

Part I

Cultural Validity in Assessment—Basic Concepts






Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. Basterra, In M. R., Trumbull, E., & Solano-Flores, G. (Eds.). *Cultural validity in assessment: A guide for educators* (pp. 3-21). New York: Routledge.

Chapter 1

Assessing the Cultural Validity of Assessment Practices

An Introduction

Guillermo Solano-Flores

Defining Cultural Validity

As with any other product of human activity, tests are cultural artifacts. They are a part of a complex set of culturally established instructional and accountability practices; they are created with the intent to meet certain social needs or to comply with the mandates and legislation established in a society; they are written in the language (and the dialect of that language) used by those who develop them; their content is a reflection of the skills, competencies, forms of knowledge, and communication styles valued by a society—or the influential groups of that society; and they assume among test-takers full familiarity with the contexts used to frame problems, the ways in which questions are worded, and the expected ways to answer those questions.

Viewing tests as cultural artifacts (see Cole, 1999) enables us to appreciate that, to a great extent, the ways in which students interpret test items and respond to them are mediated by cultural factors that do not have to do necessarily with the knowledge assessed. This is a matter of validity—scores obtained by students on a test should not be due to factors other than those that the test is intended to measure (Messick, 1995). This is as true of classroom assessments as it is of large-scale assessments.

While key normative documents on testing (e.g., AERA, NCME, & APA, 1999; Hambleton, 2005) recognize the importance of factors related to culture and language as a source of measurement error, current testing practices address culture as a threat to validity rather than the essence of validity. Culture is part of the discourse on test validity but is not viewed as the essence of a form of validity in its own right.

In 2001, we (Solano-Flores & Nelson-Barber, 2001, p. 555) proposed the concept of cultural validity, which can be defined as:

> the effectiveness with which […] assessment addresses the socio-cultural influences that shape student thinking and the ways in which students make sense of […] items and respond to them. These socio-cultural influences include the sets of values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, and the socioeconomic conditions prevailing in their cultural groups.

Along with this definition, we contended that the cultural factors that shape the process of thinking in test-taking are so complex that culture should not be treated as a factor to correct or control for, but as a phenomenon intrinsic to tests and testing. We argued that both test developers and test users should examine cultural validity with the same level of rigor and attention they use when they examine other forms of validity.

The notion of cultural validity in assessment is consistent with the concept of multicultural validity (Kirkhart, 1995) in the context of program evaluation, which recognizes that cultural factors shape the sensitivity of evaluation instruments and the validity of the conclusions on program effectiveness. It is also consistent with a large body of literature that emphasizes the importance of examining instruction and assessment from a cultural perspective (e.g., Ladson-Billings, 1995; Miller & Stigler, 1987; Roseberry, Warren, & Conant, 1992). Thus, although cultural validity is discussed in this chapter primarily in terms of large-scale assessment, it is applicable to classroom assessment as well. This fact will become more evident as the reader proceeds through the book.

In spite of its conceptual clarity, translating the notion of cultural validity into fair assessment practices is a formidable endeavor whose success is limited by two major challenges. The first challenge stems from the fact that the concept of culture is complex and lends itself to multiple interpretations—each person has their own conception of culture yet the term is used as though the concept were understood by everybody the same way.

As a result of this complexity, it is difficult to point at the specific actions that should be taken to properly address culture. For example, the notion of "cultural responsiveness" or "cultural sensitivity" is often invoked by advocates as critical to attaining fairness (e.g., Gay, 2000; Hood, Hopson, & Frierson, 2005; Tillman, 2002). However, available definitions of cultural sensitivity cannot be readily operationalized into observable characteristics of tests or their process of development.

The second challenge has to do with implementation. Test developers take different sorts of actions intended to address different aspects of cultural and linguistic diversity. Indeed, in these days, it is virtually impossible to imagine a test that has not gone through some kind of internal or external scrutiny intended to address potential cultural or linguistic bias at some point of its development. Yet it is extremely difficult to determine when some of those actions are effective and when they simply address superficial aspects of culture and language or underestimate their complexities. For example, the inclusion of a cultural sensitivity review stage performed by individuals from different ethnic backgrounds is part of current standard practice in the process of test development. While necessary, this strategy may be far from sufficient to properly address cultural issues. There is evidence that teachers of color are more aware than white, mainstream teachers of the potential challenges that test items may pose to students of color; however, in the absence of appropriate training, teachers of color are not any better than white teachers in their effectiveness in identifying and addressing specific challenges posed by test items regarding culture and language (Nguyen-Le, 2010; Solano-Flores & Gustafson, in press).

These challenges underscore the need for approaches that allow critical examination of assessment practices from a cultural perspective. While assessment systems and test developers may be genuinely convinced that they take the actions needed to properly address linguistic and cultural diversity, certain principles derived from the notion of cultural validity should allow educators and decision-makers to identify limitations in practices regarding culture and language and ways in which these practices can be improved.

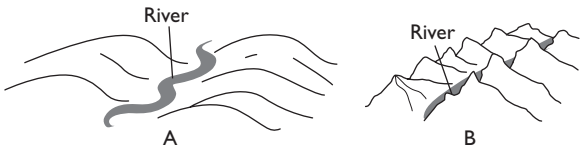
This chapter intends to provide educators, decision-makers, school districts, and state departments of education with reasonings that should enable them to answer the question, "What should I look for in tests or testing programs to know if appropriate actions have been taken to address culture?" These reasonings are organized according to four aspects of cultural validity: theoretical foundations, population sampling, item views, and test review.

