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.

Examining Language in Context: The Need for New Research and Practice Paradigms in the Testing of English-Language Learners

Guillermo Solano-Flores and Elise Trumbull

Concerns about how to ensure the valid and equitable assessment of English-language learners (ELLs) and other students from culturally non-mainstream backgrounds are longstanding. This article proposes that new paradigms in the research and practice related to ELL testing are needed to address the complexities of language and culture more effectively. Three main areas are identified as key to this paradigm shift: test review, test development, and treatment of language as a source of measurement error. Research examples are provided that illustrate that the proposed paradigm shift is not only necessary but also possible. The authors propose the combined use of generalizability theory and research designs in which ELLs are given the same items in both English and their native languages—an approach that has the potential to reveal more fine-grained understandings of the interactions among first and second language proficiency, student content knowledge, and the linguistic and content demands of test items.

A fundamental notion in test validity is that low test scores should not occur because of factors that are irrelevant to the construct an instrument intends to measure (Messick, 1989, 1995a). Yet, assessments often confound the language skills of examinees with their academic aptitudes (Durán, 1989; Valdés & Figueroa, 1994). The longstanding concern that any test is to some degree a test of language proficiency (e.g., Sanchez, 1934) continues to the present day (American Educational Research Association, 1999; LaCelle-Peterson & Rivera, 1994).

Despite legislation intended to ensure that test items address the needs of English-language learners (ELLs) and despite efforts to increase their participation in national testing (see Pellegrino, Jones, & Mitchell, 1999), the progress toward more valid and equitable testing of linguistic minorities has been small (Hakuta & Beatty, 2000). In the *Evaluation of the Voluntary National Tests* (Wise, Hauser, Mitchell, & Feuer, 1999), the Board on Testing and Assessment concludes that existing plans for including and accommodating the needs of ELLs "are sketchy and do not yet break new ground with respect to maximizing the degree of inclusion and the validity of scores for all students" (p. 3).

Practices in the testing of ELLs are often driven by policies rather than theory (Hakuta & McLaughlin, 1996)—a fact that casts doubts on the validity of measures of ELL

academic achievement (Rivera, Vincent, Hafner, & LaCelle-Peterson, 1997). As a result, definitions and classifications of ELLs vary both across and within states (Casanova & Arias, 1993; Council of Chief State School Officers, 1992) and are often based on inaccurate criteria such as ethnicity, immigrant status, or the number of years lived in the United States (see Aguirre-Muñoz & Baker, 1997; Steele & Aronson, 1995).

This dearth of effective testing approaches marginalizes ELLs in an educational system that increasingly relies on test-based accountability as a promoter of educational change (see No Child Left Behind, among others). The nearly 4.5 million ELL students who constitute approximately 9.3% of the public school enrollment of students in Pre-K through Grade 12 in the United States (Kindler, 2002) will continue to be adversely affected by flawed assessment practices because no better methods exist for their testing.

In this article, we contend that existing approaches to testing ELLs do not ensure equitable and valid outcomes because current research and practice assessment paradigms overlook the complex nature of language, including its interrelationship with culture. Although we do not mean to disparage the commendable efforts made in recent years in the field of test accommodations for ELLs (e.g., Butler & Stevens, 1997, and the work of Abedi and his associates), we do see a need to expand the theoretical framework that guides ELL research and testing. Our effort is a response to the call for new research on assessment design that includes exploration of systematic and fair methods for taking into account the processes by which culture influences student performance (Pellegrino, Chudowsky, & Glaser, 2001).

Three notions are critical to our discussion. First, like many others before us (cf., Bruner, 1993; Calfee & Berliner, 1996; Cocking & Mestre, 1988; Jacob et al., 1996; Rogoff, 1990; Saxe, 1991; Snow & Lohman, 1989), we recognize that contextual factors play a critical role as mediators of cognitive processes. It has been proved that items elicit different response processes from different tests takers (Kupermintz, Le, & Snow, 1999). Affective and conative variables shape these differences as they interact with the knowledge and skills students bring to the testing situation (Snow, 1993). Tests are cultural products (Cole, 1999; Estrin & Nelson-Barber, 1995), and taking a test is an event for which each student has a "conceptual frame" (Swisher & Deyhle, 1992). Students' varying cultural and linguistic backgrounds may prepare them with different "scripts" (schemes) or principles for approaching such an event. Second, we believe that the notion that culture and society shape minds (e.g., Vygotsky, 1978; Wertsch, 1985) deserves more attention in the field of testing. A growing body of research shows that the cognitive activity of students when they take tests is an important source of evidence of test validity (e.g., Baxter, Elder, & Glaser, 1996; Hamilton, Nussbaum, & Snow, 1997; Ruiz-Primo, Shavelson, Li, & Schultz, 2001). Whereas it has been recognized that cognitive validity can provide evidence on construct-relevant and irrelevant sources of score variance (Messick, 1995b), this perspective on validity still reflects a view of cognition as independent of culture (Resnick, 1991). A sociocultural view of cognition is needed that allows for approaching cognitive processes in test taking with the support of theories of language and culture (Lee & Fradd, 1998; Lee, 1999, 2002). Third, bilingual proficiency can be thought of as a continuum along which bilingual individuals fall at different points, depending on the varying strengths and cognitive characteristics they exhibit in their two languages (Valdés & Figueroa, 1994). ELLs can be viewed as (at least incipient) bilingual individuals whose proficiencies in English and in their native

languages vary considerably. They may have different patterns of language dominance (Genesee, 1994; Stevens, Butler, & Castellon-Wellington, 2000), and their strengths may be expressed differently in different contexts (e.g., home or school) and in the written and oral modes (Hakuta, Ferdman, & Diaz, 1987). Our discussion focuses on three aspects of ELL testing that we believe are critical to a paradigm shift: test review, test development, and treatment of language as a source of measurement error.

