Language, Dialect, and Register: Sociolinguistics and the Estimation of Measurement Error in the Testing of English Language Learners

GUILLERMO SOLANO-FLORES
University of Colorado at Boulder

This article examines the intersection of psychometrics and sociolinguists in the testing of English language learners (ELLs); it discusses language, dialect, and register as sources of measurement error. Research findings show that the dialect of the language in which students are tested (e.g., local or standard English) is as important as language as a facet that influences score dependability in ELL testing. The development, localization, review, and sampling of items are examined as aspects of the process of test construction critical to properly attaining linguistic alignment: the correspondence between the features of the dialect and the register used in a test, and the features of the language to which ELLs are exposed in both formal and instructional contexts.

Though well recognized, the impact of language on the validity of tests is yet to be properly addressed. Since testing typically depends on the use of language, to a large extent an achievement test is a test of language proficiency (American Educational Research Association/American Psychological Association/National Council on Measurement in Education, 1999). Language remains the prime construct-irrelevant factor in testing—a factor that an instrument does not intend to measure yet affects test scores (see Messick, 1989).

Although language is always an issue in testing, it becomes a much more serious problem when students are not proficient in the language in which they are tested. Efforts in the field of testing accommodations for English language learners (ELLs) have rendered results that speak to the difficulty of addressing this challenge. The effectiveness of the linguistic simplification of items is limited by factors such as the ELL students' language backgrounds (e.g., Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000). Moreover, language interacts with mathematics achievement in tests in

Teachers College Record Volume 108, Number 11, November 2006, pp. 2354–2379

0161-4681

ways that are different for ELL students and their non-ELL counterparts (Abedi, 2002).

The issue of language as a construct-irrelevant factor in ELL testing is aggravated by inappropriate or inconsistent testing practices and policies. Information on the linguistic proficiency of ELLs is usually fragmented or inaccurate (De Avila, 1990), and the criteria and instruments used to classify students as ELLs are not the same across states (Abedi, 2004; Aguirre-Muñoz & Baker, 1997). Even attempts to characterize the linguistic proficiency of ELLs based on the kind of bilingual programs in which they are enrolled (or whether they are in any bilingual program at all) may be flawed because these programs vary considerably in type and fidelity of implementation (Brisk, 1998; Gandara, 1997; Krashen, 1996), and their success is shaped by a multitude of contextual factors (Cummins, 1999).

Several authors (e.g., LaCelle-Peterson & Rivera, 1994; O. Lee, 1999, 2002, 2005; Lee & Fradd, 1998; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003) have asserted that existing approaches to dealing with diversity are limited because they lack adequate support from current theories of language and culture. This gap between disciplines is well illustrated by results from a recent review of surveys of ELL testing practices (Ferrara, Macmillan, & Nathan, 2004). This study revealed that among the accommodations reported for ELLs are actions of dubious relevance to language—such as providing enhanced lighting conditions—borrowed from the set of accommodations created for students with disabilities (see Abedi, Hofstetter, & Lord, 2004). Although these irrelevant accommodations are well intended and may contribute to enhancing testing conditions for any student, they do not target characteristics that are critical to the condition of being an ELL, and they ultimately lead to obtaining invalid measures of academic performance for ELLs.

Although linguists have seriously questioned current ELL testing practices (e.g., Cummins, 2000; Hakuta & Beatty, 2000; Hakuta & McLaughlin, 1996; Valdés & Figueroa, 1994), this criticism has not brought with it alternative approaches. Unfortunately, this dearth of alternative approaches becomes more serious in the context of the No Child Left Behind Act (2001), which mandates that ELLs be tested in English after a year of living in the United States or of being enrolled in a program for ELLs. Unfortunately, ELLs will continue to be tested for accountability purposes in spite of both the flaws of the new accountability system (see Abedi, 2004) and the body of evidence from the field of linguistics that shows that individuals need more time to acquire a second language before they can be assumed to be fully proficient in that language (Hakuta, Goto Butler, & Witt, 2000).

This article addresses the need for research in the field of language from which new and improved methods for the testing of ELLs can be derived (see August & Hakuta, 1997). It addresses the fact that tests, as cultural

artifacts, cannot be culture free (Cole, 1999) and that constructs measured by tests cannot be thought of as universal and are inevitably affected by linguistic factors (see Greenfield, 1997). It establishes the intersection of two disciplines: (1) sociolinguistics, which is concerned with the sociocultural and psychological aspects of language, including those involved in the acquisition and use of a second language (see Preston, 1999) and (2) psychometrics, which in the context of education is concerned with the design and administration of tests and the interpretation of test scores with the intent of measuring knowledge and skills.

This article is organized in two parts. In the first part, I discuss the link between two key concepts in sociolinguistics: dialect and register; and two key concepts in psychometrics: sampling and measurement error. These concepts are critical to the development of new, alternative psychometric approaches that address the tremendous heterogeneity that is typical of populations of ELLs.

In the second part, I discuss the notion of linguistic alignment: the correspondence between the dialect and the register used in a test and the characteristics of the language to which ELLs are exposed. I then discuss ways in which linguistic alignment can be addressed in different areas of the testing process.

LEVELS OF ANALYSIS IN THE TESTING OF ELLs

Current approaches to testing ELLs are mainly based on classifications of students according to broad linguistic groups, such as students whose first language is English, or students whose first language is Spanish. This view is reflected in the designs used traditionally in ELL research. These designs focus on test score differences between populations of ELLs and mainstream non-ELLs, or on test score differences between subgroups within a given population defined by some kind of treatment. For example, in the field of testing accommodations for ELLs, an "ideal study . . . is a 2×2 experimental design with both English language learners and native speakers of English being randomly assigned to both accommodated and nonaccommodated conditions" (Shepard, Taylor, & Betebenner, 1998, p. 11).

In some cases, the classifications used in these studies may be inaccurate because of the wide variety of types of ELLs or bilingual students (see Aguirre-Muñoz & Baker, 1997; Casanova & Arias, 1993; Council of Chief State School Officers, 1992). In addition, these classifications do not always refer to the students' academic language proficiencies in either English or their native language (see Cummins, 2000).

An additional level of analysis can be used that comprises two additional closely related components: dialect and register (Figure 1). *Level* refers to

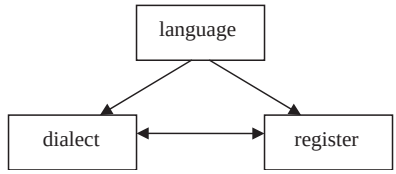

Figure 1. Levels of Analysis in the Testing of ELLs

the fact that dialect and register are considered to be subordinate categories of a language in the sense that there may be many dialects of the same language and many registers within the same language (see Wardhaugh, 2002).