In discussing these aspects, I share lessons learned and provide examples from three projects funded by the National Science Foundation (for a discussion of the methodological aspects of these projects, see Solano-Flores & Li, 2006, 2008, 2009; Solano-Flores & Trumbull, 2008). The first project, "Assessing the Cultural Validity of Science and Mathematics Assessments," investigated cultural influences on test-taking. Grade 4 students from 13 cultural groups (each defined by a unique combination of such factors as ethnicity, geographical region, ancestry, socioeconomic status, and linguistic influences) verbalized their thinking as they responded to the NAEP items shown in Figure 1.1 (see National Assessment of Educational Progress, 1996) and responded to interview questions on the ways in which they interpreted them.

The second project, "Cognitive, Sociolinguistic, and Psychometric Perspectives in Science and Mathematics Assessment for English Language Learners," examined how scores obtained by English language learners vary when they are tested in English or in their native language or in the local or standard dialects of those languages. The third project, "Teacher-Adapted Versus Linguistically Simplified Items in the Testing of English Language Learners" investigated the advantages of using language adaptations made by teachers on tests as a form of testing accommodation English language learners (ELLs) tested in English. These two projects addressed language from a sociolinguistic perspective that takes into consideration the social aspect of language and the fact that language use varies across social groups. More specifically, these projects examined the extent to which the scores of ELL students in tests vary due to language and dialect variation (Solano-Flores, 2006). As a part of the activities for these two projects, we worked with teachers from different linguistic communities with the purpose of adapting tests so that their linguistic features reflected the characteristics of the language used by their students.

For discussion purposes, throughout this chapter, language is regarded as part of culture. However, when appropriate, language or linguistic groups are referred to separately when the topics discussed target language as a specific aspect of culture. Also, the terms "assessment" and "test" are used interchangeably.

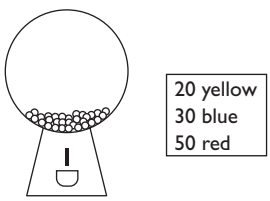
Mountains item:
The pictures below show the same river and mountains, but one picture shows how they looked millions of years ago, and the other picture shows how they look now. Circle the letter under the picture that shows how the river and mountains look now. Explain how you can tell this.



Lunch Money item:
Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that cost 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?

Metals item:
Many things are made of metal, such as pots, pans, tools, and wire. Give two reasons why metals are used to make many different things.

Gumball Machine item:
Think carefully about the following question. Write a complete answer. You may use drawings, words, and numbers to explain your answer. Be sure to show all of your work.



The gum ball machine has 100 gum balls; 20 are yellow, 30 are blue, and 50 are red. The gum balls are well mixed inside the machine.
Jenny gets 10 gum balls from this machine.
What is your best prediction of the number that will be red?

Figure 1.1 Items used in the project (source: "Assessing the Cultural Validity of Science and Mathematics Assessments." National Assessment of Educational Progress (1996). *Mathematics Items Public Release*. Washington, DC: Author).

Theoretical Foundations

Testing practices should be supported by theories that address cognition, language, and culture. Regarding cognition, testing practices should address the fact that cognition is not an event that takes place in isolation within each person; rather, it is a phenomenon that takes place through social interaction. There is awareness that culture influences test-taking (Basterra, Chapter 4, this volume). Indeed, every educator has stories to tell on how wording or the contextual information provided by tests misleads some students in their interpretations of items. However, not much research has been done to examine the ways in which culture influences thinking during test-taking.

There is a well-established tradition of research on the cognitive validation of tests that examines students' cognitive activity elicited by items (Baxter, Elder, & Glaser, 1996; Hamilton, Nussbaum, & Snow, 1997; Megone, Cai, Silver, & Wang, 1994; Norris, 1990; Ruiz-Primo, Shavelson, Li, & Schultz, 2001), as inferred from their verbalizations during talk-aloud protocols in which they report their thinking while they are engaged in responding to items, or after they have responded to them (Ericsson & Simon, 1993). Surprisingly, with very few exceptions (e.g., Winter, Kopriva, Chen, & Emick, 2006), this research does not examine in detail the connection between thinking and culture or has been conducted mainly with mainstream, white, native English speaking students (see Pellegrino, Chudowski, & Glaser, 2001).

Regarding culture and language, testing practices should be in accord with current thinking from the culture and language sciences. Unfortunately, many actions taken with the intent to serve culturally and linguistically diverse populations in large-scale testing are insufficient or inappropriate. For example, many of the accommodations used by states to test ELLs do not have any theoretical defensibility, are inappropriate for ELLs because they are borrowed from the field of special education, or are unlikely to be properly implemented in large-scale testing contexts (Abedi, Hofstetter, & Lord, 2004; Rivera & Collum, 2006; Solano-Flores, 2008).

In attempting to develop approaches for ELL testing that are consistent with knowledge from the field of sociolinguistics, we (Solano-Flores & Li, 2006, 2008, 2009) have tested students with the same set of items in both English and their first language and in two dialects (standard and local) of the same language (varieties of the same language that are distinguishable from each other by features of pronunciation, grammar, and vocabulary, among others; see Crystal, 1997; Wolfram, Adger, & Christian, 1999). Rather than testing these students with bilingual formats, the intent is to determine the extent to which ELL students' performance varies across languages or dialects. Generalizability theory (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991, 2009)—a psychometric theory of measurement error—allows examination of language as a source of measurement error and, more specifically, the extent to which student performance varies across languages or dialects.