Test Review

We address the need for a comprehensive conceptual framework that takes into account the cognitive, semantic, communicative, and sociolinguistic factors involved in testing (August & Hakuta, 1997; Bransford, Brown, & Cocking, 1999). We believe that this conceptual framework must be grounded in the practice of closely examining test items from the perspectives of different disciplines, each at a different level of complexity, from the structural linguistics perspective to the sociocultural perspective. Attempts to promote more equitable testing at the item level have been mainly oriented to ensuring accurate interpretation of ELL student written responses (e.g., Kopriva & Saez, 1997; Shaw, 1997). However, although laudable, these approaches are not multidisciplinary and focus on scoring rather than test review.

Item microanalysis is a highly useful technique for test review. We define *item microanalysis* as the set of reasonings used to examine how the properties of items and students' linguistic, cultural, and socioeconomic backgrounds operate in combination to shape the ways in which students make sense of test items. Item microanalysis is intended to be comprehensive in its scope, which is necessarily multidisciplinary. It attempts to address the fact that the way a student construes the testing situation itself is relevant.

The microanalysis of an item focuses on three kinds of item properties: *Formal* properties are identified from examining the language and wording used in the item. *Empirical* properties are identified from examining student think-aloud protocols (which include reading the item aloud) and interviews conducted with the intent to see how students make sense of items. *Differential* prop-erties are identified from examining how observed and formal properties operate in combination with the students' linguistic and cultural backgrounds to shape their interpretations of items.

The following example illustrates the value of item microanalysis. It arises from a project that investigated how students' cultural backgrounds influence the way they interpret test items (Solano-Flores & Nelson-Barber, 2001). Several items from the National Assessment of Educational Progress (NAEP) were used in this project. These items (used previously in standardized testing in the past and deemed psychometrically sound) were given to Grade 4 and Grade 5 students from different cultural backgrounds. The results reported here are from both the pilot and field samples, which included, respectively, ELLs and students not classified as ELLs by their school districts but whose native language was not English. Students read the items aloud and responded to them; then they were interviewed individually to determine the reasoning they used to respond to the items and how they connected items' content to their everyday lives.

The Lunch Money item (National Assessment of Educational Progress, 1996) is one of the items used in that study:

> Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?

Different disciplinary perspectives allowed us to identify properties of the Lunch Money item that might affect ELLs adversely (Appendix). In our analysis of the item's formal properties, we focused on linguistic properties, which have been proven relevant in research on accommodations for ELLs (e.g., Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, Hofstetter, & Baker, 2001). However, in addition to structural linguistics, we used discourse analysis (see Brown & Yule, 1983), which allowed us to examine how cohesion across sentences in the item is maintained (Lagunoff, Solano-Flores, Sexton, & Nelson-Barber, 2001).

As a second step, the analysis of student verbal protocols allowed us to examine the item's empirical properties, especially in relation to words students had problems reading or interpreting correctly. While understanding some of these words in varied contexts may be critical to acquiring a "mathematics register" in English, many of the students did not interpret them correctly even if they were used colloquially. Of particular interest was the word *only* in the third sentence ("His mother has only $1.00 bills"), which was interpreted by some students as restricting the number of dollar bills rather than the number of dollar denominations.

As a third step, a sociocultural perspective allowed us to identify the item's differential properties. We suspected that student interpretations of *only* in the third sentence might reflect different ways of making sense of the item that were influenced by socioeconomic background. This possibility did not seem unreasonable, given that the effect of poverty on student performance is well documented (e.g., Krashen, 1996; Oakes, 1990), as is the influence of psychological factors in test taking (e.g., Steele, 1997). In addition, there is evidence that students from certain cultural groups tend to incorporate emotion and personal perspectives in their arguments (Kochman, 1989). Moreover, the way they relate past experience to the testing context can affect their performance on tests (Heath, 1983, 1986).

Consistent with these kinds of observations, a closer analysis of the readings of the students from different groups revealed that students of different socioeconomic status tended to interpret the words (and the item) differently (Solano-Flores, Li, & Sexton, 2002). In this part of the study, we focused on three groups of students: (a) White, suburban, high income; (b) American Indian, rural, low income; and (c) African American, inner city, low income. Notice that, although these groups may have had English dialect differences, none of them included ELLs. This comparison allowed us to test more formally whether different cultural backgrounds associated with poverty may produce different interpretations of one word in a specific sentence.

We found that whereas 84% of White students read the third sentence as intended, only 56 and 52%, respectively, of American Indian and African-American students read the sentence as intended. This analysis also revealed that 10 and 18%, respectively, of the American Indian and African-American students interpreted the word *only* as restricting the number of dollars ("His mother has only one dollar"). This incorrect interpretation was not observed among White students.

