

## On the Cultural Validity of Science Assessments

Guillermo Solano-Flores, Sharon Nelson-Barber

*WestEd, 4200 Farm Hill Boulevard, Redwood City, California 94061*

*Received 10 January 2000; accepted 15 December 2000*

**Abstract:** We propose the concept of cultural validity as a form of test validity in science assessment. The conceptual relevance of cultural validity is supported by evidence that culture and society shape an individual's mind and thinking. To attain cultural validity, the process of assessment development must consider how the sociocultural context in which students live influences the ways in which they make sense of science items and the ways in which they solve them. These sociocultural influences include the values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, as well as the socioeconomic conditions prevailing in their cultural groups. We contend that current approaches to handling student diversity in assessment (e.g., adapting or translating tests, providing assessment accommodations, estimating test cultural bias) are limited and lack sociocultural perspective. We find that attaining cultural validity may conflict with current basic principles and assumptions in testing, such as item independence and standardization. We discuss the ways in which adopting cultural validity as a criterion for test validity makes it necessary to shift assessment paradigms and adopt new procedures for assessment development. © 2001 John Wiley & Sons, Inc. *J Res Sci Teach* 38: 553–573, 2001

For decades, cultural psychologists and child development researchers and theoreticians have acknowledged that culture and society play a critical role in cognitive development (e.g., Vygotsky, 1978; Wertsch, 1985; Wertsch, Del Río, & Alvarez, 1995). Culture influences the ways in which people construct knowledge and create meaning from experience (how they think about things, reason, and solve problems) (Greenfield, 1997a), which relates directly to the ways in which individuals learn and teach in both informal and school settings (Lipka, 1991).

Greenfield (1997a) provided an example that illustrates how ignoring the epistemologies inherent to cultures may lead to misjudgments about an individual's capabilities. Drawing on procedures developed in Cambridge, Massachusetts (an extension of Piaget's work in Geneva), Greenfield administered conservation of quantity tests to unschooled Wolof children in Senegal as she studied the Piagetian stage of concrete operations. After the children transferred water from a short fat beaker to a long thin one, they were asked in their native language whether the

---

*Correspondence to:* G. Solano-Flores; [wsolano@wested.org](mailto:wsolano@wested.org)

Contract grant sponsor: NSF; Contract grant number: REC-9909729

quantity of water was the same, more, or less. According to Greenfield, the children were asked to justify their quantity judgments by responding to the Cambridge interview protocol. Questions such as, “Why do you *think* it is the same (or more, or lesser) amount of water?” and “Why do you *say* it is the same (or more, or lesser) amount of water?” yielded silence on the part of the children.

Not until the question was changed to “Why *is* the water the same (or more or less)?” were justifications for the original quantity judgment successfully elicited. At that point, the unschooled children gave reasons for their judgments that were as articulate as those given to Piaget and his colleagues in Geneva.

These children had an epistemology of mental realism. According to their implicit theory of mind, they were not making a distinction between the nature of reality and their knowledge of it. Consequently, the idea of explaining a statement was meaningless; only the external event could be meaningfully explained [ . . .]. Implicit in this theory of mind was an assumption that there was only a single way to perceive the event of water transfer and its results.

Had an exact translation of the Cambridge conservation procedure been used, it would have been erroneously concluded that the unschooled Wolof children were not able to explain the reasoning behind their quantity judgments. Their theory of mind would have been confounded with their reasoning about the world. The research publication would have incorrectly concluded that unschooled Wolof children had a major cognitive lack in reasoning skill. Instead, the conclusion from pilot testing was that unschooled Wolof children had a different epistemology and therefore required a different interview procedure. When tested with an epistemologically appropriate procedure, the cognitive deficit in reasoning about the world disappeared. (p. 313)

Greenfield’s experience illustrates how assessing across cultures depends on assumptions that may be inaccurate and lead to misinterpretations. Because culture and society shape mental functioning, individuals have predisposed notions of how to respond to questions, solve problems, and so forth. It follows that these predispositions influence the ways in which students interpret material presented in tests and the ways in which they respond to test items. Surprisingly, this view has not been incorporated into the set of actions required to develop valid assessments. Current approaches in assessment give little consideration to understanding how these sociocultural predispositions influence student thinking.

Greenfield’s experience also illustrates how current paradigms in assessment may limit the possibility of obtaining accurate information about individuals outside the mainstream population. From a traditional testing perspective, posing the same prompt in different ways based on the population of students being tested (e.g., Why do you say it is the same amount of water? and Why is it the same amount of water?) is not a valid practice because it violates the principle of standardization in testing. However, from a sociocultural perspective, posing the same prompt in different ways is not only acceptable but also necessary if one is to make valid generalizations about the thinking processes of students from different cultures.

In this article we propose the concept of assessment cultural validity as a form of assessment validity that should be incorporated into science assessment practices. The concept of cultural validity has serious implications for both classroom-based and large-scale science assessment as a reasoning tool that contributes to obtaining more accurate information about the science achievement of cultural minorities. We contend that, to attain cultural validity, the cultural influences that shape the ways in which students interpret and solve science problems must be

considered throughout the entire process of assessment development. Understanding these cultural influences must go beyond interacting with students and may require understanding the procedures and styles of cultural groups. We also contend that addressing cultural validity may imply the adoption of new paradigms in assessment development.

Rather than engaging in a theoretical discussion on validity, we offer empirical evidence of the relevance of the concept of cultural validity. First, we define the concept of cultural validity and discuss how it differs from current approaches to handling cultural diversity in science assessment. Then we discuss several areas in which the reasonings derived from the notion of cultural validity can contribute to improving science assessment and provide examples that support our discussion. Finally, we discuss how attaining cultural validity conflicts with basic principles in testing and stress the need for a paradigm shift in assessment.

### The Concept of Cultural Validity

By cultural validity we refer to the effectiveness with which science assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of science items and respond to them. These sociocultural 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.

Cultural validity is relevant to testing in today's multicultural society, with ethnic and cultural minorities estimated to compose 40% of the overall public school population of the United States (Darling-Hammond, 1997). It is also relevant in these times in which countries are using the results of international comparisons such as the Third International Mathematics and Science Study to make decisions that affect their educational policies. The concept of cultural validity can also provide the conceptual foundation for an era of globalization of the economy, in which assessments across languages, cultures, and countries are becoming increasingly frequent (Hambleton, 1994; Van de Vivjer & Hambleton, 1996).

The concept of cultural validity is relevant to testing cultural and linguistic minorities, given the lack of conceptual clarity about the role that the social implications of testing should play in judging the validity of an assessment (Popham, 1997; Shepard, 1997). Current validity theory addresses the consequential basis of test interpretation (value implications) and test use (social consequences) (Messick, 1989). Low scores "should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected students to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected students' demonstration of competence" (Messick, 1995, p. 7).