An important finding from our studies is that, contrary to what simple common sense would lead us to believe, ELLs do not necessarily perform better

if they are tested in their native languages. Rather, their performance tends to be unstable across languages and across items. Depending on the item, some students perform better in English than in their native language, and some other students perform better in their native language than in English. Our explanation of this finding is that each item poses a unique set of linguistic challenges in each language and, at the same time, each ELL has a unique set of strengths and weaknesses in each language.

Another important finding from those studies speaks to the complexity of language. If, instead of being tested across languages, students are tested across dialects of the same language (say, the variety of Spanish used in their own community and the standard Spanish used by a professional test-translation company), it can be observed that their performance across dialects is as unstable as their performance across languages.

These findings underscore the fact that no simple solution exists if we are serious about developing valid and fair assessment for linguistic minorities (Solano-Flores & Trumbull, 2008). Testing ELLs only in English cannot render valid measures of achievement due to the considerable effect of language proficiency as a construct-irrelevant factor. The same can be said about testing ELLs in their native language. Valid, fair testing for ELLs appears to be possible only if language variation due to dialect is taken into consideration.

Our findings also speak to the fact that, even within broad linguistic groups (e.g., native Haitian-Creole speakers or native Spanish speakers), every group of ELLs is unique as to the sensitivity to the language or dialect in which it is tested. We have observed that the minimum number of items needed to obtain dependable scores may vary with dialect (i.e., more items are needed to obtain dependable scores if students are tested in one dialect than in another). Also, we have observed that groups of students within the same group (e.g., ELLs, native Spanish speakers who live in different regions in the United States), may vary considerably on the number of items needed to obtain dependable measures of their achievement.

Notice that the studies described were the first to use generalizability theory with ELLs. Appreciating the possibilities of using this theory in the testing of ELLs was possible because we were aware that linguistic variation is critical in the field of sociolinguistics. We reasoned that, because bilingual populations are heterogeneous and dynamic rather than homogenous and static, better ELL testing approaches could be developed by using this theory, since it allows examination of multiple sources of score variation. This experience illustrates the notion that testing practices should be in accord with theories on content and knowledge. As the findings from the studies discussed show, the instability of student performance across languages (or dialects) and items is consistent with the well-known notion in sociolinguistics that the use of a first language or a second language among bilingual individuals is shaped by context and content (Fishman, 1965).

Population Sampling

Statistically appropriate samples of students from different cultural and linguistic groups should participate at all stages of the process of test development. "Inclusion" is a term used these days to refer to the fact that ELL students are included in large-scale testing programs. However, inclusion itself does not ensure fairness or validity if it is limited to test administration and does not involve the entire process of test development. Sampling is key to ensuring cultural validity in testing practices.

Three aspects of sampling need to be discussed: population specification, population representation, and community participation. Population specification is critical to identifying the types and sizes of samples of culturally and linguistically diverse students that should be included in the process of test development. It refers to the ways in which cultural groups are defined and, therefore, the criteria used to determine when an individual belongs to a certain cultural group. A sufficient number of relevant attributes, such as ethnicity, first language, locale, socio-economic status, etc., should lead to proper population specification. Serious threats to the validity of tests for ELLs arise when cultural groups are defined in terms of irrelevant attributes or relevant but insufficient attributes (see Solano-Flores, 2009). Examples of population misspecification are: defining a cultural group based on race; inferring the proficiency of individuals in the predominant language based on their national origin; collapsing ELLs and special education students in the same category; and using one broad linguistic group of ELLs (e.g., those who are native Spanish speakers) as representative of all the broad linguistic groups of ELLs.

Population representation refers to the extent to which appropriate samples of diverse cultural groups are used in the process of testing. The samples of individuals from different cultural groups used in the process of testing (e.g., as pilot students) should reflect the cultural make-up of the target population (e.g., the population of fourth-grade students in a state) and the sizes of these samples should be consistent with their proportions in that population.

Unfortunately, culturally and linguistically diverse students are usually included only in the terminal stages of the process of test development, or not included at all. For example, it is not customary practice to include ELLs in the samples of pilot students who are asked about the ways in which they interpret draft versions of the items and how these items should be worded so that students understand them as intended. Test developers may wrongly believe these students do not have much to contribute as pilot students, due to their limited proficiency in English. However, a large segment of the ELL population has sufficient communicative skills in non-academic English. Indeed, there is evidence that, as many as two-thirds of ELL students chose to use English in the talk-aloud protocols and interviews conducted to determine how they interpret items; they communicate in English sufficiently well to allow the interviewer to obtain valuable information on their thinking and test-taking strategies (Prosser & Solano-Flores, 2010).

Community participation refers to the fact that decisions concerning the linguistic features of test items (e.g., their wording) should be sensitive to language

usage among the communities. Items can be thought of as samples of the features of the language (and the dialect of that language) in which they are written (Solano-Flores, 2006; Solano-Flores & Li, 2006; Solano-Flores & Trumbull, 2003). By ensuring that communities participate in the process of testing, we can ensure that the ways in which language is used in their communities are properly represented in tests.

Traditionally, a panel of experts makes decisions about the ways in which tests should be written or the ways in which their wording should be modified to ensure that ELLs gain access to the content of items. While useful, this approach may not suffice to ensure that language factors are properly controlled for, especially because, as discussed above, the performance of ELL students in tests may be extremely sensitive to the dialect of the language in which tests are administered (Solano-Flores & Li, 2006, 2008, 2009).