While reading proficiency cannot be dissociated altogether from socioeconomic status

in this study, the results suggest that low-income students may project their own concerns and experiences onto the way they solve the Lunch Money problem. Low-income students may be more accustomed to solving money problems by forgoing purchase than by planning how to spend money in advance. As a result, they may interpret the item as if they were being asked, "What can Sam buy with $1.00?" and give survival-oriented responses.

The story of Sam's buying lunch at school may have been included as part of the item with the intent of providing contextual information that would be meaningful to all. Ironically, the teachers from low-income schools that participated in this study regarded this setting as unsuitable for their students because most of them are on free-lunch programs and do not buy lunch at school (Sexton & Solano-Flores, 2001; Solano-Flores, 2002). As the following excerpt suggests, in interpreting the item, not having $1.75 to buy lunch may be a more pressing issue in the mind of a low-income student than figuring out how many bills of the same denomination are needed for one week:

> Researcher (R)*:* Now, what do you think this question is asking from you? What is it about?
>
> Student (S): It's about Sam and he wants to buy his juice, his sandwich and his fruits.
>
> R: Mm-hm.
>
> S: For lunch. Maybe he was hungry. But, I think his mom didn't have enough money.
>
> R: Why?
>
> S: Because she only had one dollar bill.
>
> (R asks a question that S does not understand; R rephrases.)
>
> R: So, what did you need to know to be able to answer this problem?
>
> S: I had to know, um, do, um, I had to do the math problems, like, how much money needed, um, check how many money he needed for five days and how much, uh, juice and sandwich and fruit costs and his mother only, his mother only had one dollar bill and, and that's all.

The microanalysis of the Lunch Money item shows that we cannot reasonably think about language without considering sociocultural contexts. Nor we can reasonably think about linguistic diversity without considering socioeconomic issues because many linguistic and cultural minority students live in poverty. One may wonder why the unnecessary linguistic challenges of the Lunch Money item and their potentially adverse impact for linguistic minorities could not be detected before this item made its way into a national standardized test. One good reason is that currently accepted assessment development practices fail to address language and culture in any depth.

Test Development

It is well known that student performance is extremely sensitive to wording, even if assessments are constructed for a population of non-ELLs in their own language (e.g., Baxter, Shavelson, Goldman, & Pine, 1992). Because of the lack of systematic and comprehensive approaches to test construction (Frederiksen, 1990; Shavelson, Carey, &

Webb, 1990; Ruiz-Primo, 2002), test developers' efforts tend to be limited largely to refining wording with the intent of ensuring that students understand what they are asked to do and can communicate their responses effectively (Solano-Flores & Shavelson, 1997).

When testing involves linguistically diverse populations, wording issues become even more serious. The position that it is impossible to measure the same construct across languages (Greenfield, 1997) may seem extreme or, at least, not helpful for practical purposes. However, it underscores the difficulty of attaining accuracy when tests are adapted for students who are not from the original target linguistic group. For example, even if we assume that a test is translated in accordance with existing, well-established norms (e.g., Geisinger, 1994; Hambleton, 1994), the test in the source language and the test in the target language are, strictly speaking, developed with very different procedures (Tanzer, in press). Whereas the wording of the former is carefully refined and the test is tried out with pilot students, the translated version is developed in a much shorter period of time and with fewer review iterations (Solano-Flores & Nelson-Barber, 2001).

During the past few years, we have investigated the effectiveness of a test development approach specifically oriented to addressing linguistic diversity. This approach enables test developers to consider language issues in depth, with reference to a specific sociocultural context throughout the entire process of test development.

The concurrent assessment development model was created as an alternative to translation, with the intention of promoting more equitable testing (Solano-Flores, Trumbull, & Nelson-Barber, 2002). According to this model, two language versions of the same assessment are developed concurrently and interactively by two teams of assessment developers, each responsible for one of the language versions. *Concurrently* means that the two versions evolve together; both versions undergo the same number of review-tryout-revise iterations and are piloted with students the same number of times. *Interactively* means that any modifications and improvements made on one language version must also be made on the other version, upon agreement of members of both development teams.

With this model, we generated Grade 4 mathematics con-structed-response items. We conducted a series of development sessions with bilingual teachers from a district that serves large numbers of Latino, native Spanish-speaking students. The teachers broke up into two teams, each responsible for developing one language version of the same assessment. These teams were provided with a template that specified the characteristics of a type of mathematics communication task linked to the state's communication standard (Washington State Commission on Student Learning, 1999), similar to the communication standard of the National Council of Teachers of Mathematics (1989). This template provided a formal, abstract description of the items to be generated and was used as a reference by the developers during their discussions throughout the entire process of development.

As a part of the cyclical process of review, teachers piloted the first versions of the items with small numbers of students, who were asked why they solved the problems as they did and what they would change in the wording of each item to make it better. On each iteration of this process and upon examining the students' responses to the items, each team proposed a series of changes that were discussed and negotiated across languages.

We observed that as this process went on, test developers reached increasingly deeper levels of analysis in their discussions of language issues relevant to their mathematics items (Table 1). We identified four stages in the teachers' thinking and actions during their discussions. In Stage 1, issues addressed included formal characteristics of language and formal equivalence of technical terms across languages. Although there was awareness that language issues cannot be addressed without considering culture, attempts to address culture were based on overgeneralizations about cultural groups.

In Stage 2, developers realized that to address language more effectively, they needed to identify which aspects of language and culture are relevant to testing. They attempted to ensure that in addition to being formally equivalent, the two language versions of the same test needed to be functionally equivalent.