Whereas *dialect* refers to a variation of a language that is characteristic of the users of that language, *register* refers to a variation of a language that is determined by use—a situation or context. Dialects are different ways of saying the same thing; they reflect social structure (e.g., class, gender, and origin). Registers are ways of saying different things; they reflect social processes (e.g., division of labor, specialty, contexts, content areas, and specific activities; Halliday, 1978). Dialects are associated with the linguistic and cultural characteristics of the students who belong to the same broad linguistic group; registers are associated with the characteristics of the language (especially academic language) used in tests.

This section discusses how the sociolinguistic perspective and the psychometric perspective can be used in combination to examine language, dialect, and register as sources of measurement error.

LANGUAGE AND MEASUREMENT ERROR

The idea of linking psychometrics to sociolinguistics originated in a project whose main goal was to assemble a sample of responses given by ELL students, whose first languages were Spanish, Chinese, and Haitian Creole, to the same set of open-ended science and mathematics items administered in English and in their native language (Solano-Flores, Lara, Sexton, & Navarrete, 2001). Our intent was to show, side by side, the students' responses to each item in both languages.

In selecting the response samples, we observed that the quality of the students' responses was inconsistent across both items and languages. A given student might perform better in his first language than in English for some items but better in English than in his first language for other items. If we wanted to determine whether these students should be tested in English or in their first language, comparing the mean scores they obtained in each language would not render valuable information because the score differences would cancel each other. What we needed was an estimation of the

amount of score variation due to this interaction between student, item, and language.

To accomplish our goal, we used generalizability (G) theory (Brennan, 1992, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). G theory is a psychometric theory that deals with measurement error. It distinguishes between two kinds of sources of score variation. One is student (s), the object of measurement; the other comprises facets—sources of measurement error. The facets in our study were item (i), rater (r), and language (l). G theory allowed us to estimate score variance due to (1) the main effect of student (s); (2) the main effect of the facets (i, r, l); and (3) the interaction effect of all sources of score variation (si, sr, sl, ir, il, rl, sir, sil, srl, and srl,e; the *e* in "srl,e" denotes the measurement error that cannot be accounted for and that is due to unknown sources).

Our analyses revealed that the sil interaction produced the largest score variation. The performance of ELLs was considerably inconsistent both across items and across languages. These results indicated that, in addition to their knowledge of the content area assessed, ELLs had different sets of strengths and weaknesses in English and in their native language, and in addition to their intrinsic cognitive demands, test items posed different sets of linguistic challenges.

A series of studies performed with larger samples of students confirmed those results (Solano-Flores & Li, 2006). In this new series of studies, we examined the ρ^2 and φ coefficients, which respectively express the extent to which student achievement measures can be generalized depending on whether they are intended to rank individuals or to index their absolute level of performance (Shavelson & Webb, 1991). Based on these coefficients, we were able to determine the number of items that would be needed to obtain dependable scores by testing ELLs in English and in their native language.

We observed that the number of items needed to obtain dependable scores varied within the same broad linguistic group. We also observed that testing ELLs in their native languages does not necessarily produce more dependable scores than testing them in English. For example, in order to obtain more dependable measures of academic achievement, some groups of native Haitian Creole speakers might need to be tested in English, while others might need to be tested in Haitian Creole. A similar pattern was observed among native speakers of Spanish.

The considerable score variation due to the interaction of student, item, and language is consistent with two well-known facts about bilinguals. First, ELLs have different patterns of language dominance that result from different kinds of language development in English and their native languages (see Stevens, Butler, & Castellon-Wellington, 2000). Second, ELLs' linguistic proficiencies vary tremendously across language modes (i.e., writing,

Table 1. Differences Between Approaches to ELL Testing Based on Item Response Theory (IRT) and Approaches Based on Generalizability (G) Theory

	Item Response Theory	Generalizability Theory
Focus	Scaling, score differences between linguistic groups	Measurement error, score variation due to language
Information produced	Differential item functioning	Test score dependability
Designs	Between groups	Within groups
Reference groups	Non-ELLs or ELLs who do not receive testing accommodations	No reference groups
Level of analyses	Item	Test
Characteristics of linguistic groups	Clear-cut differences assumed	No clear-cut differences assumed

reading, listening, speaking) and contexts (e.g., at home, at school, with friends, with relatives); they are shaped by schooling (e.g., bilingual or full immersion programs) and the way in which language instruction is implemented (e.g., by emphasizing reading or writing in one language or the other; Bialystok, 2001; Genesee, 1994; Valdés & Figueroa, 1994).

An approach based on viewing language as a source of measurement error addresses the fact that bilingual individuals do not typically replicate their capacities across languages (Bialystok, 1991; Heubert & Hauser, 1999). This approach differs substantially from other approaches to ELL testing. For example, approaches based on item response theory (IRT) allow interpretation of mean score differences between linguistic groups (e.g., Camilli & Shepard, 1994; Ercikan, 1998; van de Vijver & Tanzer, 1998); they examine bias due to language based on differential item functioning (DIF): the extent to which individuals from different populations (e.g., ELLs and non-ELLs) have different probabilities of responding correctly to an item despite having comparable levels of performance on the underlying measured attribute (Hambleton, Swaminathan, & Rogers, 1991). In contrast, an approach based on G theory does not necessarily compare groups (Table 1).

A detailed discussion of the characteristics of G theory and IRT cannot be provided in this article for reasons of space. However, it should be said that, rather than being alternate approaches to the same measurement problems, the two theories serve different sets of purposes and can be used complementarily. An approach to ELL testing that integrates the two theories may not take place soon because efforts to address some methodological and theoretical issues to link them are still in progress (e.g., Briggs & Wilson, 2004). In addition, although it has been used to examine error variance due to facets such as rater, task, occasion, and context in the testing of second-language proficiency (Bachman, Lynch, & Mason, 1995; Bolus,

Hinofotis, & Bailey, 1982; Brown & Bailey, 1984; Y. W. Lee, 2005; Molloy & Shimura, 2005; Stansfield & Kenyon, 1992), G theory has not been used before to examine language as a source of measurement error.

DIALECT AND MEASUREMENT ERROR

A dialect is defined by linguists as a variety of a language that is distinguished from other varieties of the same language by its pronunciation, grammar, vocabulary, discourse conventions, and other linguistic features. Dialects are rule-governed systems, with systematic deviations from other dialects of the same language (Crystal, 1997). Research on the linguistic characteristics of several non-standard-English dialects has found that these dialects are "as complex and as regularly patterned as other varieties of English, which are considered more standard" (Farr & Ball, 1999, p. 206). Thus, although the term *dialect* is frequently used to refer to the language used by people from a particular geographic or social group or to mean a substandard variety of a language, in fact everyone speaks dialects (Preston, 1993). Standard English is one among many English dialects (Wardhaugh, 2002).

Different dialects may originate from contact with other languages or from the fact that certain features of a language shared by its speakers evolve among some communities but are kept among others (Wolfram, Adger, & Christian, 1999). Thus, ELLs from the same broad linguistic group but from different geographical areas within the United States can be thought of as speakers of different dialects of their own language and speakers of different English dialects.