Although we agree with this view, we contend that current approaches to assessing cultural minorities do not enable assessment developers to identify with accuracy what in an assessment prevents students from a given cultural group from demonstrating their competence. We contend that critical to attaining cultural validity is the process of assessment development. We also contend that, to attain cultural validity, the process of assessment development needs to incorporate a sociocultural perspective. This perspective should allow assessment developers to identify subtle, important ways in which sociocultural influences and interactions determine student perceptions of what science items are about, what they believe they are expected to do, and what problem solution strategies they use to solve them.

Although current approaches to handling cultural diversity intend to ensure equitable testing, they fail to address the fact that culture shapes the mind. For example, estimating sys-

tematic score differences across cultural groups (Camilli & Shepard, 1994) focuses on correcting for item bias and ensuring assessment comparability across cultures (Tanzer, 1999; Van de Vivjer & Hambleton, 1996; Van de Vivjer & Leung, 1997; Van de Vivjer and Tanzer, 1998). The process of assessment development is not the focus of these approaches. Rather, the finished version of a test originally developed for a particular sector of the population is adapted for a different population or translated into another language (Ercikan, 1998; Hambleton, 1994; Van de Vivjer & Poortinga, 1997).

Current approaches to handling cultural diversity in assessment also may be limited in their intent to promote equitable testing because they use an assimilationist perspective. For example, major science content standards documents define science and science achievement solely in terms of the Western science tradition and disregard alternative views of science and ways of knowing (Lee, 1999a). This assimilationist perspective is also reflected in the ways in which cultural minority students are classified as English language learners and treated in large-scale assessment programs such as the National Assessment of Educational Progress (NAEP). For example, criteria such as ethnicity, immigrant status, the number of years one has lived in this country, or the number of years enrolled in mainstream, English-only classes, help determine whether students are considered to be English language learners. Although these sets of criteria are associated with English proficiency, they may yield inappropriate classifications (Steele & Aronson, 1995).

This assimilationist perspective also figures into the reporting of studies on the criteria for deciding whether an individual should be included or excluded from a NAEP assessment or provided certain accommodations (e.g., extra test administration time). In these studies, results for English language learners are frequently reported along with the results for students with disabilities (e.g., Olson & Goldstein, 1997)—a practice that implies a cultural deficit model in the way in which English language learners are viewed.

Cultural validity is consistent with evidence that worldviews and learning styles are critical in education (e.g., Lee, Fradd, & Sutman, 1995; Lee, 1999b; Steele, 1992). What knowledge is valued, how it is taught and learned, and how success is defined and perceived within a school system may reflect the values, knowledge, and styles of the mainstream population and can alienate cultural, linguistic, and ethnic minorities (Delpit, 1988). For example, Yup'ik children in Alaska learn important skills (such as fishing, building fish racks, and using stars to navigate on land) from observing experienced adults and actively participating as apprentice-helpers. Children and adults engage in the same activity for long periods of time in which verbal interactions are not central to the learning process. This style may not be optimal within a traditional Western school system that organizes teaching and learning around short and frequent class periods in which students are expected to listen passively to teachers, follow directions, and give long verbal answers to questions (Lipka, 1998).

Cultural validity questions approaches in instruction and assessment based on cultural stereotyping, which affects intellectual identity and performance (Steele & Aronson, 1995; Steele, 1997) and has been extensively documented, along with poverty and ineffective and segregating tracking systems (e.g., Darling-Hammond, 1995; Knapp, Shields, & Turnbull, 1995; Oakes, 1985, 1990) as responsible for low academic performance of cultural minorities in the United States. To attain cultural validity, the sociocultural context in which students live must be properly understood. Therefore, assumptions about students based solely on criteria such as ethnicity and native language are unacceptable because they do not provide detailed, accurate information about the specific sociocultural influences that shape student thinking.

Cultural validity goes beyond the concept of fairness in testing as a criterion for test validity (Linn, Baker, & Dunbar, 1991). Correcting for cultural bias, promoting the participation of

ethnic minorities in pilot student samples, and providing accommodations for linguistic minorities are remedial strategies that address cultural differences not considered in an assessment's original plan. Ideally, if cultural validity issues were addressed properly at the inception of an assessment and throughout its entire process of development, there would be no cultural bias and providing accommodations for cultural minorities would not be necessary. Cultural validity thus establishes the need for incorporating the reasonings inherent to a sociocultural perspective in assessment practices.

### Cultural Validity as a Reasoning Tool in Assessment Development

We identify five areas in which the reasonings derived from the notion of cultural validity can contribute to improving science assessment: (a) student epistemology, (b) student language proficiency, (c) cultural world views, (d) cultural communication and socialization styles, and (e) student life context and values. From the perspective of cultural validity, these areas have been either neglected or inadequately addressed by current assessment practices.

#### *Student Epistemology*

Recent studies on assessment validation have focused on the cognitive demands of mathematics performance tasks (Magone, Cai, Silver, & Wang, 1994) and the cognitive activity elicited by science hands-on and concept map tasks (Baxter, Elder, & Glaser, 1996; Hamilton, Nussbaum, & Snow, 1997; Baxter & Glaser, 1998; Ruiz-Primo, Shavelson, Li, & Schultz, in press). These studies have shown that the complexity of the knowledge and reasonings used by students influence their cognitive activity when they take science tasks. Concurrent and retrospective student verbalizations (Ericsson & Simon, 1993) are coded according to categories of cognitive activity such as problem representation, monitoring, strategy use, and content-based explanation (cf. Chi, Feltovich, & Glaser, 1981).

Although cognitive approaches are beginning to gain acceptance among the measurement community, they fail to address the fact that knowledge is a social construction. A cognitive perspective can provide valuable information about the students' use of content knowledge and problem-solving strategies but cannot provide information on the students' epistemologies and the sociocultural factors that influence how students construct knowledge.

The following example comes from an investigation that examines how culture influences the way in which students interpret science and mathematics items and respond to them (Solano-Flores & Nelson-Barber, 2000). It illustrates how understanding student epistemologies can help to assess the quality of a science item.

A Latino girl is given a NAEP fourth-grade science exercise on erosion (Figure 1) (National Assessment of Educational Progress, 1996). The exercise shows two pictures, A and B, of the same river and mountains. In A, the mountains look low and round and the river wide. In B, the mountains are high and pointy and the river narrow. The item asks the student to circle the letter under the picture that shows how the river and mountains look now (as opposed to how they looked millions of years ago).