In Stage 3, they came to understand that knowledge of certain principles for testing linguistic minorities and awareness of certain general characteristics of a broad cultural group do not suffice to effectively address language issues. Their analysis then focused on the characteristics of the specific group of students for which they were developing the items. In addition to taking the identity of their students into consideration (e.g., immigrant Latino, native Spanish-speaking students), their discussions reflected an attempt to identify the sociocultural context in which these students live and to take into consideration the dialect they speak. Underlying this analysis was the realization that any action intended to ensure accuracy and fairness should be based on accurate knowledge of their cultural and linguistic context.

In Stage 4, developers realized that the wording used in an item might be reflecting the communication styles of a particular cultural group. This notion is consistent with current thinking in cultural psychology and anthropology, which recognizes that cultural groups may differ considerably in the ways they communicate (e.g., Heath, 1986; Philips, 1983), the ways arguments are structured (Kaplan, 1988; Tsang, 1989), and the ways language integrates social discourse and "academic" discourse (Diaz, Moll, & Mehan, 1986; Greenfield, 2000; Trumbull, Diaz-Meza, & Hasan, 2000).

Thus, in addition to ensuring the use of proper syntactic structures, a way of attaining item equivalence across languages may consist of making the overall structure of items parallel to culturally determined discourse patterns. Structural concerns may also extend to visual information. For example, in one language, a given item might read as follows: "Figure out how much money

Table 1
Stages in the Process of Developing Tests Concurrently Across Languages

Stage	Test developer action and reasoning	Example of issues raised
1. Literal equivalence across languages	Developers' actions are guided by personal perceptions or superficial characteristics of culture.	Is technical terminology accurate in both languages? Are spelling and sentence structure correct?
2. Appropriateness of language to a broad cultural group	Developers attempt to identify specific aspects of culture that are relevant to testing.	Should decimal metric system units be used in the exercises in Spanish? If English system units are kept in the Spanish version, should English or Spanish abbreviations be used?
3. Appropriateness of language to a sociocultural context	Developers attempt to identify aspects of the students' everyday life experiences that are relevant to the items being developed.	Are these students more familiar with kilograms and grams (which are used in their home countries) or with pounds and ounces (which are part of their everyday life experience in the United States)?
4. Correspondence between item structure and discourse patterns	Developers become aware that the structure of some items might need to reflect the patterns of discourse inherent to the target students' dialect.	Should the sequence of the item components (e.g., contextual information, table with numeric information, space for computations) be the same for both languages, or should the sequence be different for each language?

your classroom will need to build a bird feeder. Use the information provided in the table below." In another language, the same item might display the table first, then ask the following: "Look at the information provided in the table above. Figure out how much money your classroom will need to build a bird feeder." The effectiveness of the concurrent assessment development model is still being investigated. However, the fact that Stages 3 and 4 are rarely reached in current testing practices speaks to the level of

specificity and depth at which language can and needs to be addressed during the process of test development.

Treatment of Language as a Source of Measurement Error

Bilingual individuals perform differently from monolingual speakers on tests, as there are certain mental processes and abilities that are specific to the condition of being bilingual (Bialystok, 1997; Valdés & Figueroa, 1994). This unique condition has not been properly taken into consideration by current research and practice in the testing of ELLs. As in cross-cultural research, which focuses on differences between groups (see Van de Vijver & Poortinga, 1997), designs used in ELL testing are based mainly on comparing score differences between ELLs and mainstream students (Shepard, Taylor, & Betebenner, 1998).

A simple conceptual framework shows the limitations of current research designs used in the testing of ELLs. Research in the field of ELL testing uses four types of comparisons based on whether they involve one or two linguistic groups and one or two test languages (Table 2). For example, to examine the performance of students who were tested in English, Abedi, Lord, Boscardin, and Miyoshi (2001) used three comparisons: ELLs versus non-ELLs tested with and without accommodations (Type I); ELLs tested with accommodations versus ELLs tested without accommodations (Type III-a); and non-ELLs tested with accommodations versus non-ELLs tested without accommodations (Type III-b). Type II comparisons are the same as those typical of international assessments, such as the Third International Mathematics and Science Study (TIMSS), which examined performance differences between groups of speakers of different countries, each tested in its own native language (e.g., Ercikan, 1998). In Type IV-a comparisons, one group of the same population of ELLs is tested in English and the other in its native language.

Except for Type IV-b, in all comparisons used in ELL testing, a given ELL student is tested in either English or his or her native language, yet

> When a bilingual individual confronts a monolingual test, . . . both the test taker and the test are asked to do something that they cannot. The bilingual test taker cannot perform like a monolingual. The monolingual test cannot 'measure' in the other language. (Valdés & Figueroa, 1994, p. 87)

In contrast, in Type IV-b comparisons the same students take the same items in both English and their native language. In this type of comparison, language is viewed as a source of measurement error, an approach that allows for examining the quality of the scores obtained by ELLs. Surprisingly, Type IV-b comparisons have not been used in the research on ELL testing, with the exception of the study described subsequently.

We (Solano-Flores, Lara, Sexton, & Navarrete, 2001) assembled a sampler of responses given by ELLs to the same set of items when they were tested in English and when they were tested in their native language. (Appropriate actions were taken to control for the possible effects of sequence or learning resulting from taking the same item twice in different languages.) As a part of the strategies for ensuring the adequacy of the translations to the students' developmental level in their own languages, the teachers of the participating students acted as test translators. These teachers were bilingual, native

speakers of the target native languages.