We (Solano-Flores, Li, Speroni, Rodriguez, Basterra, & Dovholuk, 2007; Solano-Flores, Speroni, & Sexton, 2005) have investigated the advantages of using a sociolinguistic approach in the linguistic modification of tests. In this approach, teachers modify the linguistic features of test items based on their knowledge of the characteristics of the language used in their own schools. This approach takes into consideration the fact that language use and language usage vary across social groups (Wardhaugh, 2002). According to this sociolinguistic perspective, to minimize language as a source of measurement error, the process of language adaptation must be sensitive to differences in which language is used by different communities. Critical to this approach is the notion of localization, which we use to refer to the process of adapting the linguistic features of test items to the local English dialects used by the students' communities. Frequently used in the context of translation, the notion is also applicable in the context of dialects within a language. It refers to the "process of adapting text and cultural content to specific target audiences in specific locations" (WorldLingo, 2004). Originating in the jargon of globalization economy, the concept of localization recognizes that every location is unique by virtue of a series of linguistic and cultural factors. Thus, efforts to adapt text to a given target group must go beyond the pure formal level (Muzzi, 2001).

To examine the possibilities and limitations of using teacher adaptation as an approach to facilitating ELLs gain access to the content of items, we (Solano-Flores et al., 2007) conducted a study that compared teacher adaptation and linguistic simplification as forms of testing accommodation for ELLs. We converted the original version of a test into two test versions, teacher-adapted and linguistically simplified. Using a design that controlled for the effects of sequence, we gave ELL students the same set of test items in two test version combinations, teacher-adapted and original version or teacher-adapted and linguistically simplified. The teacher-adapted version of the items was created by using the approach described above. The linguistically simplified version of the items was created by using linguistic simplification procedures similar to those used by Abedi and his associates (e.g., Abedi, Lord, Hofstetter, & Baker, 2001).

A comparison of the teacher-adapted and linguistically simplified versions revealed that, in terms of their psychometric properties, the two forms of modi-

fied tests (internal consistency reliability and mean score differences) were similar. Mean score differences between test versions were either not statistically significant, or their effect sizes were small. In addition, we observed a negligible score variability across test versions, which indicates that teacher adaptation allows modification of test items in ways that do not alter their technical properties. However, we found that the two forms of accommodation are effective for different reasons. The teacher-adapted approach allowed specialists to focus more on the functional aspects of language; in contrast, the linguistically simplified approach allowed specialists to focus more on formal aspects of language.

We concluded that both the teacher adaptation and linguistic simplification approaches are effective in minimizing proficiency in the language of testing as a form of testing accommodation for ELLs. These results speak to the advantages of including teachers from the communities of the ELL populations in the process of ELL testing. Teacher adaptation-based approaches can be successfully used in combination with or as an alternative to linguistic simplification as a form of accommodation in the testing of ELL students.

Item Views

As part of the process of test development, the views that students from different cultural groups have of items should be carefully examined. Item views can be thought as ways in which students tend to make sense of items and which are influenced by their cultural experience (Solano-Flores, 2001). The notion of item views is an extension of the notion of worldviews—culturally determined ways of making sense of experience (see Lee, 1999).

To examine the item views of different cultural groups, we (Solano-Flores, 2001; Li, Solano-Flores, Kwon, & Tsai, 2008) asked students to examine one of the four items shown in Figure 1.1. Then we asked them the following two questions intended to probe their item views: "What is this item about?" and "What do you have to know or be able to do to answer it?"

A conceptual framework for examining the observed students' responses to these questions is shown in Figure 1.2. According to this conceptual framework, students' views of test items can be characterized along two dimensions and four resulting quadrants. The dimension, content–context (vertical axis) refers to whether a student's view of the item is based on either the content addressed by the item or the contextual information provided by it. For example, when asked what the Gumball Machine item is about, some students identify a broad knowledge domain (e.g., *the item is about math*) whereas others focus on the characters and situations included in the items with the intent to make them meaningful (e.g., *the item is about a gumball machine*).

The dimension, specificity–generality (horizontal axis) refers to the level of specificity with which students think about the skills or knowledge assessed by an item. For example, when asked what they have to know or what they are able to do to answer the Gumball Machine item correctly, some students relate the item to a specific topic (e.g., *you need to know about probability*) whereas others invoke general skills they think are critical to responding to the item (e.g., *you need to be smart*).

