

CHAPTER ONE

Issues in Culturally Appropriate Psychoeducational Assessment

Amado M. Padilla Graciela N. Borsato

There is a long-standing debate regarding appropriate assessment strategies for use with women, the economically disadvantaged, racial/ethnic minorities, language minorities, and individuals with disabilities. Indeed, psychoeducational assessment is an area of professional practice that has been often criticized for its contribution to the perpetuation of social, economic, and political barriers confronting these various groups (Gregory and Lee, 1986; Mensh and Mensh, 1991). It has also been argued (Stobart, 2005) that fairness in assessment is, above all, a sociocultural rather than a technical issue and that the consideration of the curriculum and educational opportunities of students should therefore be included in discussions about fair assessment.

The recognition that no knowledge is acultural warrants close examination of the content we choose to assess and the extent to which it privileges the dominant culture's values, beliefs, and interests over those of other groups. In terms of educational opportunities, it is necessary to examine whether there are differences in the resources available for different groups. For example, that there is a shortage of well-qualified teachers in schools serving low-income communities is well known. Given the importance of teacher expertise to student achievement (Darling-Hammond, 2004a), it should not be surprising that students attending schools with a high proportion of underqualified teachers do not perform as well on academic achievement tests as do students from more privileged socioeconomic backgrounds. The critical issue at hand is that standardized tests are extensively used for selecting and screening. Thus,

if these tests are biased against particular groups because of curricular issues or unequal access to the knowledge the tests assess, such groups are unfairly denied access to educational and career opportunities. We fully acknowledge in this chapter that “fair assessment is inseparable from fairness in access opportunities and in what the curriculum offers” (Stobart, 2005, p. 278). Neglect of the role of the sociocultural context in which testing takes place absolutely collides with the ideal of equity in assessment.

This chapter is based on the one authored by Amado Padilla and Antonio Medina that appeared in the first edition of the *Handbook of Multicultural Assessment* and was subsequently updated by Amado Padilla for the second edition of the *Handbook*. In this third edition, we have incorporated more recent materials, but the arguments are essentially the same since issues of culturally appropriate assessment have not changed since the original chapter was written. What has evolved from the first edition is the increased importance of high-stakes testing and concerns about accountability. Together these conditions make the challenges of culturally appropriate testing more relevant than ever before. In what follows, we discuss the role of culture in test construction, administration, and use. We also highlight the consequences of poor testing practices and offer recommendations for using assessment instruments and procedures in ways that are culturally sensitive.

TEST CONSTRUCTION

We take the position in this chapter that psychoeducational assessment is made culturally appropriate through a continuing and open-ended series of substantive and methodological insertions and adaptations designed to mesh the process of assessment and evaluation with the cultural characteristics of the group being studied. The insertions and adaptations must span the entire assessment and evaluation process, from the development or adaptation of instruments, including translation, to the administration of the measure, and ultimately to the appropriate scoring and interpretation of the outcomes. Thus, assessment is made culturally appropriate through a continuing, intentional, and active preoccupation with the culture of the group or individual being assessed.

The search for culturally fair assessment strategies has long been hampered by the lack of consensus on the definition of *culture*. As Frisby (1998) points out, the common practice of equating culture with geographical, racial, and ethnic differences, albeit expedient, is entirely arbitrary. There are other culturally relevant constructs, such as English-language proficiency, age of migration, generational status, length of residence in the United States, ethnic identity, and extent of acculturation, that have not traditionally been viewed

as significant concerns in test theory or development and yet are likely to function as moderators of psychoeducational assessment. In terms of extent of acculturation, the more acculturated the person is to the host society, the greater the likelihood is that standardized norms can be applied with confidence to that individual (Cuéllar, 2000). With regard to language proficiency, Geisinger (2003) argues that when nonnative English speakers are tested, each individual's relative language abilities in the first and second languages need to be considered before deciding whether to test in English or the first language. Sometimes the markers to guide test selection are not readily apparent, however. Consider a Latino adolescent who to a majority group test examiner appears acculturated but who nonetheless is more adept in Spanish than in English (Padilla, 1992). Such an adolescent will likely have more difficulty on a speed test (limited time) if it is administered in English than in an assessment situation where time is not a critical element in the test administration.