Initially, our efforts focused on side-by-side comparisons of responses given by the same students to each item in both languages. However, it soon became clear that this approach alone could not render a complete picture of how ELL student performance varies across languages. The quality of the students' responses was inconsistent across both items and languages. Some students performed better in their native language than in English for some items but better in English than in their native language for other items.

We examined this score variability more closely from the perspective of generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), a psychometric theory of measurement error. Accordingly, we treated language, item, and rater as *facets* (i.e., sources of measurement error) and computed the estimated variance components (statistical

Table 2
Main Types of Comparisons Used in the Testing of ELLs

Type I Across linguistic groups, within test language	Type II Across linguistic groups, across test languages	Type III Within lingustic group, within test language	Type IV Within linguistic group, across test languages
ELLs versus non-ELLs; both groups tested in English	ELLs versus non-ELLs; each group tested in its own native language	(a) ELLs versus ELLs; both groups tested in English	(a) ELLs tested in English versus ELLs tested in their native language
		(b) Non-Ells versus non-ELLs; both groups tested in English	(b) ELLs tested in both English and their native language

estimates of the magnitude of measurement error) due to the main and interaction effects of student, item, language, and rater. This analysis allows for examining the percentage of score variation produced by each of these sources. It also allows for examining how the patterns of score variation due to these sources are similar or different across linguistic groups. For example, Table 3 compares the patterns of score variation observed in two groups of students, native Mandarin and native Spanish speakers. (The source [silr,e] is the residual—the interaction effect of student, item, language, rater, and error due to other, unknown sources of score variation.)

Consistent with results from research on science assessment (e.g., Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Pine, 1992; Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999), we found that rater (r) was not an important source of

measurement error. In contrast, the magnitude of the interaction of student and item (si) varied considerably across groups (4 and 24%, respectively, for the Mandarin and the Spanish native speakers). A large score variation due to this interaction means that a student who performed well on one item did not necessarily perform well on another. The implication of this is that to make valid generalizations about students' knowledge in a given knowledge domain, they must be given a considerable number of items (see Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993; Shavelson et al., 1992).

Examining the interaction of language and rater (lr) addresses the concern that raters may misinterpret ELL students' responses because of a cultural knowledge gap (see Eriks-Brophy & Crago,

Table 3

Rounded Percentage of Score Variation Due to the Main and Interaction Effects of Student, Item, Language, and Rater on Three Science Items

	Native language	
Source	Mandarin (n = 26)	Spanish (n = 18)
student (s)	16	6
item (i)	25	29
language (l)	0	0
rater (r)	0	0
si	4	24
sl	0	0
sr	0	0
il	0	0
ir	0	1
lr	0	1
sil	42	13
sir	1	5
slr	2	2
ilr	1	1
silr,e	9	17
Total	100	99*
ρ2	.62263	.32828
φ	.46906	.21253

* Percentage does not add up to 100 due to rounding. e = error

1993). A large score variation due to this interaction should be observed when raters are incapable of scoring the student responses reliably (and fairly) in both languages. In our study, in which the raters were familiar with the characteristics of the students, the score variation due to this interaction was small for the three groups (0 to 1%), which indicates that the raters' scoring was not systematically higher in either language.

Whereas the main effect of language is not an important source of measurement error (0% for both groups), the interaction of this facet with student and item (sil) deserves special consideration. A large score variation due to this interaction indicates that some ELLs perform better on some items in one language and better on other items in the other language. This score variation is considerable for the native Mandarin speakers (42%) and moderate but still important (13%) for the native Spanish speakers. These results show that, in addition to content knowledge, each ELL student has a unique set of weaknesses and strengths in English and a unique set of weaknesses and strengths in his or her native language. Also, in addition to its intrinsic cognitive demands and the academic knowledge it taps, each item poses a different set of linguistic challenges depending on the language in which it is administered. ELL student performance varies considerably not only across items but also across languages.

The groups differed in the magnitude of their coefficients ρ^2 (G theory's equivalent to classical test theory's reliability coefficient) and φ (dependability of absolute decisions—a domain-reference reliability). This indicates that the number of items needed to produce dependable scores is different for each group—a fact that could not be detected with conventional comparisons used in ELL testing. Additional analyses (not reported here) performed with a larger sample of Mandarin speakers also revealed considerably different patterns of score variation between groups of students within the same broad linguistic group.

Although the sample sizes in this study were small, the results show the potential of this approach, in which validity is addressed by examining the dependability of scores obtained for a given group of ELLs, not by comparing their performance to the performance of non-ELLs. Methods intended to ensure test validity in ELL testing are mainly based on item response theory (IRT) (see Lord & Novick, 1968; Hambleton & Swaminathan, 1985), which allows for detecting and controlling for language or cultural bias by means of differential item functioning (e.g., Camilli & Shepard, 1994; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 1998). However, IRT is a theory of scaling item scores, not a theory of measurement error (Shavelson & Ruiz-Primo, 2000).

In contrast, G theory addresses the fact that student scores are due to multiple factors and their interactions. In the context of ELL testing, these factors include the students themselves, the items, the language used in administering the items, and the raters who score the responses. G theory allows for determining how much information about ELL students' knowledge is gained and missed when they are tested in English and when they are tested in their native languages.