The girl circles the letter B, the wrong letter. During the interview, we learn that she does not remember learning about mountains at school. Nor has she had experiences such as climbing mountains or hiking in the countryside. But then she recalls one first-hand experience related to mountains:

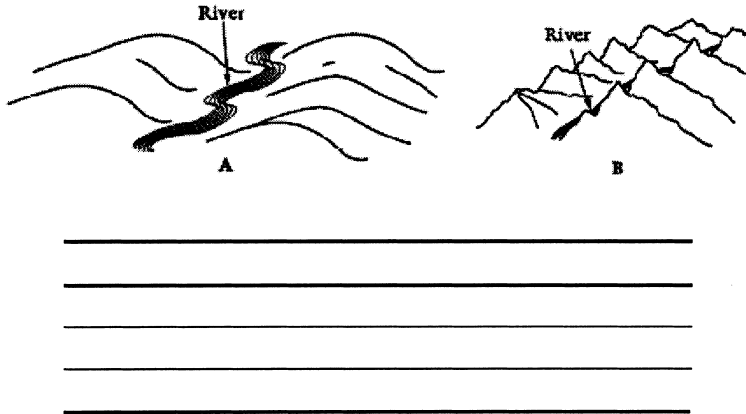


Figure 1. NAEP Erosion item.

[When we] went to Salinas, and . . . when we were passing by there, I saw like a mountain . . . like a . . . mountain and then there was rain . . . water rain through it. And then I saw it and then [a relative] said, “look” and I saw it and it was like that [Picture B].

To see how she interprets the item we ask her how she would write it to make it easier for other kids like her to understand it. She responds:

*I would write, uh, “Circle the . . . circle the . . . the letter of the mountain and the river you think was the past and think . . . yeah, was the past, and explain how you know.”*

We ask her if she thinks she can respond to the item in its new form.

*Yeah... This one [Picture A]. . . . I think this one is the past and this one [Picture B] is today.*

“Explain why,” we ask.

*[. . .] Because this one looks familiar [Picture B] and this one [Picture A] . . . I don’t know, I’ve never seen it.*

Certainly, insufficient opportunity to learn about erosion in a formal science curriculum can account for the incorrect response. The student does not use the concept of erosion in her reasonings. She is not aware that weathering can transform the shape of mountains over time. However, from a sociocultural perspective, a more complex picture emerges. The student seems to have an epistemology that links what happens in the present to her personal experience. According to this epistemology, mountains like those in Picture A are from the past because she has never seen them. Her first-hand experience with mountains is scant. Moreover, her experience with mountains is limited to mountains on the West Coast—young mountains, geologically speaking, and pointy in shape.

It is clear from this analysis that the item should be revised. By exploring how this student related the content of the item to her own experiences, we discerned that the item implies that all mountains that exist in the present are pointy. This student would probably still give an incorrect answer even after correcting for that overgeneralization in the item. One could argue that a student who is familiar enough with the concept of erosion should be able to respond correctly without needing to rely on personal experiences and personal epistemology. However, even if this were true, in its current version the item may be privileging students whose first-hand

experience is with flat and round mountains over students whose first-hand experience is with pointy mountains.

We have preliminary evidence that the way students interpret science items and respond to them may be more influenced by personal experience than formal school learning experience. Frequently, everyday-life experiences seem to be what first comes to students' minds when they respond to science items. In some cases, students who give incorrect responses to the items are able to demonstrate, during the interview, good understanding of the scientific concepts addressed when those concepts are discussed in relation to their personal experiences. This preliminary evidence is consistent with findings that everyday-life experiences are an important influence in student performance. For example, outside-of-school experiences can account for gender differences on science achievement tests (e.g., Hamilton, 1998, Jovanovic, Solano-Flores, & Shavelson, 1994).

### *Student Language Proficiency*

In addition to the natural difficulties of using a language that is not their native language, students who are tested in a second language must deal with the fact that there is a specialized knowledge within that second language that is specific to the discipline. Students who may understand a scientific phenomenon or a mathematical principle may still not be able to demonstrate that knowledge because of the lack of appropriate academic vocabulary. Even if they possess that vocabulary, the way in which they use it may not look or sound as technical as it would look or sound for native speakers (Kusimo et al., 2000).

Approaches to testing linguistic minorities purport to reduce the impact of limited English proficiency on the skills being measured (Pellegrino, Jones, & Mitchell, 1999) and ensure the validity of measures of academic achievement for English-language learners (August & Hakuta, 1997). Still, there are three major challenges to attaining these goals. First, assessment depends on the use of language and is based on assumptions about students' capabilities to understand written and spoken language and to communicate their ideas in writing (García & Pearson, 1994). Second, even within populations of students for whom language proficiency is not an issue, student performance is extremely sensitive to the ways in which prompts are worded (Baxter, Shavelson, Goldman, & Pine, 1992). Third, student scores cannot be considered valid if they confound language comprehension and the academic skills addressed by a test (Durán, 1989).

Unfortunately, the basic notion that important cultural influences shape language and language use (e.g., Greenfield, 1997a) has not permeated current thinking and approaches to testing linguistic minorities. As a result, procedures used with these students are based on overgeneralizations or assumptions that may be erroneous.

One example illustrates that even procedures considered sound among the measurement specialist community may still fail to serve linguistic minority students if cultural issues are not properly considered. A study on the use of bilingual, English-and-Spanish formats in science assessment (Solano-Flores, Ruiz-Primo, Baxter, & Shavelson, 1991) found that some Latino, English language learners were unfamiliar with some words the researchers believed to be part of the students' everyday language. These words made their way into the Spanish version despite use of an experienced, native Spanish-speaker translator, a panel of bilingual scholars who reviewed the Spanish version and translated it back into English to monitor retention of the original meaning, and testing the Spanish version with some students to make refinements. Although the translation was accurate and its grammar impeccable, it actually reflected the researchers' thinking and preconceptions about the target students.

The tremendous influence that wording has on the way students respond to science exercises is not properly addressed when assessments are translated. For example, in the Paper Towels assessment, students are asked to carry out an experiment to find out which of three paper towel brands “holds, soaks up, or absorbs the most water and which one holds, soaks up, or absorbs the least water.”

[If the word] *soaks* is used, students pour water on the table and use a towel to soak it up [...]. If [...] *holds* is used, students place a towel across a container and see how much water it will support. *Absorb* is not a familiar word to many students (Baxter et al., 1992, p. 16).

Proper wording can be accomplished only through a series of review-revise iterations in which wording is refined based on the observed performance and the verbalizations of pilot students (Solano-Flores & Shavelson, 1997). Notwithstanding, when an assessment is translated, this delicate process of language refinement does not necessarily take place. To make things worse, in practice, translations are not always made in accordance with published guidelines (e.g., Van de Vivjer & Hambleton, 1996). For example, a test publisher may allow only one week for translating a test, whereas its development in the source language may have taken much longer. Because of these flaws and serious deficiencies in the construction and administration of translated tests, testing in the first language might not be fair if intrinsic language properties are not taken into account (Figueroa, 1980).