In thinking about dialect and testing, it must be kept in mind that distinguishing between two dialects of a given language or defining a language or a dialect may be problematic and even a matter of opinion. For example, the differences between Mandarin and Cantonese are more profound than the differences between Danish and Norwegian, yet Mandarin and Cantonese are considered dialects of Chinese, whereas Danish and Norwegian are treated as different languages (see Haugen, 1966). However, although dialects may be difficult to characterize, what is relevant to our discussion is the notion that dialect can be an important source of measurement error. We adopt a pragmatic approach in our research. Rather than trying to characterize the dialect of each community, we assume that communities of users of the same language may have dialect differences important enough to affect student performance on tests.

For the purpose of this discussion, the term *community* is used here in an ample manner to refer to a group of users of the same dialect. This group can be, for example, a school community or a group of individuals who speak the same language and live in the same neighborhood. It is not

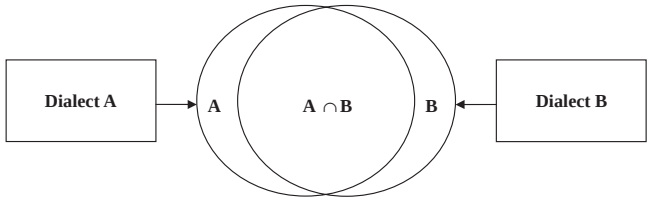

Figure 2. Commonalities and Differences Between Dialects

necessarily used as synonym of *speech community*, a concept that is somehow controversial because it assumes constancy across social contexts in the way in which individuals use a speech style (see Hymes, 1974; Wardhaugh, 2002).

Although the dialects spoken by different communities are mutually intelligible (Rickford & Rickford, 1995)—they tend to differ in phonetics and phonology but not in semantics (Halliday, 1978)—in the absence of opportunities for clarification, body language, and certain physical clues, tests limit the possibilities for understanding test items. Moreover, because young ELLs are developing both their first and second languages or because their own native language is poorly developed, their performance in tests can be especially sensitive to dialect variations, regardless of the language in which they are tested. Subtle but important differences in elements such as word usage, word frequency, syntax, and the use of certain idiomatic expressions may limit the capacity of standard dialect versions of tests to properly assess ELL students in either English or their native language (Solano-Flores, Trumbull, & Kwon, 2003).

Figure 2 illustrates the commonalities and differences that may exist between dialects of the same language. For the sake of simplicity, this example assumes that there are only two dialects, A and B, of the same language. The circles represent the sets of linguistic features (e.g., grammar, vocabulary, word use frequency, and discourse conventions) that define Dialect A and Dialect B. The intersection of A and B ($A \cap B$) represents all the features that the two dialects have in common; the areas outside the intersection represent all the features that the dialects do not have in common and that might pose a challenge for communication.

In the field of testing, it is common to say that a test is written in the standard form of a language (as in Standard English) to imply that it can be understood by all the users of that language. In our example, this is true only if all the linguistic features of the test are in $A \cap B$. However, the reality might be otherwise. Dialects are associated with various social groups or classes (Coulmas, 2005); *standard* is actually used to refer to the mainstream or most socially acceptable dialect in a society (Wardhaugh, 2002). If Dialect A is the dialect used by the mainstream segment of a society, then a test

written in the standard dialect reflects all the linguistic features of Dialect A but only the linguistic features of Dialect B that are in A ∩ B. As a consequence, Dialect B users are more likely than Dialect A users to face linguistic challenges that are not related to the construct that the test is intended to measure.

Solano-Flores and Li (2006) have performed a series of G studies in which both dialect and language have been examined as sources of measurement error. These studies have provided empirical evidence that dialect can be as important as language in the testing of ELLs.

The participants in these studies were fourth- and fifth-grade ELL students whose first language was Haitian Creole. They were assigned to two treatment groups. In Group 1, students were tested across languages with the same set of National Assessment of Educational Progress (NAEP) mathematics items (drawn from NAEP 1996, 2000) in both Standard English (the original version of the test) and the standard dialect version of their Haitian Creole, created by professional translators. In Group 2, students from two communities of Haitian Creole speakers were tested with the same set of items in two dialect versions of Haitian Creole, standard and local—the dialect of Haitian Creole used in their communities.

To create the local-dialect versions of the test, a series of test translation sessions were facilitated with a team of teachers from each community. These teachers were asked to translate the items into a version of Haitian Creole that reflected the language used in their community and that could be understood by their own students.

G theory analyses revealed a considerable score variation due to the interaction of student, item, and dialect for Group 2. Moreover, the magnitude of score variation due to this interaction was as large as the magnitude of score variation due to the interaction of student, item, and language. These results indicate that ELL students do not necessarily perform better if they are tested in a standard version of their native language than if they are tested in Standard English. In addition, the results indicate that the dialect of the language in which they are tested (whether it is English or the first language) is a powerful influence that shapes student performance. Whether tested in English or in their native language, ELLs are tested in some dialect of that language. Regardless of what language is used to test ELLs, dialect can be crucial to obtaining valid measures of their academic achievement.

REGISTER AND MEASUREMENT ERROR

Linguists distinguish between the linguistic skills used by ELLs in informal conversation and the linguistic skills inherent to learning content (Cummins, 1996; Hamayan & Perlman, 1990). Although there has been debate

around the nature of this distinction (see, for example, Cummins, 2000; Edelsky, 1990; MacSwan & Rolstad, 2003; Rivera, 1984), there is agreement that school and testing pose more linguistic demands to ELLs than the demands posed by using a second language in informal settings.

The concept of register as a variety of a language is particularly useful in conceptualizing the challenges that ELLs face in testing:

> A *register* [italics added] can be defined as the configuration of semantic resources that the member of a culture typically associates with a situation type. It is the meaning potential that is accessible in a given social context. Both the situation and the register associated with it can be described to varying degrees of specificity; but the existence of registers is a fact of everyday experience—speakers have no difficulty in recognizing the semantic options and combinations of options that are "at risk" under particular environmental conditions. Since these options are realized in the form of grammar and vocabulary, the register is recognizable as a particular selection of words and structures. But it is defined in terms of meanings; it is not an aggregate of conventional forms of expression superposed on some underlying content by "social factors" of one kind or another. It is the selection of meanings that constitutes the variety to which a text belongs. (Halliday, 1978, p. 111)

Thus, performing well on a standardized test requires ELLs to know more than the content area assessed by the test or the language in which it is administered. It also requires from them the use of the register of that discipline and the register of tests. This combined register is defined by the activity in which individuals are engaged at a given time (e.g., taking a test). Among other things, this register differs from other registers on features such as semantics (e.g., *root* has different meanings in colloquial language and in mathematics); word frequency (e.g., *ion* is mostly restricted to the content of science); idiomatic expressions (e.g., the phrase *None of the above* is used almost exclusively in multiple-choice tests); notation (e.g., *A divided by B* is represented as *A/B*); conventions (e.g., uppercase letters are used to denote variables); syntactical structures (e.g., the structure of multiple-choice items in which an incomplete sentence [the stem] is followed by several phrases [the options]); and ways of building arguments (e.g., *Let A be an integer number*).