New approaches—not used by any assessment system yet— based on this view of language as a source of measurement error need to be investigated. One possible approach consists of giving ELLs the same items in both English and their native languages.

Unlike certain accommodations in which the same items are administered simultaneously in two languages side by side (e.g., Anderson, Liu, Swierzbin, Thurlow, & Belinski, 2000; Solano-Flores, Ruiz-Primo, Baxter, & Shavelson, 1992), in this approach the two language versions of the same item would be administered at different times and their scores would be treated as different. This approach differs substantially from what is deemed acceptable according to current paradigms in ELL testing because it

requires testing ELLs with more items than those given to non-ELLs.

Concluding Remarks

To date, efforts to address linguistic diversity have been less effective than researchers have hoped. These efforts can be characterized as attempts to eliminate the effects of non-mainstream language and non-mainstream culture as a way to ensure test validity. Under new paradigms in the testing of ELLs, test development efforts should be oriented in the opposite direction. Culture-free tests cannot be constructed because tests are inevitably cultural devices. Therefore, understandings of non-mainstream language and non-mainstream culture must be incorporated as part of the reasoning that guides the entire assessment process. Whereas test use and test interpretation are considered as the two basic aspects of test validity (see Shavelson & Towne, 2002), under new research and practice paradigms test development and test review would be recognized as equally important.

Closer attention to the contextual factors that shape student performance is critical to the emergence of these new paradigms. Students' responses to the Lunch Money problem (an item judged psychometrically sound by NAEP) should make us skeptical of the ability of existing methods to certify items as appropriate for a culturally and linguistically diverse population. Under the existing paradigms of ELL testing, in the best case that item would either be eliminated or "corrected" with the intent of avoiding the sensitive issue of poverty by changing the story of Sam's receiving money from his mother. In contrast, under new paradigms, that item would be adapted to each specific context. It is possible (why not?) that the item could even capitalize on students' sensitivity to issues related to survival. The problem involving addition and rounding to the next higher monetary unit could be set in a story that emphasizes the fact that mathematical reasoning can be used strategically to sort out tough situations.

We have discussed three areas in which the shift toward new paradigms in the testing of ELLs should occur: test review, test development, and treatment of language as a source of measurement error. They are not the only areas in which important transformations should occur but are probably the most important. In discussing them and providing examples of alternative new approaches, we have made an effort to show that multiple disciplines must contribute to the reasonings that guide testing decisions. Psychometrics is just one of those disciplines.

Of course, one important challenge to a paradigm shift in the testing of ELLs is possible resistance from the measurement community. Our experience with item microanalysis and with the concurrent assessment development model indicates that attaining equity and validity in the testing of linguistic minorities may require the use of items that, though intended to measure the same construct, differ considerably in wording, syntactic structure, and even discourse structure. Moreover, our experience using Type IV-b comparisons (see Table 3) shows that the unique condition that results from being bilingual can be better addressed when research designs incorporate language as a source of measurement error—a facet that is specific to ELLs. Attaining more equitable and valid measures of ELL academic achievement may require using different numbers of items with different segments of the population (or even sub-groups within

the same broad linguistic group).

These approaches look unacceptable because they seem to violate the basic principle of standardization. But let us be a little cynical: Many testing practices currently accepted violate this principle anyway (see Kopriva, 1999). Examples are computer adaptive testing, in which students take different sets of items, depending on their performance, and the accommodations for ELLs, which imply a differential treatment of groups of students. Even testing across languages (e.g., in international comparisons) is based on administering tests that differ radically in language.

The most important challenge, however, has to do with the implementation, not the popularity of the proposed ideas. As in any reform in education (see Greeno, Collins, & Resnick, 1996), shifting to new paradigms in the testing of ELLs, as we envision them, will require a systemic transformation. For example, in order to properly perform item microanalysis, adapt the wording of items according to specific local contexts, and identify dialect and language proficiency differences within broad linguistic groups, local educators should be engaged in developing assessments for their own communities of students and participate directly in the review, piloting, and adaptation of items. These tasks should be carried out in accordance with item content specifications and in alignment with state and national standards. Such a system differs considerably from a centralized system in which a relatively small team of test developers develops the items that are given to students across an entire state or across the entire country. It assumes a coordinated effort of professional development programs and the support of different kinds of professionals, from assessment development experts to cognitive scientists to linguists to cultural anthropologists.

A paradigm shift in the testing of ELLs is both necessary and possible. Assessment is a multidisciplinary endeavor. Factors that at first glance seem impractical, if not impossible, to address are shown to be addressable through methods that have been tried in the field. While these methods need refining and streamlining in order to be feasible for large scale efforts, we cannot, as a society, afford to overlook them if they hold promise for greater validity and equity in testing and accountability systems.

Appendix. Microanalysis of the Lunch Money Item: Some Findings.

Formal Properties

1. 1. The noun phrase “one dollar” as an adjective modifying “bill” may cause problems for students. It requires recognizing that one structural category is functioning as another.
2. 2. The use of “least” as “the smallest or lowest in importance” is somewhat archaic. It does not often apply to nouns, such as “number [of dollars],” but modifies adjectives, as in “least expensive.”