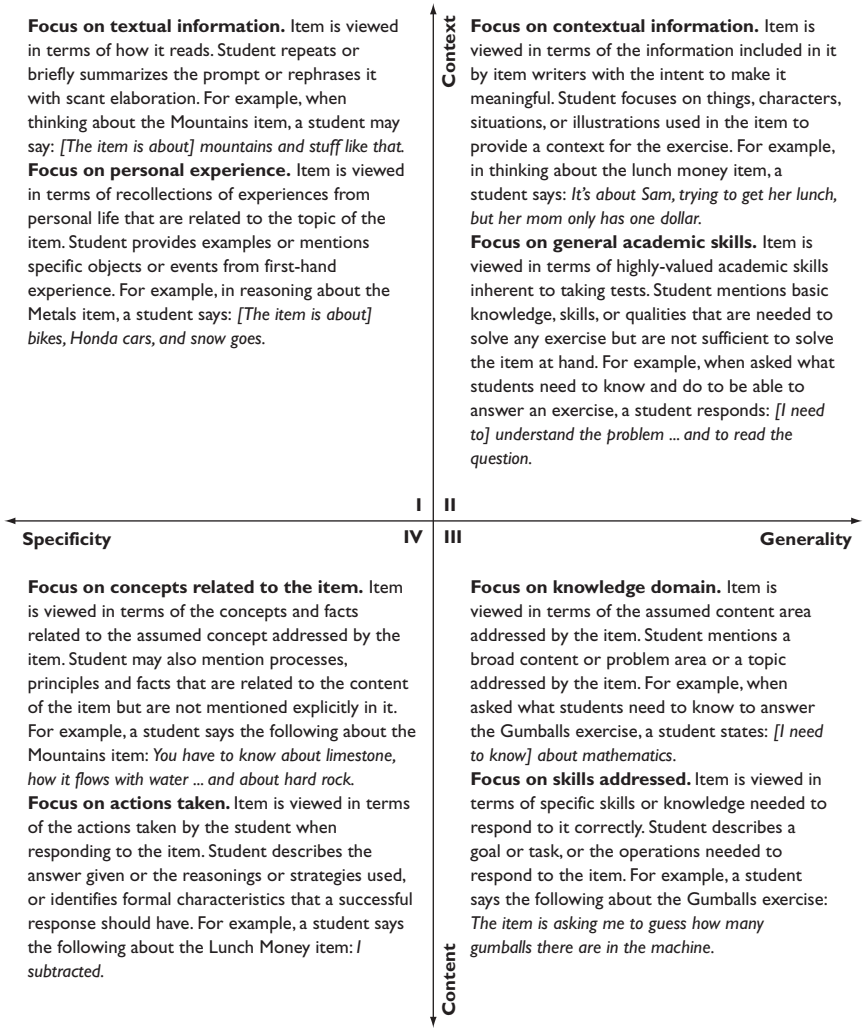


Figure 1.2 Item Views Quadrants that Result from the Combination of Two Dimensions of Student Item Views: Specificity–Generality and Content–Context. I. Context-Oriented, Specific; II. Context-Oriented, General; III. Content-Oriented, General; IV. Content-Oriented, Specific (source: "Assessing the Cultural Validity of Science and Mathematics Assessments." National Assessment of Educational Progress (1996). *Mathematics Items Public Release*. Washington, DC: Author).

Figure 1.3 shows the percentage of students whose responses were coded as belonging to each of the four quadrants for a sample of cultural groups (Solano-Flores, 2001). For example, 53% of the responses given by Painas[1] students to the two questions above were coded as belonging to Quadrant 1 (Specific, Context-Oriented).

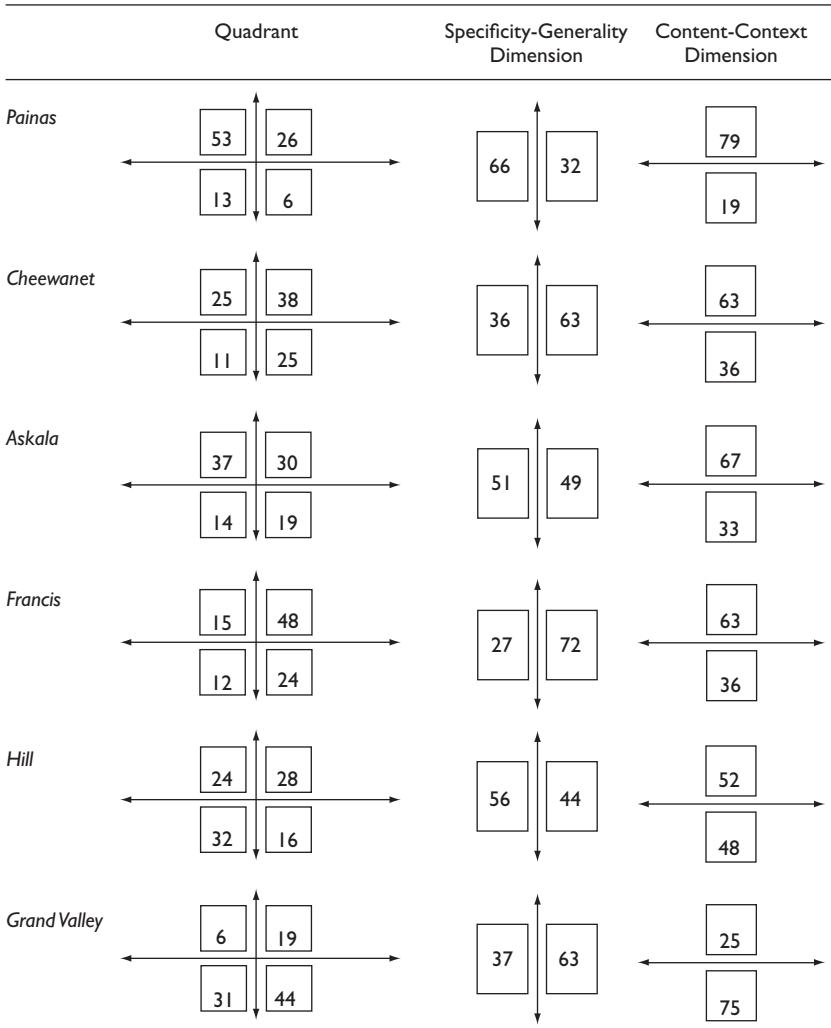


Figure 1.3 Test Views Quadrant Case Frequencies in Percentages for Six Cultural Groups (Percentages Do Not Add Up to 100 in All Cases Due to Rounding) (source: "Assessing the Cultural Validity of Science and Mathematics Assessments." National Assessment of Educational Progress (1996). *Mathematics Items Public Release*. Washington, DC: Author).