Another example illustrates how assessing knowledge across languages necessarily implies assessing across cultures. It illustrates how a sociocultural perspective allows assessment developers to make informed decisions involving word equivalence across languages. The example comes from a project that evaluates our model for the concurrent development of assessments in two language versions, as an alternative to the traditional approach of translating tests originally created for a mainstream population of native English speakers (Solano-Flores & Nelson-Barber, 1999a, 1999b; Solano-Flores, Trumbull, & Nelson-Barber, 2000).

In this project, seven bilingual teachers from a district-level bilingual program were trained to use our concurrent development model. The teachers developed English and Spanish language versions of the same sets of exercises. Because both language versions were developed simultaneously, the issue of meaning across languages was at all times factored into the discussion of the exercises. For instance, the following sentence is part of a constructed-response item involving unit conversion:

[In your response, you] *can use pictures, numbers and/or words.*

In the Spanish version of the item, “o” (“or”) is used instead of “y/o” (“and/or”). From the discussion during the assessment development sessions, teachers concluded that their native English-speaking students and their native Spanish-speaking students have different ways of interpreting the words “or” and “o.” Native English speakers tend to interpret “or” as an exclusive disjunction; native Spanish speakers tend to interpret “o” as an inclusive disjunction. According to this reasoning, using “and/or” in the English version of the item can benefit native English speakers, but using “y/o” in the Spanish version can be confusing to native Spanish speakers.

This evidence supports the notion that, to attain equitable assessment across two languages, both languages must receive the same treatment throughout the entire process of assessment development. It also indicates that effective testing of linguistic minorities is possible only if the

ways in which culture shapes language are properly considered throughout the process of assessment development.

### *Cultural World Views*

It has been proposed that efforts to promote high academic standards for all students must include a process of mediating academic content with students' cultural experiences to make such content accessible, meaningful, and relevant for diverse students (Lee & Fradd, 1998).

So-called culturally responsive pedagogy is based on developing educational methods that are situated in students' cultural experiences (Bartolomé, 1994) and, in the case of science, views transitioning from a student's life-world to a science classroom as a cross-cultural experience (Aikenhead & Jegede, 1999). Such an approach is especially relevant in the case of cultural groups for whom knowledge is highly contextualized. Understanding a given culture's ways of knowing and traditional knowledge is the basis for establishing connections between cultural content and academic content (Nelson-Barber & Estrin, 1996). To accomplish this for a given cultural group, its world views must be properly understood (Kawagley, 1995) and its communities must be empowered to take part in deciding what knowledge is relevant for their children to learn (Lipka, 1991, 1994).

Consistent with these reasonings, assessments developed for a specific cultural group should be sensitive to its ways of knowing and traditional knowledge. However, how accurately assessment developers understand the ways of knowing and the traditional knowledge is critical to attaining cultural validity. Stereotypes and unrealistic expectations about the students' knowledge and skills can result from the simplistic assumption that all the individuals from a culture are familiar with certain knowledge associated with it.

The following example illustrates that understanding the ways of knowing and the traditional knowledge relevant to assessing students from a given cultural group is a complex endeavor. It comes from a project for the development of a curriculum based on the knowledge of Yup'ik elders (Lipka, 2000). In this curriculum, content is taught through everyday problems that are related to traditional Yup'ik activities such as making baskets, fishing, building fish racks, and navigating across the tundra using landmarks and the positions of the stars.

The content of the module, *Star Navigation* (Bradley-Kawagley, 1998; Lipka, 2000) intends to reflect the knowledge of Frederick George, from Akiachak, one of the last living Yup'ik expert navigators. From the work done by Lipka (1991, 1998), we know that Mr. George uses a mental representation of the sky as a big circle with Polaris, the North Star, at its center, dividing the circle into angles (to which he refers as hours, as in the face of a clock). This mental representation includes the motion and position of the main stars and constellations throughout the year. With this knowledge, Mr. George is capable of solving complex location and orientation problems based only on the position of the stars at any given time of the year.

One hands-on assessment aligned with the module on star navigation is *Tunturyuk* (Figure 2) (Solano-Flores, 2000). *Tunturyuk*, the caribou, is the Yup'ik name of the constellation known as the Big Dipper in Western culture; the caribou turns around Polaris, the North Star, and is always oriented to Polaris such that it appears to be walking toward it.

In *Tunturyuk*, students are given two sky maps, A and B. In Sky Map A, students can rotate each of two transparent discs around Polaris. One of the discs, the larger one, is a protractor. The other disc has the constellation of *Tunturyuk*. Students are given the position of *Tunturyuk* in December 10, 1998, at 10 P.M. Based on that position, their task is to determine the exact position of the constellation in the sky when it was May 10, 1998, at 6:20 P.M. and draw it in Sky Map B.

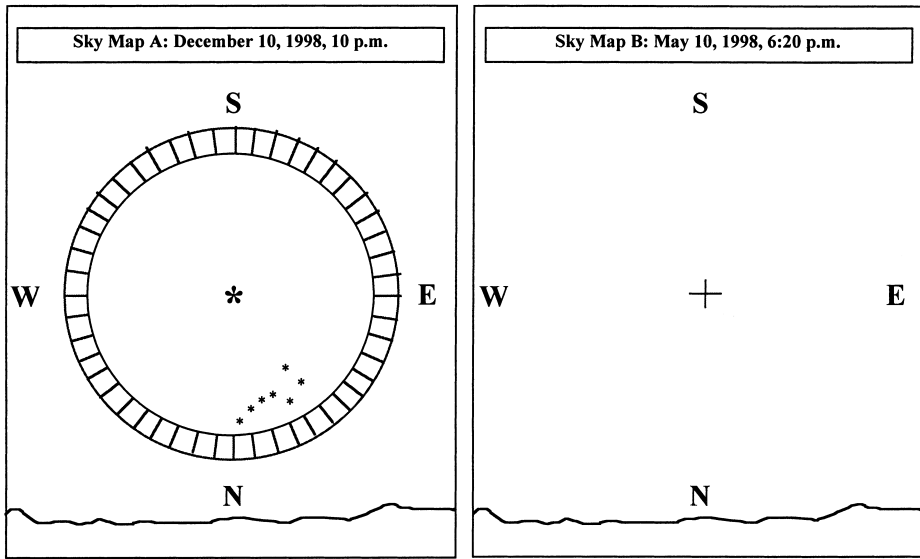


Figure 2. Tunturyuk assessment sky maps.

Critical to solving the problem correctly is the knowledge that Tunturyuk moves counter-clockwise around Polaris at a speed of  $15^\circ/\text{h}$  and that the position of Tunturyuk on any day and exactly 30 days later at the same time of the day changes  $30^\circ$ . Also critical to solving the problem correctly is the appropriate use of angular measures.