1. 3. The first two sentences may naturally lead the student to expect that the next sentence will tell how much money Sam has or how much his mother has given him.
2. 4. The last sentence is complex. It contains a main independent clause in the form of a question (“What is the least number”) with a prepositional phrase (“of $1.00 bills”) followed by a dependent clause that modifies the first clause (“that his mother

should give him"), a subordinate (dependent) clause ("so he will have enough money") followed by an infinitive phrase ("to buy lunch") with a prepositional phrase at the end ("for 5 days").

Empirical Properties

1. 1. *Purchase* was unfamiliar to many students, although many are able to figure out its meaning from reading the item.
2. 2. Although students may know what *enough* means, they sometimes had trouble reading it.
3. 3. *Least* is not difficult to read, but often students did not seem to understand its meaning in the context of the item (e.g., some students read it as *last*).
4. 4. Many students misinterpreted the word *only* in the third sentence ("His mother has only $1.00 bills") as restricting the number of dollar bills or the number of dollars, rather than the number of dollar denominations.

Differential Properties

1. 1. Low-income students were more likely than high-income students to misinterpret *only* in the third sentence. In reading the item aloud, some of them also inserted words that did not appear in the sentence and that modified the story (e.g., "His mother has given him only one dollar").
2. 2. The interpretation of *only* as restricting the number of dollars ("His mother has only one dollar") was observed only in low-income students.
3. 3. Some low-income students solved the item by computing the total cost of the lunch and then eliminating lunch items to find out what can be bought with only one dollar.
4. 4. When asked what the item is about, low-income students gave answers such as, " it's about Sam, trying to get her lunch, but her mom only has one dollar, and she needs more for five days, so I think she should give her a dollar ninety-five."

NOTE

The writing of this article is supported by the National Science Foundation. We are grateful to Ursula Sexton, Min Li, Rebeca Díaz, Jo Ann Izu, and Rachel Lagunoff for their participation in the data collection or analysis of one of the projects and to Julia Lara and Sharon Nelson-Barber for their important collegial support. We are grateful to Evelyn Jacob,
C. Stephen White, and three anonymous reviewers for their insightful comments. The opinions here expressed are not necessarily those of the funding agency or our colleagues.

REFERENCES

Abedi, J., Lord, C., Boscardin, C. K., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of limited English proficient (LEP) students in the National Assessment of Educational Progress (NAEP)*
(CSE Technical Rep. 537). Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education

& Information Studies, University of California, Los Angeles.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students; NAEP math performance.* Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2001). 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 Technical Rep. 462). Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Belinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2* (Minnesota Rep. No. 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 10, 2002, from http://education.unm.edu/NCEO/OnlinePubs/MnReport31.html

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

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133–140.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1–17.

Bialystok, E. (1997). Effects of bilingualism and biliteracy on children's emerging concepts of print. *Developmental Psychology*, *3*(3), 429–440.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge, UK: Cambridge University Press.

Bruner, J. (1993). Do we "acquire" culture or vice versa? *Behavioral and Brain Sciences, 16*, 515–516.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Technical Rep. 448). Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education

& Information Studies, University of California, Los Angeles.
Calfee, R. C., & Berliner, D. C. (1996). Introduction to a dynamic and relevant educational psychology. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 1–11). New York: Simon & Schuster Macmillan.
Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage.
Casanova, U., & Arias, B. (1993). Contextualizing bilingual education. In M. B. Arias & U. Casanova (Eds.), *Bilingual education: politics, practice, and research: Ninety-second yearbook of the National Society for the Study of Education, Part II*. Chicago: University of Chicago Press.
Cocking, R. R., & Mestre, J. P. (1988). *Linguistic and cultural influences on learning mathematics*. Hillsdale, NJ: Lawrence Earlbaum Associates.
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: The MIT 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.
Diaz, S., Moll, L. C., & Mehan, H. (1986). Sociocultural resources in instruction: A context-specific approach. In *Beyond language: Social and cultural factors in schooling language minority students* (pp. 187–230). Bilingual Education Office, California State Department of Education: Los Angeles: Evaluation, Dissemination, and Assessment Center, California State University.
Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.
Durán, R. P. (1989). Testing of linguistic minorities. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). New York: American Council of Education, Macmillan.
Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543–553.
Eriks-Brophy, A., & Crago, M. (1993, April). *Transforming classroom discourse: Forms of evaluation in Inuit IR and Ire routines*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.
Estrin, E. T., & Nelson-Barber, S. (1995). *Issues in cross-cultural assessment: American Indian and Alaska Native students*. San Francisco: Far West Laboratory.
Frederiksen, N. (1990). Introduction. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. ix–xvii). Hillsdale, NJ: Lawrence Erlbaum Associates.
Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304–312.
Genesee, F. (Ed.). (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*(10), 1115–1124.
Greenfield, P. M. (2000). Three approaches to the psychology of culture: Where do they come from? Where can they go? *Asian Journal of Social Psychology, 3*, 223–240.
Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York: Simon & Schuster Macmillan.
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., Ferdman, B. M., & Diaz, R. M. (1987). Bilingualism and cognitive development: Three perspectives. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics volume II: Reading, writing and language learning* (pp. 284–319). Cambridge, UK: Cambridge University Press.
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.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229–244.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181–200.
Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge, UK: Cambridge University Press.
Heath, S. B. (1986). Sociocultural contexts of language development. In *Beyond language: Social and cultural factors in schooling language minority students* (pp. 143–186). Bilingual Education Office, California State Department of Education: Los Angeles: Evaluation, Dissemination, and Assessment Center, California State University, Los Angeles.
Jacob, E., Johnson, B. K., Finley, J., Gurski, J., & Lavine, R. (1996). One student at a time. The cultural inquiry process. *Middle School Journal, 27*(4), 29–35.
Kaplan, R. (1988). Contrastive rhetoric and second language learning: Notes toward a theory of contrastive rhetoric. In A. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 275–304). Newbury Park, CA: Sage.
Kindler, A. L. (2002). *Survey of the states, limited English proficient students and available educational programs and services: 1999-2000 summary report*. Washington, DC: The George Washington University and National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.
Kochman, T. (1989). Black and white cultural styles in pluralistic perspective. In B. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 259–296). Boston: Kluwer Academic.
Kopriva, R. (1999). *A conceptual framework for the valid and comparable measurement of all students*. Washington, DC: Council of Chief State School Officers.
Kopriva, R., & Saez, S. (1997). *Guide to scoring LEP student responses to open-ended mathematics items*. Washington, DC: Council of Chief State School Officers.