In addition to the percentages of responses coded by quadrant, the table shows the percentages of responses coded by dimension. Two facts stand out. First, each group seems to have either predominantly specific or predominantly generic item views. The only exception is Askala (51% and 49% respectively for the specific and generic components of the dimension). Second, as the third column in Figure 1.3 shows, the item views of students from Painas, Cheewanet, Askala, and Francis are predominantly context-oriented; the percentages of

responses coded on the context component of the dimension for these groups range from 63% to 79%. In contrast, the students from Hill have comparable percentages of responses on content and context (48% and 52%, respectively), and the students from Grand Valley are predominantly content-oriented (with 75% and 25% of responses coded respectively on the content and context components of the dimension).

The students from Hill and Grand Valley can be characterized as white, high socio-economic status, suburban. Students from both cultural groups obtained the highest average scores on the four NAEP items used in this study. These facts suggest that there is an important link between academic performance and content-oriented reasonings. They also suggest that the items reflect and favor content-oriented, de-contextualized thinking.

Test Review

The review of test items with the purpose of examining their cultural and linguistic appropriateness should take into consideration multiple sources of information. Two aspects of test review are discussed here: the use of judgmental and empirical procedures for reviewing the cultural appropriateness of test items, and the use of alternatives of form of representation of data with the purpose of comparing multiple cultural groups on their interpretation of items.

Regarding the use of judgmental and empirical procedures, while teachers who belong to ethnic/cultural minorities are more aware of cultural issues in testing, they do not necessarily make more accurate judgments than their mainstream counterparts about the cultural appropriateness of test items. Rather, the two types of teacher tend to address only superficial features of the items when they are asked to propose ways of minimizing the likelihood for an item to be culturally or linguistically biased again their students (Nguyen-Le, 2010; Sexton & Solano-Flores, 2001).

To examine this issue more carefully, we have evaluated the challenges that a specific item may pose to students due to linguistic or cultural factors. We used the Lunch Money item (see Figure 1.1), whose correct solution involves addition, multiplication, and rounding. We used this item because we have evidence that several interpretation and reading errors observed for this item can be attributed to four kinds of linguistic features: vocabulary (meaning of individual words), semantics (meaning of words in a sentence), syntax (grammatical structure of sentences), and pragmatics (interpretation of words and sentences in context) (Solano-Flores & Trumbull, 2003). Also, the written responses and computations produced by some students who live in poverty suggest that they interpret the item as if they were asked, "What can Sam buy with $1.00?" For example, students wrote solutions such as, "He should buy only the sandwich" or "He can buy only the juice and the fruit" (Solano-Flores & Trumbull, 2003).

We examined how consistently four review approaches identified vocabulary, semantics, syntax, pragmatics, meaningfulness, and appropriateness issues as critical to addressing the linguistic and cultural appropriateness of the Lunch Money item. Two of these four review approaches were judgmental, the judg-

ment of a linguist and whether the issue had been identified by at least 20% of the teachers who reviewed the item. The other two approaches were empirical, whether the issue had been observed among at least 20% of the students who read the item aloud (and whom we interviewed about their interpretations of the item) and whether statistically significant differences were observed between groups on the issues identified by the other criteria.

The results showed that the item features that may have a negative impact on student performance due to linguistic or cultural bias are difficult to anticipate based solely on judgmental procedures. Teachers produced a review of the Lunch Money item that did not reflect entirely the difficulties their own students had in interpreting the item.

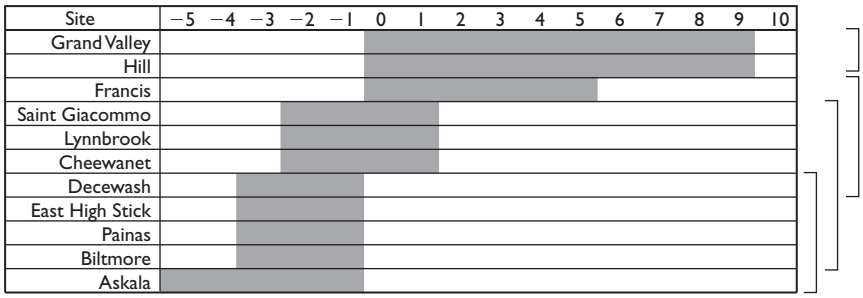
These results speak to the need for revising current test development practices, which depend heavily on teachers' judgment for test review. Teachers' judgments are necessary but not sufficient to identify subtle ways in which items may be biased due to language. Teachers' judgments should be used in combination with empirical review procedures that examine the psychometric properties of items.

Regarding the use of alternatives of form of representation of data, pattern charts (Solano-Flores, 2008b) illustrate ways in which different forms of information can be used in combination to examine cultural influences in test-taking. Pattern charts can be defined as devices for visually representing and linking information with different levels of statistical power on information obtained with different cultural groups.

Figure 1.4 shows an example of pattern charts. They display the relative standing of each cultural group with respect to the others in terms of the number and direction of statistically significant differences with other groups. The chart at the top of the figure shows group performance differences on the item, Gumball Machine; the chart below shows group differences on item meaningfulness, a concept advanced by Brenner (1998) when she examined how students relate school mathematics and everyday life. In our investigation, item meaningfulness was defined as the tendency of students to relate the content and/or context of an item to activities in which they are actors,[2] as reflected by the response to the question, "How do you see [the content of the item] as part of what you do when you are not at school?" This interview question was included with the intent to determine if students from different cultural backgrounds vary in their ability to relate the content of the item to activities that take place out of the school context in which they are actors.