At the very least, if it is not possible to eliminate cultural influences on standardized test scores, specific cultural variables should be considered in order to interpret findings from these tests (Anastasi, 1988, as cited in Cuéllar, 1998). Consistent with Anastasi's recommendation and noting that the impact of acculturation on performance in standardized intelligence measures is well documented in the psychological assessment literature, Gopaul-McNicol and Armour-Thomas (2002) suggest using an acculturation instrument in conjunction with psychometric tests of intelligence so as to obtain a more appropriate measure of cognitive ability. This is a sensible recommendation, yet there is to date no consensus as to how most effectively to measure acculturation (Cabassa, 2003).

PSYCHOMETRIC CONSIDERATIONS

Many tests used in education are norm referenced. In norm-referenced testing, a person's test score is compared to the scores of a group of people who have already taken the same test, called the norming group. As Sireci (2005) points out, the utility of norm-referenced tests depends on how recent the norms are and the extent to which the norm group is representative of the population with which the test is currently being used. Psychologists have long argued that instruments normed on majority group populations cannot be indiscriminately used with individuals who differ from the normative population. The reason is that the validity (the extent to which the test measures what it claims to measure) and reliability (the extent to which the test is internally consistent or gives consistent results over time) of a norm-referenced test when used with individuals whose cultural or linguistic backgrounds were not adequately represented in the norming group are questionable. The implication is that the

test performance of an individual who comes from a nondominant cultural background or is lower in social status may be affected in ways not intended by the test creator, unless this individual's group was included in the norming of the test.

Whereas in norm-referenced tests a student's test score is compared to the performance of similar students, in criterion-referenced tests a student's test performance is compared with carefully designed standards of expected performance (Sireci, 2005). The issue is that those standards often reflect the values of the dominant culture as to what constitutes acceptable language, behavior, and valued knowledge. As Scheuneman and Oakland (1998) point out, the competencies and cognitive abilities of children raised in a minority culture are often different from those required by the tests, and therefore these children may not perform as well in those tests as children from the dominant culture. Scheuneman and Oakland add that in general, "observed performance differences suggest that the greater the similarity of an individual's socioeconomic and cultural background to that of the majority population, the better the test performance will be" (p. 86). Although this is a well-established fact in the psychometric literature, the pressure associated with accountability and test use often results in failure to acknowledge it when interpreting differences in test outcomes.

Questions about the reliability and validity of aptitude and achievement tests were discussed by a Latino psychologist more than seventy years ago (Sanchez, 1932a, 1932b), but little attention was given to Sanchez's critique at that time despite his intimate knowledge of Latino culture and the fact that he published his critiques in reputable journals of the day. We might ask why, if Sanchez raised questions about the lack of cultural sensitivity of tests over seventy years ago and such concerns have continued to the present, so little attention has been given to these problems. The answer to this rhetorical question probably is that people of color have not had the political clout, in either society generally or the field of psychological assessment specifically, to insist that their concerns be taken seriously (Padilla, 1992).

TRANSLATION AND ADAPTATION OF TESTS

Over the years, numerous scholars have criticized the administration of tests in English to language-minority children because of issues of validity. Indeed, research results indicate that test scores of English learners are substantially lower than those of native English speakers and that as the language demands of individual test items decrease, so does the achievement gap between English-learning and English-proficient students (Abedi, Hofstetter, and Lord, 2004). Given the pitfalls of testing language-minority individuals in English,

consideration has been given to translating tests from a source language (English) into a second target language. However, as Sandoval and Durán (1998, p. 195) point out, this practice is “fraught with hazards and issues,” the most salient of which is the difficulty of maintaining construct equivalence when tests are translated (Abedi et al., 2004).