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

Kupermintz, H., Le, V. H., & Snow R. (1999). *Construct validity of mathematics achievement: Evidence from interview procedures*. (CSE Technical Rep. 493). University of California, Los Angeles, Center for the Study of Evaluation.

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*(1), 55–75.

Lagunoff, R., Solano-Flores, G., Sexton, U. N., & Nelson-Barber, S. (2001, February). *English language ability and math and science assessments*. Paper presented at the annual meeting of California Association for Bilingual Education, Los Angeles.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English language backgrounds. *Educational Researcher, 27*(4), 12–21.

Lee, O. (1999). Equity implications based on the conceptions of science achievement in major reform documents. *Review of Educational Research, 69*(1), 83–115.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Messik, S. (1995b). Validity of psychological assessments. Validation of inferences from person's responses and performances as scientific inquiry into scoring meaning. *American Psychologist*, *50*, 741–749.

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

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social, class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: RAND.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the na-tion's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Philips, S. U. (1983). *The invisible culture: Communication in classroom and community on the Warm Springs Indian Reservation*. New York: Longman.

Resnick, L. B. (1991). Shared cognition: Thinking as social practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 1–20). Washington, DC: American Psychological Association.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). Statewide

assessment programs: Policies and practices for the inclusion of limited English proficient students. *Practical Assessment, Research & Evaluation, 5*(13).

Rogoff, B. (1990). *Apprenticeship in thinking*. New York: Oxford University Press.

Ruiz-Primo, M. A. (2002, February). On a seamless assessment system. In *Seamless Science Education.* Symposium conducted at the annual meeting of the American Association for the Advancement of Science, Boston.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41–53.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E., (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99–141.

Sanchez, G. I. (1934). Bilingualism and mental measures: A word of caution. *Journal of Applied Psychology, 18*, 756–772.

Saxe, G. B. (1991). *Culture and cognitive development: Studies in mathematical understanding*. Hillsdale, NJ: Erlbaum.

Sexton, U., & Solano-Flores, G. (2001). *Cultural validity in assessment development: A cross-cultural study on the interpretation of math and science items*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*(9), 692–697.

Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. Mintzes, J. Wandersee, & J. Novak (Eds.), *Assessing science understanding* (pp. 303–341). San Diego, CA: Academic Press.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

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

Shaw, J. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching, 34*(7), 721–743.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English proficient students in Rhode Island's Grade 4 mathematics performance assessment* (CSE Technical Rep. 486). Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennet & W. C. Ward (Eds.). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp.

263–331). New York: American Council of Education, Macmillan.
Solano-Flores, G. (2002, April). *Cultural validity: A sociocultural perspective in educational measurement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.
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., & Sexton, U. (2002). *Assessing the impact of poverty on student performance in a mathematics constructed-response item*. Manuscript in preparation.
Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.
Solano-Flores, G., Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1992). *Science performance assessments: Use with language minority students*. Unpublished manuscript, University of California, Santa Barbara.
Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16–25.
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*(2), 107–129.
Steele, C. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629.
Steele, C. M., & Aronson, J. (1995). Stereotpye threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.
Stevens, R. A., Butler, F. A. & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English Language Learners (ELLs)* (CSE Technical Rep. 552). National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.
Swisher, K., & Deyhle, D. (1992). Adapting instruction to culture. In J. Reyhner (Ed.), *Teaching American Indian students* (pp. 81–95). Norman: University of Oklahoma Press.
Tanzer, N. K. (in press). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence Erlbaum.
Trumbull, E., Diaz-Meza, R., & Hasan, A. (2000, April). *Using cultural knowledge to inform literacy practices: Teacher innovations from the Bridging Cultures Project*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
Tsang, C. L. (1989). Informal assessment of Asian Americans: A cultural and linguistic

mismatch? In B. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 231–254). Boston: Kluwer Academic.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex. Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, &

J. Pandey (Eds.), *Handbook of cross-cultural psychology: Theory and method* (Vol. 1, 2nd ed.). Needham Heights, MA: Allyn & Bacon.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29–37.

Van de Vijver, F. J. R., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263–279.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press. Washington State Commission on Student Learning. (1999). *Essential academic learning requirements*. Olympia, WA: Author. Wertsch, J. V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Wise, L. K., Hauser, R. M., Mitchell, K. J., & Feuer, M. J. (1999). *Evaluation of the voluntary national tests: Phase 1*. Washington, DC: National Academy Press.