The Tunturyuk assessment was developed based on the knowledge of Mr. George's thinking, as captured by Star Navigation. As a part of the development process, Mr. George reviewed the assessment to ensure that it reflected his thinking. Conventional maps show the perspective of a viewer who is situated above the land. In contrast, the sky maps used in both the module and this assessment show the perspective of a viewer who is standing on the land. By using Polaris as a reference, the viewer situates himself or herself facing the North. The circle represents the sky above the viewer. At the bottom and the top of the map are represented, respectively, the North horizon and what is behind the viewer.

Tunturyuk shows us that to develop assessments that are sensitive to the characteristics of a cultural group, the cognition of its individuals must be properly understood. An assessment such as Tunturyuk would have been impossible to develop without the knowledge base provided by the research that supported the development of the module on star navigation. Even well-intentioned efforts to honor Yup'ik traditional knowledge would have failed to capture the sophistication with which a sky map can represent not only the position of the stars in the sky but also the position of the person on land as part of the environment.

### *Cultural Communication and Socialization Styles*

The need to consider culturally inherent communication and socialization styles to obtain accurate information about student knowledge and skills has been discussed thoroughly in the literature. The style in which students provide their responses and structure their writing may be influenced by the way in which discourse is organized in their native languages (Estrin, 1993). Also, some languages tend to be circular; or some cultures do not encourage long responses in

their students (Greenfield, Quiroz, & Raeff, 1998; Snow, 1983). When presented with open-ended, constructed-response items students from these cultures may tend to provide brief, rather than extended, detailed responses. Not taking into account these important cultural differences may produce inaccurate perceptions of their performance.

Unfortunately, knowledge of these facts has not influenced current assessment practices. This is likely because considering the communication and socialization styles of a culture in assessment conflicts with basic assumptions and principles in assessment. For example, it is well known that conducting individual interviews with the members of a community with a fixed-format questionnaire in which questions are totally independent from each other may not render accurate results for cultures in which knowledge is constructed collectively, decisions are made in a group, and discourse and thinking are highly contextualized activities (Greenfield, 1997b). Evidence like this questions the fairness of the principle of standardization in test administration, the assumption of item independence, and even individual testing.

Efforts to address patterns of communication and socialization in testing are oriented to making educators, item writers, and raters aware of the important linguistic and cultural influences that shape student responses (e.g., Koelsch, Estrin, & Farr, 1995; Kopriva, 2000; Kopriva & Saez, 1997; Kopriva & Sexton, 1998; Kusimo et al., 2000; Shaw, 1997). These efforts focus on item writing and the proper interpretation of the students' responses in scoring but their emphasis is not the process of assessment development.

One example illustrates how the communication and socialization styles of cultural can shape the procedures used to develop an assessment. The example comes from Kayaks, another hands-on assessment developed for the curriculum based on traditional Yup'ik knowledge.

Kayaks is about body measures. Body measures are part of the Yup'ik way of solving everyday problems. For example, a kayak's measures are determined based on the arm and forearm lengths of its owner (Figure 3a). In Kayaks, students are given two transparent charts with two characters' body parts (Figure 3b) and a sheet with the illustrations of eight kayaks of different lengths (Figure 3c). The task is to find out which kayaks belongs to which characters.

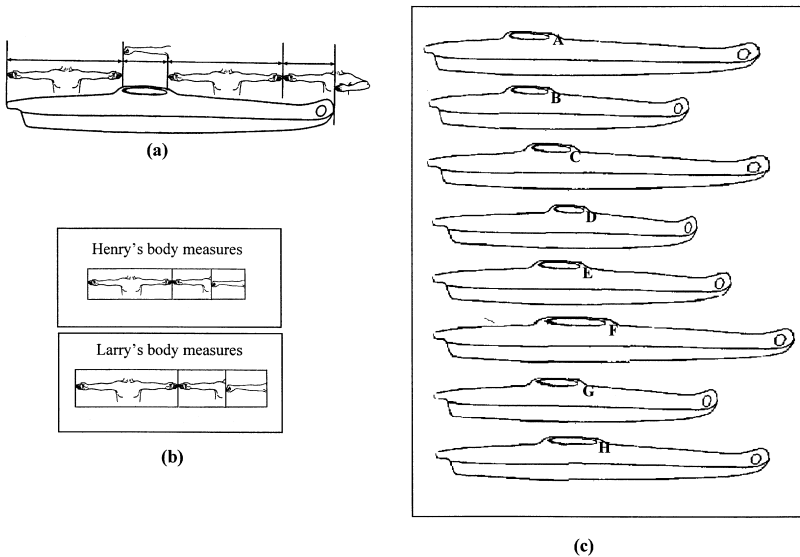


Figure 3. Kayaks assessment materials.

To solve the problem correctly, students need to use the body part charts as rulers to measure the kayaks. They also need to measure each kayak to collect both confirming and disconfirming evidence as to whether any of the kayaks belong to any of the characters.

Strategies for the review of science assessments rely on the expertise of teachers, curriculum developers, and science educators. These professionals provide feedback on issues such as information accuracy, content representation, task difficulty, and wording (Brown & Shavelson, 1996; Shavelson, 1995). One way of obtaining feedback from them is to have them take the assessment individually, as if they were the students, and then interviewing them about the kind of thinking elicited by the assessment and the kind of knowledge they used to solve the problems.

Because elders have a prominent role in Yup'ik society and their experiences have been critical to the development of the curriculum based on traditional Yup'ik knowledge, including them in the team of reviewers made perfect sense as a strategy for validating our assessment. They would judge how well the assessment reflected traditional Yup'ik knowledge.

During a meeting in which about a dozen elders gathered, we asked them to review the Kayaks assessment. As a first approach, we tried the review procedure described above, which has rendered good results in the context of urban and suburban schools in the United States. However, soon it became clear that working individually was inconsistent with the way elders relate to each other. As soon as we asked an elder to come with us, he would ask other elders to accompany him.

We changed our strategy and convened a meeting with all the elders. We sat with them around the table and gave each a copy of Kayaks. Then we asked a bilingual Yup'ik teacher to assist us as a translator because some of the elders only spoke Yup'ik. However, this approach was not productive, either. It turned out that the teacher was uncomfortable to find herself in a situation in which she appeared to be testing the elders' knowledge. In addition, we were asking the elders to work individually although that they were sitting together at the same table.

The meeting became productive only when the teacher asked the elders to solve the problem as a team—which elicited a rich discussion on the nature of the task, and when she solved the problem thinking aloud and asked the elders to provide her with guidance and to correct her when her reasoning was incorrect.