No wonder that, although ELLs can become reasonably fluent in conversational English in a relatively short time, it takes much longer for them to become fluent in academic English (Cummins, 1984, 2003; Guerrero, 1997; Hakuta et al. 2000). In addition, it takes much longer for them to deal successfully with the fact that test items tend to contain dense text and scant contextual information, use colloquial terms with unusual meanings

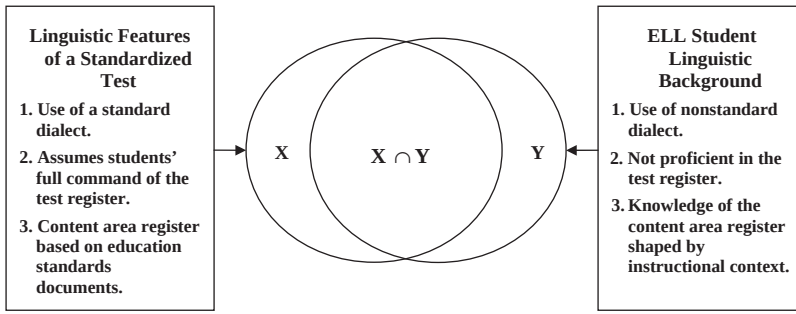

Figure 3. Commonalities and Differences Between the Linguistic Features of Standardized Tests and the Linguistic Backgrounds of ELLs

(Ferguson & Fairburn, 1985; Freedle, 2003), provide unintended clues that may lead students to misinterpret them and to use incorrect problem-solving strategies (Solano-Flores & Trumbull, 2003), and use unnecessarily complex syntactical structures (Solano-Flores et al., 2003). The fact that the linguistic simplification of items used in standardized tests (administered in English) results in improved performance for both ELL and non-ELL students (Abedi & Lord, 2001; Abedi et al., 2000) speaks to the extent to which conventional test writing style tends to be difficult even for non-ELLs.

In Figure 3, the intersection $X \cap Y$ represents the commonalities between the linguistic features of standardized tests and the linguistic backgrounds of ELL students. The area of X that is outside $X \cap Y$ represents the linguistic features of the test that may be challenging to ELLs because of differences in their backgrounds, lack of familiarity with the dialect, or the fact that the register used in the test does not match the students' instructional experiences (see Butler & Stevens, 1997; Hofstetter, 2003). For example, the terms, expressions, and notation conventions used in standardized tests may be based on national or state education standards and may not be exactly the same terms, expressions, and notation conventions used in the curriculum, the textbooks, and the lessons implemented in specific classroom contexts.

Experience from facilitating test translation and test translation review sessions with teachers confirms the notion that register and dialect cannot be examined separately because they are closely interconnected. First, registers tend to be associated with a dialect (Halliday, 1978), and the register used in tests is associated with Standard English. Second, even within the register of a discipline, there may be subtle but important variations that affect student performance in tests. For example, because of their various personal schooling experiences, students from certain Latin American countries may have learned to use commas instead of points to write fraction numbers—which may affect how they interpret an item.

We facilitated translation sessions with teachers who taught native Haitian Creole speakers and teachers who taught native Spanish speakers. Although we had originally planned to focus on dialect, it soon became apparent that the teachers' discussions also addressed register. For instance, teachers discussed at length not only how their translations should reflect the variety of Haitian Creole used in their communities but also whether these translations should keep the notation and styles used in the original English version of the items (e.g., *5 cents* as opposed to *5¢*).

Figure 4 shows an item in its original Standard English version, and the six translations used in one of our studies. The figure shows language (English, Haitian Creole, and Spanish) and dialect (e.g., Standard Spanish, Spanish-City W, and Spanish-City V) as different levels of analysis. Bolding is used to highlight a common register issue that arose during the translation sessions. Teachers from the four translation teams expressed concern that the translation of *rounded to the nearest 10 feet* might not make much sense to students in the target languages unless a more formal expression was used in the translation *a awondi nan dizèn ki pi pre a* and *redondeado a la decena más próxima* [*rounded to the nearest tenth*] in Haitian Creole and Spanish, respectively. At the same time, they were concerned that this more formal expression might increase the difficulty of the item by requiring students to be familiar with the term *tenth*. One of the teams of teachers for each target language decided to use the colloquial translation, whereas the other opted for the formal version.

This tension between construct comparability and dialect and register has also been observed in translations of items used in international comparisons (e.g., Solano-Flores & Backhoff, 2003; Solano-Flores, Contreras-Niño, & Backhoff, 2005). It confirms the notion that even a correct translation that intends to address construct comparability may not necessarily be able to capture entirely the thinking associated with language and culture (Greenfield, 1997; van de Vijver & Poortinga, 1997). Research is needed to devise approaches to effectively deal with the tension between construct comparability and the features of dialect and register.

LINGUISTIC ALIGNMENT

In the assessment literature, it is recognized that instructionally valid assessments might be expected to show "somewhat lower effects of student background characteristics on achievement than tests that are not explicitly related to the curriculum" (Yoo & Resnick, 1998, p. 2). Unfortunately, this literature has not paid sufficient attention to the linguistic correspondence of tests to instructional experiences.

