the ELL students’ native language influence the quality of their scoring of student responses when they are assessed *in English* is yet to be investigated.

In conclusion, a decision about which language to use in testing ELLs depends upon the specific group of students in question. The decision-making process is fraught with difficulties, not least of which are those posed by the limitations in existing language assessments, lack of language professionals who can evaluate students’ language, and the dearth of research-based, theoretically defensible approaches to matching testing accommodations to student needs.

In an ideal scenario, such a decision would be made for *each* student, with consideration given to many factors. Among these are the student’s particular patterns of proficiency in L1 and L2, his or her educational background, and the cultural group(s) in which he or she has been socialized.

We have shown how one alternative approach to assessment, the “optimum number of items” method, has proven to be a feasible way to obtain better information about ELL student learning and achievement. Research on this method points to the complexity of interactions among student, language, and test item. It also highlights the importance of local involvement in designing and developing assessment systems. Even in the absence of perfect information on every student, teachers and administrators have knowledge of their student populations that can guide decisions about assessment practices within a school or district. And, even in the absence of perfect language tests based on a communicative competence view, an understanding of students’ language development can be constructed through the use of existing tests, teacher observations, and communication with parents about their children and how language is used in the home. If the education community is truly committed to maximizing the validity, utility, and equity of test scores, it must seriously entertain such alternative methods.

The issues we have discussed in this chapter pertain not only to decisions about choice of language for student testing but also to the assessment of ELLs in general. Districts may not be able to craft assessment systems that are perfectly aligned to the dual needs of their student populations and public accountability. Nevertheless, with a deep understanding of the

limitations of current assessment practices for ELLs and an awareness of the kinds of possibilities that hold promise for improved assessment, they are likely to develop better visions of what is feasible within their own contexts.

Endnotes

1. Here, we use the components of language proficiency used by educators. "Language proficiency," a disputed construct, has traditionally been based on evaluation of a person's knowledge and use of oral language.
2. In reality, each individual has his or her own linguistic repertoire, based on individual life experience, so even the term "dialect" obscures many individual differences. However, it is often associated with participation in a particular social and cultural community. A dialect is a variety of a language that is distinguished from other varieties of the same language by features of pronunciation, grammar, vocabulary, and use (Crystal, 1997; Wolfram *et al.*, 1999). Cultural groups vary in their beliefs about how language ought to be used by children versus adults, what is appropriate in different settings, and for what purposes language should be used (Bialystok, 2001; Greenfield *et al.*, 2000; Heath, 1983, 1986). Although the term "dialect" is sometimes taken to mean a sub-standard version of a language, in fact everyone speaks a dialect, and all dialects follow their own consistent rules of grammar. The most socially prestigious dialect is usually labeled "standard," as in "Standard English" (Preston, 1998).
3. The term "academic language" is now used to refer to the advanced oral and written forms and functions of a language required to achieve in school including the vocabulary, grammar, and discourse conventions associated with the different disciplines (Chamot and O'Malley, 1994; Scarcella, 2003).
4. According to Katz *et al.* (2004), "proficiency" has been historically associated with the "proficiency movement" of the American Council on the Teaching of Foreign Languages and has particular associations. Nevertheless, it is the term most commonly used in legislation and education literature.
5. It should be pointed out that tests in which particular language skills are examined out of context also contribute to an understanding of a student's language proficiency (see, e.g., Katz *et al.*, 2004), despite their limitations in capturing language in use.
6. Figueroa and Hernandez (2000) argue on the basis of several studies that processing time in L2 is longer and that the difference gets magnified as a task becomes cognitively harder. In fact, a recent experimental study with Spanish-English bilingual adults showed that they took significantly longer to retrieve common words in their non-dominant language than in their dominant language (Langdon *et al.*, 2005). So extending the time for testing would, in that case, address a language need, particularly for more difficult test items. Extra time has also been shown to benefit monolingual students when scores of large groups of ELLs and monolinguals are compared. It could be argued that extra time for ELLs alone would be more equitable, to compensate for the greater ease of linguistic processing by monolinguals.
7. Notice that the score variation due to the main and interaction effect of rater is minimum. In our studies, we were very careful in selecting raters who were native speakers of the ELL students' L1 and who had experience teaching students from the same localities. A larger effect of rater might have been observed had we had included raters with other language backgrounds (e.g., bilingual teachers whose L1 was English).
8. August and Hakuta (1997) discuss the results of an investigation with translated tests, in which Anderson (cited by the same authors) observed that ELLs performed better on some items administered in L1 and better on some items administered in L2. August and Hakuta suggest that this performance inconsistency across languages was attributable to the fact that item difficulty was not controlled for. In our investigation, differences due to different difficulties across languages would be reflected as a large main effect of language and large interaction effect of the facets, item and language. However, these effects were small and a considerable effect due to the interaction of student, item, and language was observed—which supports our conclusions.