The length of the bars with respect to the zero vertical line for a given cultural group indicates the number of instances that a statistically significant difference was observed between that group and another group on the construct measured. The orientation (to the left or to the right) of each bar indicates whether the statistically significant differences are in favor of or against the cultural group.[3] For example, in the pattern chart for the Gumball Machine item the length and the orientation of the bars for Grand Valley and the two mainstream, high socioeconomic status cultural groups that participated in the comparison indicate that each of these two groups had statistically significantly higher scores than nine of

Gumball Machine item score differences



Item meaningfulness score differences

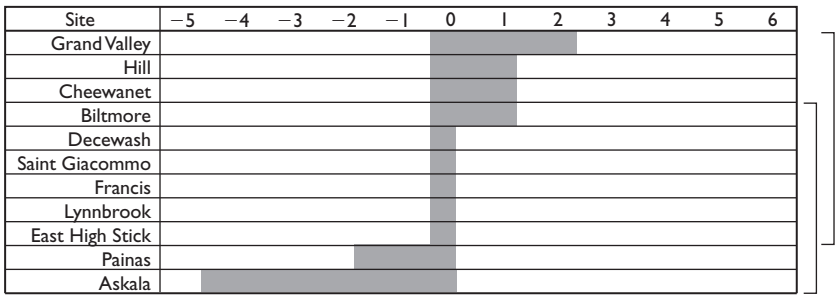


Figure 1.4 Number and Direction of Statistically Significant Item Score Differences with Other Groups: Score on the Gumball Machine and Item Meaningfulness on The Interview Question, "How Do You See [the Content of the Item] as Part of What You Do When You are Not at School?" Subsets for Alpha = 0.05. Tukey Post-Hoc Multiple Comparisons Indicated By Brackets (source: "Assessing the Cultural Validity of Science and Mathematics Assessments." National Assessment of Educational Progress (1996). *Mathematics Items Public Release*. Washington, DC: Author).

the eleven cultural groups included in the comparison. At the bottom of this chart is the group, Askala, which ranks the lowest; its performance on the item was statistically significantly lower than five of the other groups.

The brackets on the right of the charts show statistically significant score differences between sub-sets of groups. For example, the bracket on the right of the chart for Gumball Machine distinguishes one sub-set comprising two groups, Grand Valley and Hill, from the rest of the groups.

A comparison of the two pattern charts reveals similar rank orderings of the cultural groups. Overall, the cultural groups with higher item meaningfulness scores in the interview question tended to score high in the Gumball Machine item; the cultural groups with low item meaningfulness scores in that interview question tended to score low in the Gumball Machine item. The similarity of these patterns supports the notion that socio-cultural factors that take place

outside the formal instruction environment are a powerful influence that shapes performance on tests. Or, put in another way, the similarity of patterns suggests that the Gumball Machine item reflects contexts that are more familiar to mainstream, white, high socio-economic status students than students from any other cultural group.

Altogether, these results underscore the major role that informal, culturally determined experience plays as a factor that shapes student performance on test items. Also, they show the value of using alternative forms of visual representation of data and quantitative and qualitative information in combination to compare multiple cultural groups.

Summary and Conclusion

I have written this chapter with the intent to support test users (teachers, decision-makers, school districts, and state departments of education) in their reasonings about how effectively cultural validity is addressed in tests and testing practices. The chapter responds to the need for tools for objectively examining the extent to which tests and testing practices are sensitive to issues of language and culture. Four main aspects of cultural validity are discussed: theoretical foundations, population sampling, item views, and test review.

Throughout the chapter, I have shared lessons learned from National Science Foundation funded projects which have examined cultural influences in test-taking and the relationship between language variation and score variation. Findings from those projects illustrate ways in which culture can be addressed in testing practices.

As a summary, in examining tests and testing practices from the perspective of cultural validity, test users have four questions to ask:

1. To what extent are the testing practices consistent with current thinking in the culture and language sciences?
2. How accurately are culturally and linguistically diverse populations specified, and how properly are they represented throughout the entire process of test development?
3. To what extent does the process of test development take into consideration ways in which students from different cultural backgrounds interpret items?
4. To what extent are test review practices based on multiple sources of information, and how well are various forms of data analysis and data representation used in combination to examine how culture influences student performance?

Test users interested in examining the cultural validity of tests and testing programs are strongly encouraged to try to answer these questions when they examine the supporting documentation provided by test developers. Also, readers are encouraged to ask these questions as they read each of the chapters included in this book.

Author's Note

The research reported in this chapter was funded by the National Science Foundation, Grants REC-9909729, REC-0126344, REC-0336744, and REC-0450090. My sincere thanks to Elizabeth VanderPutten, Larry Suter, and Finbarr Sloane for their support. Also, thanks to my colleagues Elise Trumbull, María del Rosario (Charo) Basterra, Min Li, and Melissa Kwon. The opinions expressed here are not necessarily those of my colleagues or the funding agency.

Notes

1. To meet confidentiality requirements, the real names of the sites in which the investigation was conducted are not disclosed. Fictitious names are used instead.
2. Rogoff's (1995) theory of social participation establishes that activities in which individuals engage within a group take place in one of three planes of social participation—apprenticeship, guided participation, and participatory appropriation—which imply different levels of involvement in sociocultural activity. Being an actor corresponds to the level of participatory appropriation. It involves contributing to an activity and a substantial understanding of the activity.
3. In order to properly interpret the chart for the interview question, one must bear in mind that, due to practical limitations, it was not possible to interview all students on their interpretations of each of the four items shown in Figure 1.1. As a consequence, the number of students interviewed on their interpretations of Gumball Machine for each cultural group is small. To circumvent this limitation, the chart was constructed by aggregating the information on item meaningfulness, regardless of which of the four items shown in Figure 1.1 any given student was interviewed about. This form of aggregating data assumes exchangeability of the four items used as stimulus materials. That is, we assume that, for a given group, the cultural influences on the students' interpretations of items are the same for any of the four items.