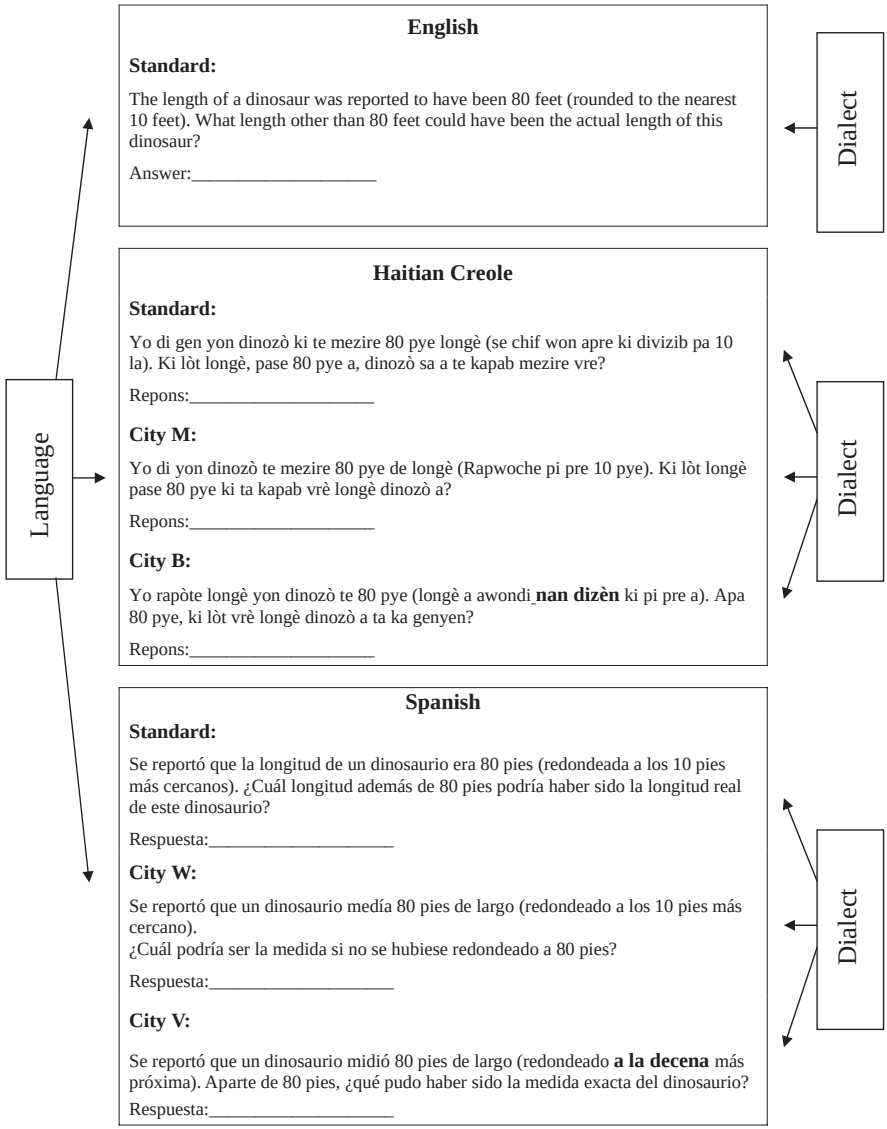

Figure 4. Original, Standard English Version and Six Translations of a Mathematics Item

Linguistic alignment can be defined as the correspondence between the features of the dialect and the register used in a test, and the features of the language to which ELL students are exposed through both informal and

formal school experiences. The counterpart of linguistic alignment is *linguistic misalignment*. Instances of misalignment (e.g., an idiomatic expression unfamiliar to the student, a word of low frequency in the student's dialect, a slight variation in notation) are shown in Figure 3 as the area of X that is outside the intersection $X \cap Y$.

The impact of linguistic misalignment on student performance is represented in Figure 5 as a probabilistic space defined by two dimensions, frequency and criticality. The former refers to the number of instances of misalignment, and the latter refers to their importance. The light and shaded areas represent, respectively, cases in which an item is likely to be linguistically sound (fair) or linguistically challenging (unfair) to an ELL student. Although, within certain limits, individuals are capable of handling misalignment—they can construct meaning from incomplete information or from information that is somewhat unfamiliar to them—too many mild, or few but severe (or both mild and severe), instances of linguistic misalignment make an item likely to affect student performance.

This section discusses how testing practices can be improved to address linguistic misalignment. The discussion addresses four areas in testing. Item writing is concerned with promoting linguistic alignment at the beginning of the process of test development; item localization is concerned with promoting linguistic alignment when tests have been already developed but before they are used with ELLs; item review is concerned with correcting for linguistic misalignment when tests have been already developed but before they are used with ELLs; and item sampling is concerned with controlling for the fact that linguistic alignment cannot be entirely eliminated from tests. Some of the ideas discussed come from experience gained in testing populations in different languages.

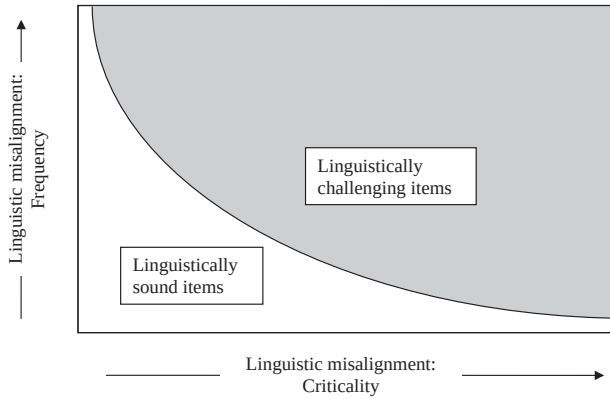

Figure 5. Linguistic Misalignment Represented as a Probabilistic Space

ITEM WRITING

Literature on cross-cultural testing shows the possibilities of alternative approaches to test development. It has been suggested that tests be developed simultaneously in different languages as an alternative to the approach of simply translating or adapting a test from one culture to another (see Tanzer, 2005). Indeed, Solano-Flores, Trumbull, and Nelson-Barber (2002) have evidence that developing two language versions of the same test concurrently makes test developers reason more deeply about how, and address the ways in which culture and context are inextricably related to language.

We facilitated test development sessions in which bilingual (English-Spanish) teachers who taught ELLs (native Spanish speakers) developed mathematics items in both English and Spanish. Teachers worked in two teams, each responsible for the development of one of the language versions of the test. In our procedure, any modification on the wording of the items was negotiated across languages; no change of the wording in an item was made on one language version without reaching a consensus across teams on how that change should be reflected in the other language version of the item. Shells (blueprints or templates) were used to mediate the teachers' discussions. These shells were "hollow," generic descriptions of the characteristics of the items to be developed, which were constructed according to the target knowledge and skills.

We observed that, as the test development sessions progressed, teachers became more sophisticated in their reasoning about language. During the first sessions, they were concerned mainly about correct formal equivalence across languages. In contrast, during the final sessions, their thinking also focused on the alignment of their translations with the dialect used by their students in their own communities.

The experience gained from developing tests concurrently allows us to devise alternate assessment systems that entrust local communities (e.g., schools or school districts) with the task of generating their own test items according to a set of content and format specifications. In these alternate assessment systems, students are tested with items generated by teachers in their own communities rather than items developed by external agencies. However, the items are generated according to shells, which may be generated by those external agencies and tap specific types of knowledge and skills. Although the items generated by different communities differ on wording, they are comparable in the sense that they all meet the specifications of the shells.

At first glance, testing students with items whose wording varies across communities challenges the notion of standardization in testing. But the idea is far from being new, and in fact is seen as natural in cross-language

testing (e.g., Allalouf, 2003; Cook & Schimitt-Cascallar, 2005; Sireci, 2005). There is no reason that testing across dialects should be less acceptable than testing across languages.

ITEM LOCALIZATION

Item procedures can be devised in which the wording of items is adapted by each community to ensure the linguistic alignment of tests with the enacted curriculum.