Geisinger (1994) advanced a set of guidelines for translating and adapting a test to a new culture and language population:

- Individuals translating or adapting an assessment instrument should not only be fluent in the two languages, but also fully knowledgeable about both cultures, the content that the instrument purports to assess, and the uses to which the instrument will be put.
- A group of individuals who meet the same rigorous requirements as the translators should review the quality of the translation or adaptation, and any recommendations for change should be incorporated subsequently.
- The instrument should be pilot- and field-tested.
- Scores need to be standardized and necessary validation research conducted.

As these guidelines suggest, if done appropriately, translating tests can be difficult, time-consuming, and very expensive. These are likely the reasons that wholesale translation of tests has not generally been found to be an acceptable solution to the pervasive problem of inappropriate assessment of English-language learners. Furthermore, research indicates that assessments using students’ first language should be used only when students have received the corresponding content instruction in their first language. In other words, the language of testing should be the same as the language of instruction (Abedi et al., 2004), but proficiency in English must match the linguistic proficiency demands of the content instruction in English. The mismatch between the two proficiencies is what frequently leads to misinterpretation of test outcomes.

TEST ADMINISTRATION: ASSESSMENT ACCOMMODATIONS

In the past, students with disabilities and English-language learners were frequently excluded from participation in large-scale assessment programs (Abedi et al., 2004; Thurlow, Lazarus, Thompson, and Morse, 2005). In the case of English-language learners, the concern was of course that poor English-language proficiency could be a confounding influence in the measurement of academic achievement. As for students with disabilities, several factors have accounted for their exclusion, but a major reason has been that the highly standardized ways in which tests need to be administered in order to

be meaningful are not always accessible to these students (Bolt and Thurlow, 2004). Recently we have seen an increased emphasis on accountability to improve student performance. Because of the 2001 No Child Left Behind (NCLB) Act, states' educational agencies are responsible for developing high academic standards and implementing assessment systems to monitor whether students are meeting those high standards (Abedi et al., 2004). The exclusion of English-language learners or students with disabilities from assessment programs would prevent these students from taking advantage of the potential positive effect of accountability practices. Therefore, under the NCLB Act, it is no longer acceptable to exclude them from mandated testing. States must include all students in their accountability systems and provide adaptations or accommodations for students with disabilities and students with limited English proficiency so the inclusion requirements can be met (Thurlow et al., 2005; Abedi et al. 2005). Test accommodations refer to changes in test format or administration that allow students with disabilities and English-language learners to demonstrate their content knowledge without giving them an advantage over students who are not receiving the accommodated assessments (Abedi et al., 2004).

In a recent review of research on test accommodation strategies for English learners, Abedi et al. (2004) cautioned against a one-size-fits-all approach given that some accommodations are more effective with some students than with others, depending on factors such as length of time in the United States and English reading proficiency. In addition, Abedi et al. found that modifying the language of test items to reduce low-frequency vocabulary and complex language structures that are incidental to the content knowledge being assessed and allowing the use of customized dictionaries constitute effective accommodation strategies. Moreover, these strategies have been found to help English-language learners while not affecting the scores of English-proficient students. As Abedi et al. point out, there is no reason, then, that these strategies cannot be used with all students. A parallel review of research conducted by Bolt and Thurlow (2004) on five frequently allowed test accommodations used with students with disabilities (dictated response, large print, braille, extended time, and sign language interpreter for instructions) revealed "mixed support and nonsupport" for providing these accommodations to students with disabilities.

Reviews of literature by Abedi et al. (2004) and Bolt and Thurlow (2004) reveal that empirical research on test accommodations is scarce. Given the emphasis on testing and accountability and the mandate to include all students in assessment programs, the urgency for research examining the appropriateness and effectiveness of accommodation strategies cannot be overstated. In the meantime, as Abedi et al. point out, "for any group of students, accommodations must be administered, used, and interpreted cautiously" (p. 8).