Had we rigorously applied the assessment review procedures used successfully in other contexts—essentially having participants complete tasks individually—we would have been unable to obtain the desired information. Understanding and following the communication and socialization styles of Yup'ik elders was critical to properly refining the Kayaks assessment.

### *Student Life Context and Values*

That school and assessment tend to reflect the thinking and experiences of the mainstream sector of the population has been identified as a factor that accounts for the low performance of certain cultural minority students (Delpit, 1988, Hidalgo, Bright, Siu, Swap, & Epstein, 1995). It is reasonable to expect, then, that more equitable testing can be attained if assessments are contextualized in the students' cultural experiences.

However, identifying the experiences that are relevant to or representative of the cultural context of a given group of students can be extremely difficult. The challenge is that often assessment developers think they know what defines the targeted students' culture—and often their perceptions are inaccurate. This can happen even when assessment developers are aware of cultural issues, are familiar with the population of students targeted, or share with students a

similar cultural heritage. Attempts to contextualize assessment based on the students' cultural experiences can be flawed or misleading if there are no procedures that allow developers to test the accuracy of their assumptions about other peoples' cultures.

Part of our experience developing the Kayaks assessment illustrates this point. During the discussion with the elders who reviewed Kayaks, we found that only one of them had experience building kayaks. As it turned out, kayaks are more part of life in coastal areas than in villages located more inland along the Kuskokwim River, where canoes are used. Although Yup'ik communities from these two areas may not be more than 100 miles apart, most communities have developed relatively independently owing to the harsh terrain and climate, which has been enough to create considerable linguistic and cultural differences over time.

Behind our erroneous assumption that kayaks are part of the everyday life of Yup'iks was an overgeneralization of their characteristics. Being a Yup'ik does not make a person a kayak user, the same way being Swiss does not make a person a banker or a clock maker. Several questions arose. Is using kayaks as a strategy for contextualizing the assessment appropriate for all Yup'ik students? Does using kayaks privilege Yup'ik students from a given area and penalize students from communities outside that area? Should we replace kayaks by another object that is equally familiar to all Yup'ik students? Or should we use different objects depending on the communities where they live?

It seems to us that even students who have never used kayaks recognize them as an element of their culture. However, we are not certain whether honoring the Yup'ik culture by using kayaks is appreciated by students and, if this is the case, whether this has an effect on their performance. These issues are yet to be investigated.

In any case, what is valuable in the assessment is not necessarily the use of kayaks, but the fact that it promotes body-based measurement skills. Whereas these skills are relevant to everyday Yup'ik life (Lipka, 1998) and many indigenous cultures (Denny, 1986), they are neglected in major standards documents (e.g., National Council of Teachers of Mathematics, 2000; National Research Council, 1996).

A second example illustrates how perceptions about the context in which students live may be based on superficial characteristics of the students' culture. It also illustrates how deep consideration of cultural context in assessment development may conflict with the principle of standardization in testing. The example comes from the project in which seven bilingual teachers developed the English and the Spanish version of the same assessment for a school district with a bilingual program (Solano-Flores, Trumbull, & Nelson-Barber, 2000). About 60% of the student population in this school district are native English speakers and about 40% of the student population are Latino, migrant, English-language learning students, the majority from the state of Michoacán, Mexico (Ross, 1999).

At the beginning of the project, the teachers' attempts to consider the students' cultural backgrounds focused on superficial characteristics. For instance, one of the exercises was about making tacos for the class, a context whose relevance to promoting equitable testing was debated. However, as the project went on, the discussion increasingly focused on the equivalence of the test across languages and cultures. For example, in discussing the complexity of information that the tables with measurement specifications should have, the issue of metric systems arose. Because Mexico uses the decimal metric system, should decimal metric system units be used in the exercises in Spanish? If the English system units were kept in the Spanish version, should English or Spanish abbreviations be used? As the discussion went on, a deeper level of analysis was reached. Are these Mexican students more familiar with kilograms and grams—which are part of their country's culture—than pounds and ounces—which may be part of their everyday life experience in this country? What is the real sociocultural context in which

these students live? The teachers switched from using cultural generalizations to support their reasonings about the linguistic or cultural adequacy of the Spanish version (e.g., “Latinos think that . . .” or “Mexicans prefer . . .”) to relying only on first-hand experience with their students in the context of the community where they live.

One unexpected outcome of the assessment development sessions was that teachers identified value systems and experience as relevant to testing. Most of their students come from rural areas in the state of Michoacán, Mexico, where large-scale, standardized testing practically does not exist. Therefore, these students may differ from mainstream American students in their knowledge about the consequences and implications of testing and their level of familiarity with the kinds of problems used by their state’s examination system. Based on the results from pilot-testing, it was hypothesized that the structure of some exercises parallel the patterns of thinking of a given cultural group better than the structure of other exercises. For example, an exercise that states the problem or goal up front and then provides contextual and ancillary information might fit better the patterns of problem solving of a given cultural group than an exercise in which those elements are presented in the opposite order.

Although many of these issues need to be investigated, it is clear that proper consideration of cultural factors in assessment cannot be accomplished without proper consideration of the context in which students live. Efforts to contextualize assessment based on the students’ cultural experiences can be ineffective or even alienating if they rely on assumptions based on race or ethnicity. Careful development work with a sociocultural perspective is needed to identify, from within the context in which students live, what is relevant or should be considered in a given assessment.

#### Conclusions: Cultural Validity and the Need for New Paradigms in Science Assessment

In this article we address how culture and society shape the ways in which individuals construct knowledge and create meaning and what this means for science assessment. We propose the concept of cultural validity as a form of validity that should be incorporated into assessment practices. Cultural validity is the effectiveness with which science assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of science items and respond to them.

Critical to attaining cultural validity is the process of assessment development. Only through this process can assessment developers identify ways in which sociocultural influences and interactions determine student perceptions of what science items are about, what they believe they are expected to do, and what problem-solution strategies they use to solve them. Existing approaches to handling student diversity do not focus on the process of assessment development. In addition to reflecting an assimilationist, cultural deficit model perspective in the way in which cultural minorities are viewed, they give little regard to the role of society and culture in thinking—which limits the effectiveness with which they promote equitable testing.

Cultural validity establishes the need for new methods of assessment development and new views of the testing of cultural minorities. We presented examples that illustrate how the perspective of cultural validity can enrich the reasoning of assessment developers in five areas: student epistemology, student language proficiency, cultural world views, cultural communication and socialization styles, and student life context and values.

The main implication of the reasonings and examples presented is that, from the perspective of cultural validity, what is being done to address cultural diversity in assessment is not sufficient to ensure equitable testing. Current approaches to handling cultural diversity do not focus on

understanding student thinking and the sociocultural influences that shape thinking. As a result, the assessment of cultural minorities is guided by simplistic assumptions about language and culture and cultural misconceptions and stereotypes, and gives little consideration to the context in which students live.