Tests can be classified as immediate, close, distal, and remote, according to the extent to which their content (topics), goals (what students should be able to demonstrate), and tasks (activities) resemble the content, goals, and tasks implemented in the curriculum (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). It is reasonable to assume that the linguistic misalignment of tests is likely to increase as their distance from the enacted curriculum increases. Thus, item localization appears to be a reasonable approach intended to reduce the linguistic misalignment of distal and remote tests.

Localization refers to the "process of adapting text and cultural content to specific target audiences in specific locations" (WorldLingo, 2004) or, more specifically, to the process of adapting a product or service to the language and cultural requirements of a specific target environment and make it look as if it had been created in that target environment (Tunick, 2003). It recognizes that every location is unique by virtue of a series of linguistic and cultural factors and stresses that efforts to adapt text to a given target group must go beyond the pure formal level (Muzzi, 2001). Surprisingly, the notion of localization has not been adopted in the field of testing.

My colleagues and I are currently investigating whether item localization reduces measurement error due to language (Solano, Speroni, & Sexton, 2005). This research is being conducted with ELLs who are tested in English. We have facilitated item localization sessions in which teachers modify the vocabulary and the wording of mathematics items according to the language used in their classrooms and in the enacted curriculum. We have tested students with the standard and localized versions of the same set of items. The procedure is the same as when we have tested students in two dialect versions of their native languages (Solano-Flores & Li, 2006). If our reasoning about linguistic misalignment holds, we should observe that students perform better in localized than in standard versions of tests.

ITEM REVIEW

Judgmental test review methods pay considerable attention to the adequacy of the wording of items to the characteristics of the target population (e.g.,

Hambleton & Rodgers, 1995). There is evidence that the effectiveness of item review procedures is shaped by factors such as the quality of the training given to reviewers (Hambleton & Jones, 1994).

Unfortunately, there is also evidence that the linguistic features identified by teachers in test items as potential sources of bias against their own students may differ considerably from the challenges observed when those students read those items aloud or when they are interviewed about them (Solano-Flores, Li, & Sexton, 2005). Teachers may have inconsistent or inaccurate perceptions of the linguistic aspects (including vocabulary, syntax, semantics, and pragmatics) and cultural aspects (meaningfulness and appropriateness) of items that are critical to properly understanding them. Moreover, being a native speaker of, or being familiar with, the native language of the students' first language does not necessarily enable teachers to identify those factors (Sexton & Solano-Flores, 2002).

Evidence from the field of test translation review in the context of cross-language comparisons underscores the need for enriching test review procedures with reasonings from sociolinguistics. Solano-Flores, Contreras-Niño, and Backhoff (2005) facilitated a series of test translation review sessions with a multidisciplinary team that included teachers in service, measurement and curriculum specialists, a certified English-Spanish translator, and a linguist. This multidisciplinary team examined the quality of the translation of a sample of items of the Mexican Spanish-language version of the TIMSS (Third International Mathematics and Science Study)-1995 science and mathematics items, Populations 1 (grades 4 and 5) and 2 (grades 7 and 8). Our review system addressed, among other aspects, register (translation accord of the items with the language used in the Mexican curriculum).

We observed that register was not properly addressed in the translation of a considerable number of items. Register issues were observed among about 18% and 17% of the mathematics items, for Population 1 and Population 2, respectively, and among about nearly 10% and 23%, respectively, of the science items for Populations 1 and 2. These errors made their way through the final version of the translation in spite of the procedure that TIMSS used to examine the translations of all participant countries before they were authorized to use them with their students (see Mullis, Kelly, & Haley, 1996).

It should be mentioned that not all the errors related to register could be considered as severe. It should also be mentioned that international guidelines for test translation and test adaptation have been revised since 1995 (compare, for example, Hambleton, 1994, and Hambleton, Merenda, & Spielberger, 2005) and keep evolving as experience with item comparability across language accrues (see Grisay, 2003). However, our findings speak to the relevance of register in test review. Review guidelines appear not to be

sufficient to guarantee linguistic alignment in the absence of adequate procedures that ensure fidelity of implementation and without explicit consideration to register.

ITEM SAMPLING

Item sampling procedures should focus on the fact that linguistic misalignment occurs despite the actions intended to address language bias during the process of test development and test review. In testing theory, test items are thought of as samples drawn from a knowledge domain (Bormuth, 1970; Guttman, 1969; Hively, Patterson, & Page, 1968). Similar reasonings can be used to address linguistic misalignment. Test items can be thought of as unintended samples of dialect and register, as Figure 6 illustrates. Dots represent the linguistic features of all the items in a test. Some of these features match the linguistic backgrounds of ELLs; they appear in the intersection $T \cap U$. The others, which are represented in the area of T that is outside $T \cap U$, are instances of linguistic misalignment.

Item sampling appears to be critical to effectively addressing linguistic misalignment. The number of items included in a test should be large enough to minimize the impact of linguistic misalignment.

The optimal number of items to be used in a test is always an issue of practical and theoretical relevance. However, what makes the proposed item-sampling perspective special is its focus on ELLs. Although having as many items as possible is always the plea of measurement specialists, this interest is always driven by concerns about item, rater, occasion, and so on,

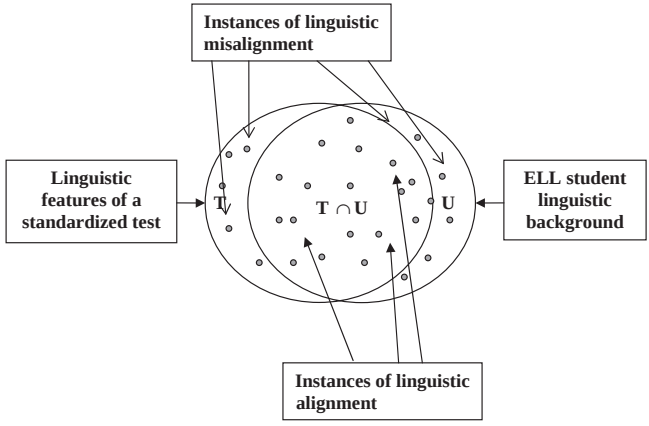

Figure 6. Linguistic Alignment and Misalignment of the Linguistic Features of a Test with the Features of the Language to Which Students Have Been Exposed

as sources of measurement error. When a population of examinees includes ELLs, this interest should also be driven by concerns about sources of measurement error related to language. Hence, in reasoning about score dependability and the number of items needed in a test, special consideration should be given to the dependability coefficients obtained for ELLs.

This focus on ELLs makes it possible to devise new testing procedures intended to both produce valid measures of ELL academic achievement and comply with regulations that make ELL testing in English mandatory. For example, procedures can be devised in which the minimum number of items used to test an entire population of students (i.e., both ELLs and non-ELLs) is determined based on the dependability coefficients obtained with ELLs. Other procedures can be devised in which ELLs and non-ELLs are tested with different numbers of items, based on the dependability of the scores obtained for each segment of the population. Yet other procedures can be devised in which different groups of ELLs within the same linguistic groups are tested with different numbers of items, based on the dependability coefficients obtained for each group.