References

Abedi, J., Hofstetter, C.H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1–28.

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.

AERA, NCME, & APA (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

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.

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

Brenner, M.E. (1998). Meaning and money. *Educational Studies in Mathematics*, 36, 123–155.

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.

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

Crystal, D. (1997). *A dictionary of linguistics and phonetics* (4th edition). Cambridge, MA: Blackwell.

Ericsson, K.A., & Simon, H.S. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fishman, J.A. (1965). Who speaks what language to whom and when? *La Linguistique*, 2, 67–88.

Gay, G. (2000). *Culturally responsive teaching: Theory, research, & practice.* New York, NY: Teachers College Press.

Gay, G. (2001). Preparing for culturally responsive teaching. *Journal of Teacher Education*, 53(2), 106–116.

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hamilton, L.S., Nussbaum, E.M., & Snow, R.E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 181–200.

Hood, S., Hopson, R., & Frierson, H. (2005). *The role of culture and cultural context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice*. Greenwich, CT: Information Age Publishing Inc.

Kirkhart, K.E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice*, 16(1), 1–12.

Ladson-Billings, G. (1995). Making mathematics meaningful in multicultural contexts. In W.G. Secada, E. Fennema, & L.B. Adjian (Eds.), *New directions for equity in mathematics education* (pp. 126–145). Cambridge: Cambridge University Press.

Lee, O. (1999). Science knowledge, world views, and information sources in social and cultural contexts: Making sense after a natural disaster. *American Educational Research Journal*, 36(2), 187–219.

Li, M., Solano-Flores, G., Kwon, M., & Tsai, S.P. (2008). *"It's asking me as if I were the mother": Examining how students from different groups interpret test items*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD, April.

Megone, M.E., Cai, J., Silver, E.A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research*, 21(3), 317–340.

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

Miller, K.F., & Stigler, J.W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, 2, 279–305.

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.

Nguyen-Le, K. (2010). Personal and formal backgrounds as factors which influence linguistic and cultural competency in the teaching of mathematics. Doctoral dissertation, Educational Equity and Cultural Diversity Program, University of Colorado at Boulder.

Norris, S.P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement*, 27(1), 41–58.

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

Prosser, R.R., & Solano-Flores, G. (2010). *Including English language learners in the process of test development: A study on instrument linguistic adaptation for cognitive*

validity. Paper presented at the Annual Conference of the National Council of Measurement in Education, Denver, Colorado, April 29–May 3.

Rivera, C., & Collum, E. (Eds.) (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Lawrence Erlbaum Associates

Rogoff, B. (1995). Observing sociocultural activity on three planes: participatory appropriation, guided participation, and apprenticeship. In J.V. Wertsch, P. del Río, & A. Alvarez (Eds.), *Sociocultural studies of mind*. New York, NY: Cambridge University Press.

Roseberry, A., Warren, B., & Conant, F. (1992). *Appropriating scientific discourse: Findings from language minority classrooms* (Working paper 1–92). Cambridge, MA: TERC.

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.

Sexton, U., & Solano-Flores, G. (2001). *A comparative study of teachers' cultural perspectives across different cultures.* Poster presented at the annual meeting of the American Educational Research Association, Seattle, WA, April 2–6.

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

Shavelson, R.J., & Webb, N.M. (2009). Generalizability theory and its contribution to the discussion of the generalizability of research findings. In K. Ercikan & W.M. Roth (Eds.), *Generalizing from educational research* (pp. 13–32). New York, NY: Routledge.

Solano-Flores, G. (2001). *World views and test views: the relevance of cultural validity*. Paper presented at the European Association of Research in Learning and Instruction, Fribourg, Switzerland, August 28–September 1.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record*, 108(11), 2354–2379.

Solano-Flores, G. (2008a). *Cultural validity and student performance on science assessments.* Paper presented at the Symposium, Culture and Context in Large-Scale Assessments: Obstacles or Opportunities? organized by Sharon Nelson-Barber and Larry Sutter. Annual meeting of the American Educational Research Association, New York, NY, April 24–28.

Solano-Flores, G. (2008b). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, 37(4), 189–199.

Solano-Flores, G. (2009). The testing of English language learners as a stochastic process: Population misspecification, measurement error, and overgeneralization. In K. Ercikan & W.M. Roth (Eds.), *Generalizing from educational research* (pp. 33–50). New York, NY: Routledge.

Solano-Flores, G., & Gustafson, M. (in press). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Handbook on large-scale assessments and secondary analyses.*

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.

Solano-Flores, G., & Li, M. (2008). Examining the dependability of academic achievement measures for English-Language Learners. *Assessment for Effective Intervention*, 333, 135–144.

Solano-Flores, G., & Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment*, 14, 1–15.

Solano-Flores, G., Li, M., Speroni, C., Rodriguez, J., Basterra, M.R., & Dovholuk, G. (2007). *Comparing the properties of teacher-adapted and linguistically-simplified test items for English language learners*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 9–13.

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., Speroni, C., & Sexton, U. (2005). *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, April 11–15.

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. (2008). In what language should English language learners be tested? In R.J. Kopriva (Ed.), *Improving testing for English language learners*. New York, NY: Routledge.

Tillman, L.C. (2002). Culturally sensitive research approaches: An African-American perspective. *Educational Researcher*, 31(9), 3–12.

Wardhaugh, R. (2002). *An introduction to sociolinguistics* (fourth edition). Oxford: Blackwell Publishing.

Winter, P.C., Kopriva, R., Chen, C.S., & Emick, J. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, 16, 267–276.

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

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