As we have seen, the notion of cultural validity implies the need for a paradigm shift in assessment. Current assessment practices used with cultural minorities should be revised. For example, because an individual's cognition cannot be properly understood without a thorough understanding of the context in which that individual lives, examining student epistemologies should be part of the actions taken to validate an assessment. Back translation no longer should be considered a guarantee for proper translation if the culture and the context in which students live is not properly considered. Or, items might need to be worded in different ways, depending on the group of students tested.

It seems that the major tension between attaining cultural validity and following current assessment practices and principles has to do with the notion of standardization. It is difficult to think of large-scale tests in which the wording and appearance of items vary with cultural group. However, we believe that in the context of a multicultural society and a global economy, item equivalence across cultures should be given precedence over standardization. After all, many current testing practices, such as assessment accommodations and computer adapting testing, in which students are given different sets of items depending on their responses to previous items, deviate to some extent from strict standardization (Kopriva, 1999). Why, then, should it be surprising that standardization gives way to cultural validity? Needless to say, considerable amount of conceptual and methodological work is needed to meet the challenges of incorporating cultural validity into science assessment practices.

Perhaps the major challenge posed by the concept of cultural validity has to do with who needs to be involved in the process of assessment development and who decides what is relevant to a given cultural group. As our experience developing assessments for Yup'ik and Latino students illustrates, the process of assessment development must be sensitive to the subtle but important differences in the context in which individuals from the same cultural group live. According to this, no valid generalizations regarding culture can be made based on criteria such as ethnicity, country of origin, native language, or in what state students live. People who are external to a cultural group tend to make overgeneralizations and rely on cultural stereotypes. As a result, they may misperceive or misrepresent what of that group's culture is relevant to an assessment. Cultural validity, then, cannot be attained if the current assessment systems remain unchanged—and only a few people write the items or develop the assessments that are administered to all students.

It has been said that if educators and policy makers are serious about equity, teachers should be viewed as the only contributors in a research agenda of diverse learners who can provide accurate information based in real-world settings with both cultural minority and mainstream students (Fradd & Lee, 1999). Thus, new assessment systems have to be devised that allow for a more genuine participation of educators from all communities in the process of developing the assessments that are given to their students. Implementing this genuine participation may require that we transform the way in which we think about large-scale testing. For example, a national test can consist of a set of item types that specify a universe of knowledge (Hively, Patterson, & Page, 1968). Each item type prescribes a specific structure; the science topic, concept, or skill addressed; and the science standards covered. This set of item types is distributed nationally. Educators from each community customize the national test to the characteristics of their students by creating items that meet the specifications of those item types according to their students' cultural backgrounds and the context in which they live. Decisions concerning the

wording of the items, what kind of contextual information is included, and test administration time, among others, are made locally within each community and even each school.

Another action consistent with the reasonings derived from the concept of cultural validity consists of rethinking what kinds of skills are needed to develop good assessments. It has been recommended that science assessment development teams include science teachers, science educators, scientists, and assessment specialists (e.g., Brown & Shavelson, 1996; Shavelson, 1995; Shavelson, Brown, Solano-Flores, & Ruiz-Primo, 1993). Although this team configuration ensures the expertise needed to produce tasks that are appropriate to the age or grade level of the students, have appropriate content coverage, and are scientifically accurate and psychometrically sound, it does not necessarily ensure that the tasks are culturally valid. Other professionals, such as cultural anthropologists, should be included in the team of assessment developers. These specialists should ensure that a sociocultural perspective is incorporated in the thinking and the decisions made by the team throughout the entire process of assessment development.

Many of the issues discussed in this article are unresolved and many of the ideas derived from the concept of cultural validity pose serious logistical and methodological challenges. They imply the need for new developments in psychometric theory and radical changes in the ways in which assessments are developed and administered. This seems to be the price that needs to be paid to attain equitable science testing.

#### Note

Most of the research reported here was supported by National Science Foundation Grant REC-9909729. The opinions expressed do not necessarily reflect those of the funding agencies. The authors are grateful to Jerry Lipka for collegial support; Elise Trumbull for participation in the design of the investigation and the facilitation of the assessment development sessions for the project on concurrent assessment development; Maricruz Díaz for carefully transcribing the concurrent assessment development sessions; Ursula Sexton and Rachel Lagunoff for refining the data collection instruments and interviewing students for the project on assessment cultural validity; and two anonymous reviewers for providing valuable and careful comments on a previous version of this article. The authors are also grateful to the teachers and students of the Wenatchee School District and to the elders, teachers, and students of the Yup'ik communities in Alaska. The illustrations in Figures 3(a) and 3(b) were made by Putt Clark, graphic artist with the University of Alaska, Fairbanks.

#### References

- Aikenhead, G.S., & Jegede, O.J. (1999). Cross-cultural science education: A cognitive explanation of a cultural phenomenon. *Journal of Research in Science Teaching*, 36, 269–287.
- August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: 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. Washington, DC: National Academy Press.
- Bartolomé, L. (1994). Beyond the methods fetish: Toward a humanizing pedagogy. *Harvard Educational Review*, 64, 173–194.
- Baxter, G.P., Elder, A.D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist*, 31, 133–140.
- Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17, 37–45.
- 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, 1–17.

Bradley-Kawagley, C. (1998). Star navigation: An NSF-sponsored module with an emphasis on geometry. Developed in cooperation with the project, "Adapting Yup'ik elders' knowledge: pre-K to 6th grade instructional materials development." Fairbanks, AK: University of Alaska, Fairbanks.

Brown, J.H., & Shavelson, R.J. (1996). *Assessing hands-on science: A teacher's guide to performance assessments*. Thousand Oaks, CA: Corwin Press.

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

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.

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

Darling-Hammond, L. (1995). Inequality and access to knowledge. In J.A. Banks & C.A.M. Banks (Eds.), *Handbook of research in multicultural education*. New York: Simon & Schuster Macmillan.

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work*. San Francisco: Jossey-Bass.

Delpit, L. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review*, 58, 280–298.

Denny, J.P. (1986). Cultural ecology of mathematics: Ojibway and Inuit hunters. In M.P. Closs (Ed.), *Native American mathematics* (pp. 129–180). Austin, TX: University of Texas Press.

Durán, R.P. (1989). Testing of linguistic minorities. In R.L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 573–587). New York: Macmillan.

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

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

Estrin, E. (1993). *Alternative assessment: Issues in language, culture and equity*. Knowledge Brief No. 11. San Francisco: Far West Laboratory.

Fradd, S.H., & Lee, O. (1999). Teachers' roles in promoting science inquiry with students from diverse language backgrounds. *Educational Researcher*, 28, 14–20.