SUMMARY AND CONCLUSIONS

In this article, I have discussed the intersection of sociolinguistics and psychometrics in the testing of English language learners (ELLs). The ideas presented have important implications for the testing of young ELLs, whose performance in tests can be especially sensitive to slight language variations (e.g., wording, use of idiomatic expressions) regardless of the language in which they are tested.

I have discussed dialect and register as subordinate and interconnected categories of language, in the sense that there may be many dialects of the same language and many registers within the same language. Dialects are varieties of a language according to the users of a language; registers are varieties of that language according the uses of that language. The concept of dialect is relevant to addressing the tremendous linguistic and cultural diversity that may exist within broad linguistic groups of ELLs who are users of the same language; the concept of register is relevant to examining academic language used in tests and the vocabulary and expressions that are specific to tests.

I have also discussed how generalizability (G) theory (a psychometric theory of measurement error) can be used in ELL testing. Unlike other psychometric approaches to ELL testing, which address language based on examining item bias or score differences between groups (e.g., ELLs vs. non-ELL students), G theory focuses on measurement error. From this perspective, validity and fairness are not only a matter of score gaps between ELLs and non-ELLs but also a matter of score dependability.

My colleagues and I have used G theory to examine how the performance of ELLs varies when they are tested with the same set of items in two languages or in two dialects of the same language. We have observed that dialect can be as important as language as a source of measurement error in the testing of ELLs. Language factors that affect the dependability of achievement measures take place at the level of dialect. No matter what language you use (either your first or your second language), you use a dialect, standard or nonstandard, of that language.

I have discussed how linguistic misalignment—the mismatch between the features of the dialect and the register used in a test, and the features of the language to which students have been exposed—can be addressed through item development, item localization, item review, and item sampling procedures. Some of the ideas proposed include entrusting local communities to generate items according to common sets of item specifications; localizing items so that their wording is adapted to the characteristics of each community but their content and format are kept unchanged across communities; enriching item-review procedures with criteria based on concepts from the field of sociolinguistics; and using the dependability coefficients obtained with ELLs to determine the number of items used to test both ELLs and non-ELLs.

Needless to say, some of these ideas challenge the notion of standardization in testing and have serious methodological implications. Research is needed to determine the costs, limitations, and advantages of their implementation. However, what is important about these ideas is the common underlying notion that effective actions oriented toward addressing language and improving ELL testing cannot be taken in isolation without changing what is done for all students. This notion is consistent with the following statement, frequently cited in documents that address the testing of minorities: "Fairness, like validity, cannot be properly addressed as an afterthought once the test has been developed, administered, and used. It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation, and use" (Heubert & Hauser, 1999, pp. 80–81).

Language, dialect, and register are critical to producing valid, fair measures of academic achievement for ELLs. But actions intended to minimize the linguistic misalignment of items should not be thought of as exclusive to ELLs; they need to involve both ELLs and non-ELLs. If we are serious about improving testing practices for ELLs, we need to improve our views of testing.

The research work reported here was supported by the National Science Foundation Grants Nos. REC-0126344, REC-0336744, and REC-0450090. I wish to thank Rich Shavelson, Barry Sloane, and Larry Sutter for their advice and comments. I am grateful to the participant

schools and students for welcoming me and my colleagues in their communities and classrooms, and to their teachers for their enthusiastic participation as translators or raters. I am grateful to Janette Klingner, Margaret LeCompte, and Alfredo Artiles for encouraging me to write this article, and to Richard Valencia, Jeff MacSwan, Phil Esra, Diane Doyal, and three anonymous reviewers for their comments. I am indebted to Min Li, Elise Trumbull, Julia Lara, and Yael Kidron for their expert collaboration in the projects here reported, and to Cecilia Speroni, Melissa Kwon, and Ursula Sexton for their enthusiastic participation in data collection and data analysis. The opinions I express in this article do not necessarily reflect the position, policy, or endorsement of the funding agency or the opinions of my colleagues.
A previous version of this article was presented at the conference, English Language Learners Struggling to Learn: Emergent Research on Linguistic Differences and Learning Disabilities, organized by the National Center for Culturally Responsive Educational Systems (NCCRESt), Arizona State University, Tempe, Arizona, November 18–19, 2004.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, *8*, 231–257.

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., 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. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16–26.

Aguirre-Muñoz, Z., & Baker, E. L. (1997). *Improving the equity and validity of assessment-based information systems* (CSE Tech. Rep. No. 462). Los Angeles: University of California, Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education and Information Studies.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, *16*, 55–73.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language minority students: A research agenda*. Washington, DC: National Academy Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*, 239–257.

Bialystok, E. (Ed.). (1991). *Language processing in bilingual children*. Cambridge, UK: Cambridge University Press.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge, UK: Cambridge University Press.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning*, *32*, 245–258.

Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Briggs, D. C., & Wilson, M. R. (2004, June.). Generalizability in item response modeling. Presentation at the International Meeting of the Psychometric Society, Pacific Grove, CA.

Brisk, M. E. (1998). *Bilingual education: From compensatory to quality schooling*. Mahwah, NJ: Erlbaum.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, *34*, 21–42.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* ((CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Casanova, U., & Arias, B. (1993). Contextualizing bilingual education. In M. B. Arias & U. Casanova (Eds.), *Bilingual education: Politics, practice, and research* (pp. 1–35). Chicago: University of Chicago Press.

Cole, M. (1999). Culture-free versus culture-based measures of cognition. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 645–664). Cambridge, MA: MIT Press.

Cook, L. I., & Schimitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139–169). Mahwah, NJ: Erlbaum.

Coulmas, F. (2005). *Sociolinguistics: The study of speakers' choices*. Cambridge, UK: Cambridge University Press.

Council of Chief State School Officers. (1992). *Recommendations for improving the assessment and monitoring of students with limited English proficiency.* Washington, DC: Author.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Crystal, D. (1997). *The Cambridge encyclopedia of language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* Clevedon, UK: Multilingual Matters.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society.* Los Angeles: California Association for Bilingual Education.

Cummins, J. (1999). Alternative paradigms in bilingual education research: Does theory have a place? *Educational Researcher*, *28*, 26–32, 41.

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

Cummins, J. (2003). BICS and CALP: Origins and rationale for the distinction. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings* (pp. 322–328). London: Blackwell.

De Avila, E. (1990). Assessment of language minority students: Political, technical, practical and moral imperatives. *Proceedings of the First Research Symposium on Limited English Proficient Student Issues*. Retrieved February 9, 2005, from http://www.ncela.gwu.edu/pubs/symposia/first/assessment.htm

Edelsky, C. (1990). *With literacy and justice for all. Rethinking the social in language and education*. London: Falmer Press.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research*, *29*, 543–553.

Farr, M., & Ball, A. F. (1999). Standard English. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 205–208). Oxford, UK: Elsevier.

Ferguson, A. M., & Fairburn, J. (1985). Language experience for problem solving in mathematics. *Reading Teacher*, *38*, 504–507.

Ferrara, S., Macmillan, J., & Nathan, A. (2004, January). *Enhanced database on inclusion and accommodations: Variables and measures* (NAEP State Analysis Project Report to the National Center for Education Statistics). Washington, DC: NCES.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*, 1–43.

Gandara, P. (1997). *Review of research on instruction of limited English proficient students*. Davis: University of California, Linguistic Minority Research Institute Education Policy Center.

Genesee, F. (1994). Introduction. In F. Genesee (Ed.), *Educating second language children: The whole child, the whole curriculum, the whole community* (pp. 1–11). Cambridge, UK: Cambridge University Press.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, *52*, 1115–1124.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, *20*, 225–240.

Guerrero, M. D. (1997). Spanish academic language proficiency: The case of bilingual education teachers in the U.S. *Bilingual Research Journal*, *21*(1). Retrieved April 1, 2004, from http://www.ncela.gwu.edu/pubs/nabe/brj/v21.htm

Guttman, L. (1969). Integration of test design and analysis. In *Proceedings of the 1969 invitational conference on testing problems* (pp. 53–65). Princeton, NJ: Educational Testing Service.

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. schools: Report and workshop summary*. Washington, DC: National Academy Press.

Hakuta, K., Goto Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Report No. 2000-1). Davis: University of California, Linguistic Minority Research Institute.

Hakuta, K., & McLaughlin, B. (1996). Bilingualism and second language learning: Seven tensions that define the research. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 603–621). New York: Simon & Schuster Macmillan.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Hamayan, E., & Perlman, R. (1990). *Helping language minority students after they exit from bilingual/ESL programs*. Washington, DC: National Clearinghouse for Bilingual Education.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229–240.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly*, *18*, 23–36.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Hambleton, R., & Rodgers, J. (1995). Item bias review. *Practical Assessment, Research & Evaluation*, *4*. Retrieved October 31, 2004, from http://PAREonline.net/getvn.asp?v=4&n=6

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Haugen, E. (1966). Dialect, language, nation. *American Anthropologist*, *58*, 922–935.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education*, *16*, 159–188.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Krashen, S. D. (1996). *Under attack: The case against bilingual education*. Culver City, CA: Language Education Associates.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids: A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, *64*, 55–75.

Lee, O. (1999). Equity implications based on the conceptions of science achievement in major reform documents. *Review of Educational Research*, *69*, 83–115.

Lee, O. (2002). Promoting scientific inquiry with elementary students from diverse cultures and languages. *Review of Research in Education*, *26*, 23–69.

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*, 491–530.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English language backgrounds. *Educational Researcher*, *27*, 12–21.

Lee, Y. W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. Princeton, NJ: Educational Testing Service.

MacSwan, J., & Rolstad, K. (2003). Linguistic diversity, schooling, and social class: Rethinking our conception of language proficiency in language minority education. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings* (pp. 329–340). Malden, MA: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education, and National Council on Measurement in Education.

Molloy, H., & Shimura, M. (2005). An examination of situational sensitivity in medium-scale interlanguage pragmatics research. In T. Newfields, Y. Ishida, M. Chapman, & M. Fujioka (Eds.), *Proceedings of the May 22–23, 2004, JALT Pan-SIG Conference* (pp. 16–32. Retrieved September 8, 2005, from www.jalt.org/pansig/2004/HTML/ShimMoll.htm.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation verification procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection*. Chestnut Hill, MA: Boston College.

Muzzi, A. (2001). Challenges in localization. *The ATA Chronicle*, *30*(11), 28–31.

National Assessment of Educational Progress. (1996). *Mathematics items public release*. Washington, DC: Author.

National Assessment of Educational Progress. (2000). *Mathematics items public release*. Washington, DC: Author.

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

Preston, D. R. (Ed.). (1993). *American dialect research*. Philadelphia: John Benjamins.

Preston, D. R. (1999). Sociolinguistics. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 65–70). Amsterdam: Elsevier.

Rickford, J. R., & Rickford, A. E. (1995). Dialect readers revisited. *Linguistics and Education*, 7(2), 107–128.

Rivera, C. (1984). *Language proficiency and academic achievement*. Clevedon, UK: Multilingual Matters.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393.

Sexton, U., & Solano-Flores, G. (2002, April). *A comparative study of teachers' cultural perspectives across different cultures: Preliminary results*. Poster presentation at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English proficient students in Rhode Island's Grade 4 mathematics performance assessment* (CSE Tech. Rep. No. 486). Los Angeles: University of California; Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education and Information Studies.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Mahwah, NJ: Erlbaum.

Solano-Flores, G., & Backhoff, E. (2003). *Test translation in international comparisons: A preliminary study*. Mexico City, Mexico: Mexican Department of Education, National Institute for Educational Evaluation.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2005, April). *The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English language learners: A sampler of student responses to science and mathematics test items*. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, *25*, 13–22.

Solano-Flores, G., Li, M., & Sexton, U. (2005). *On the accuracy of teacher judgments of the linguistic and cultural adequacy of test items*. Unpublished manuscript.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*, 553–573.

Solano-Flores, G., Speroni, C., & Sexton, U. (2005, April). *The process of test translation: Advantages and challenges of a socio-linguistic approach*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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.

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April). *The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners*. Paper presented at the 2003 Annual Meeting of the American Evaluation Research Association, Chicago, IL.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*, 107–129.

Stansfield, C. W., & Kenyon, D. M. (1992). Research of the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*, 347–364.

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

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–263). Mahwah, NJ: Erlbaum.

Tunick, L. (2003). Language translation, localization, and globalization. *The ATA Chronicle*, *32*(2), 19–21.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

van de Vijver, F., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29–37.

van de Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*, 263–279.

Wardhaugh, R. (2002). *An introduction to sociolinguistics* (4th ed.). Oxford, UK: Blackwell.

Wolfram, W., Adger, C. T., & Christian, D. (1999). *Dialects in schools and communities*. Mahwah, NJ: Erlbaum.

WorldLingo. (2004). Localization. *Glossary of terms*. Retrieved May 24, 2004, from http://www.worldlingo.com/resources/glossary.html

Yoo, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the California Mathematics Renaissance* (CSE Tech. Rep. No. 484). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

GUILLERMO SOLANO-FLORES is associate professor of bilingual education and English as a second language at the University of Colorado, Boulder. His research examines ways in which linguistics and psychometrics can be used in combination to develop improved models for the testing of linguistically diverse populations in international test comparisons and in the testing of linguistic minorities.