Figueroa, R.A. (1980). Intersection of special education and bilingual education. In J.E. Alatis (Ed.), *Georgetown University roundtable on languages and linguistics* (pp.147–161). Washington, DC: Georgetown University Press.

García, G.E., & Pearson, P.D. (1994). Assessment and diversity. *Review of Research in Education*, 20, 337–390.

Greenfield, P.M. (1997a). Culture as process: Empirical methods for cultural psychology. In J.W. Berry, Y.H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd Ed), Vol. 1: Theory and method. (pp. 301–346). Needham Heights, MA: Allyn & Bacon.

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

Greenfield, P.M., Quiroz, B., & Raeff, C. (1998). Cross-cultural conflict and harmony in the social construction of the child. In S. Harkness, C. Raeff, & C.M. Super (Eds.), *The social construction of the child: Nature and sources of variability. New directions in child psychology*. San Francisco: Jossey-Bass.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.

Hamilton, L. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis*, 20, 179–195.

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

Hidalgo, N.M., Bright, J.A., Siu, S.-F., Swap, S.M., & Epstein, J.L. (1995). Research on families, schools, and communities: A multicultural perspective. In J.A. Banks & C.A.M. Banks (Eds.), *Handbook of research in multicultural education* (pp. 498–524). New York: Simon & Schuster Macmillan.

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

Jovanovic, J., Solano-Flores, G., & Shavelson, R.J. (1994). Science performance assessments: Will gender make a difference? *Education and Urban Society*, 26, 352–366.

Kawagley, A.O. (1995). *A Yupiaq worldview: A pathway to ecology and spirit*. Prospect Heights, IL: Waveland Press.

Knapp, M.S., Shields, P.M., & Turnbull, B.J. (1995). Academic challenge in high-poverty classrooms. *Phi Delta Kappan*, 76, 770–776.

Koelsch, N., Estrin, E.T., & Farr, B. (1995). *Guide to analyzing linguistic and cultural assumptions in assessment*. San Francisco: Far West Laboratory for Educational Research and Development.

Kopriva, R. (1999, April 19–23). A conceptual framework for the valid and comparable measurement of all students. Paper presented at the American Educational Research Association annual meeting, Montreal, Canada.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. 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.

Kopriva, R., & Sexton, U.M. (1998). *Guide to scoring LEP student responses to open-ended science items*. Washington, DC: Council of Chief State School Officers.

Kusimo, P., Ritter, M.G., Busick, K., Ferguson, C., Trumbull, E., & Solano-Flores, G. (2000). *Making assessment work for everyone: How to build on student strengths*. San Francisco: Regional Educational Laboratories.

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

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

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

Lee, O., Fradd, S.H., & Sutman, F.X. (1995). Science knowledge and cognitive strategy use among culturally and linguistically diverse students. *Journal of Research in Science Teaching*, 32, 797–816.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 5–21.

Lipka, J. (1991). Toward a culturally based pedagogy: A case study of one Yup'ik Eskimo teacher. *Anthropology and Education Quarterly*, 22, 203–223.

Lipka, J. (1994). Culturally-negotiated schooling: Toward a Yup'ik mathematics. *Journal of American Indian Education*, 33, 14–30.

Lipka, J. (1998). *Transforming the culture of schools: Yup'ik Eskimo examples*. Mahwah, NJ: Erlbaum.

Lipka, J. (2000). *Adapting Yup'ik elders knowledge: Pre-K to 6th math and instructional materials development*. Fairbanks, AK: University of Alaska, Fairbanks.

Magone, 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, 317–340.

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

Miles, M.B., & Huberman, A.M. (1994). *Qualitative Data Analysis* (2nd Ed.). Thousand Oaks, CA: Sage.

National Assessment of Educational Progress. (1996). *The nation's report card: Assessment science-public release*. [www.ed.gov/NCES/NAEP](http://www.ed.gov/NCES/NAEP).

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Nelson-Barber, S., & Estrin, E. (1996). *Culturally responsive mathematics and science education for native students*. San Francisco: Far West Laboratory for Educational Research and Development.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

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

Olson, J.F., & Goldstein, A.A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress*. Washington, DC: U.S. Department of Education.

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.) (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: Committee on the Evaluation of National and State Assessments of Educational Progress. National Academy Press.

Popham, W.J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16, 9–13.

Ross, M.J. (1999). *Mathematics problem solving strategies and the bilingual student: A study of strategy usage between monolingual and bilingual mathematics students*. Unpublished master's thesis, City University School of Education.

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

Shavelson, R.J. (1995). On the romance of science curriculum and assessment reform in the United States. In D.K. Sharpes & A.-L. Leino (Eds.), *The dynamic concept of curriculum: Invited papers to honour the memory of Paul Hellgren* (Research Bulletin 90). (pp. 57–76). Helsinki, Finland: University of Helsinki Department of Education.

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

Shavelson, R.J., Brown, J., Solano-Flores, G., & Ruiz-Primo, M.A. (1993, December). Development of performance assessments in science. Staff development workshop. University of California, Santa Barbara, and National Science Resources Center, Smithsonian Institution, Washington, DC.

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

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 5–8, 13, 24.

Snow, C.E. (1983). Literacy and language: relationships during the preschool years. *Harvard Educational Review*, 53, 165–169.

Solano-Flores, G. (2000). Astronomy tasks that promote geometrical, contextualized reasoning. Unpublished manuscript.

Solano-Flores, G., & Nelson-Barber, N. (1999a, March 28–31). Developing culturally-responsive science assessments. Workshop paper presented at the 1999 Meeting of the National Association for the Research of Science Teaching. Boston, Massachusetts.

Solano-Flores, G., & Nelson-Barber, N. (1999b, January 20–23). Promoting equity and fairness in testing from the start: Developing culturally-responsive assessments. Presented at the 1999 meeting of the Center for Research of Students Placed at Risk, El Paso, Texas.

Solano-Flores, G., Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1991). Science performance assessment with language minority students. Santa Barbara, CA: 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, 16–25.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2000, April). Evaluation of a model for the concurrent development of two language versions (English and Spanish) of a mathematics assessment in a bilingual program. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Steele, C. (1992, April). Race and the schooling of Black Americans. *Atlantic Monthly*, 269, 68–76.

Steele, C. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613–629.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.

Tanzer, N.K. (1999). 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: Erlbaum.

Van de Vivjer, F., & Hambleton, R.K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1, 89–99.

Van de Vivjer, F., & 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* (2nd Ed.). Vol. 1: Theory and method (pp. 257–301). Needham Heights, MA: Allyn & Bacon.

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

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

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wertsch, J.V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wertsch, J.V., Del Río, P., & Alvarez, A. (Eds.). (1995). *Sociocultural studies of mind*. New York: Cambridge University Press.