HIGH-STAKES DECISIONS

In today's test-conscious environment where tests have acquired such prominence for diagnosis, selection, certification, and accountability, it is vitally important to contextualize our discussion in the language of low- and high-stakes decision making. In low-stakes testing, there are typically no consequences associated with performance. For example, weekly classroom quizzes to determine what the student knows are low-stakes testing. Results from these quizzes are used to inform subsequent instructional practices to promote and support learning. High-stakes testing carries important and often long-term consequences for the individuals being assessed. Scores from high-stakes tests play an important role in decisions made with regard to academic placement, scholarship awards, certification, and professional and graduate school entry. An example of high-stakes testing is the use of a test outcome to determine whether a student will receive a diploma following thirteen years of public school attendance. High school exit exams have received considerable public attention and best illustrate what is meant by a high-stakes test.

Identical treatment, the condition most consonant with accurate prediction, presupposes access to the same experiences; this is especially true of cognitive type testing (Williams, 1983). It can be argued that this prerequisite is not met in our educational system or in most aspects of daily life where minority individuals are concerned. For example, although attendance of early education programs has increased significantly in the United States, enrollment rates vary substantially by race/ethnicity, poverty status, and maternal education (Karoly and Bigelow, 2005). This enrollment disparity is of great concern given the benefits to children's academic outcomes associated with participation in high-quality preschool programs (Magnuson, Meyers, Ruhm, and Waldfogel, 2004). It is important to keep in mind that once test scores become numbers in a file, they provide the basis for high-stakes decisions concerning selection, placement, and promotion that are made without consideration of potential inequities imposed by the original testing situation (Lam, 1993).

Notwithstanding the criticism that their use is incompatible with the goal of equity, standardized tests are used regularly to make key educational decisions, such as placement in special education classes, assignment to low-track or high-track classes, and grade retention (Gopaul-McNicol and Armour-Thomas, 2002). With regard to special education, Losen and Orfield (2002) cite "unjustifiable reliance on IQ and other evaluation tools" (p. xviii) as one of the factors contributing to the overrepresentation of minority children in special education classes. Padilla (1988) warned against the use of assessment instruments in English to diagnose learning disabilities of English-language learners. Of note is that numerous lawsuits centered on this issue were brought

against the educational establishment. In one of the early legal cases, *Diana v. California State Board of Education* (1970), the suit questioned the practice of testing children in English when this was not their first or dominant language. Padilla noted that although this case was settled out of court and in favor of the student, the ruling has had little impact on professional practice and placement decisions. Indeed, recent results (Artiles, Rueda, Salazar, and Higuera, 2002) indicate that Hispanic students identified as English-language learners are still disproportionately placed in special education in the upper grades. The overrepresentation of minority students in special education is troubling, especially taking into account that students in special education are more likely to experience lower teacher expectations and to drop out of school at higher rates than their nondisabled peers (Ferri and Connor, 2005).

Standardized tests are also used to make decisions about grade retention (Gopaul-McNicol and Armour-Thomas, 2002). The use of high-stakes testing in grade retention has been called into question by some educational specialists. For example, Darling-Hammond (2004b) reports that in 1997 and 1998, more than twenty thousand students were retained in grade in Chicago under a policy requiring test passage at grades 3, 6, and 8. Darling-Hammond also points out that an evaluation of the subsequent academic performance of these students confirmed past research findings indicating that retaining students not only does not help them to catch up with their peers, but also contributes to high rates of academic failure and behavioral problems.

It is common practice in education to use assessment procedures to track students into vocationally or academically oriented classes (Oakes, 1985). Not surprisingly, high school minority students are disproportionately found in lower academic, non-college-oriented tracks (Mickelson, 2003). Proponents of tracking argue that achievement tests are necessary for determining which students have the relevant knowledge and to what degree, and who is thus adequately prepared to complete college-oriented, high-track classes (Loveless, 1999). Critics of tracking (such as Heubert, 2001) contend that as tracking is currently practiced, the emphasis in lower class tracks is on basic skills, and these classes are often taught by the least qualified teachers. Oakes (2000) further argues that minority and White children are exposed to different curricula through the practice of ability tracking, which begins as early as elementary school. Thus, poor learning environments lead directly to poor performance on achievement tests, which are used in turn to make other high-stakes decisions, such as high school graduation and admission to college.

In the area of college admissions, scores on a test of scholastic achievement measuring developed ability to reason with words and numbers to solve college-like problems inform the user about how much trouble a student may expect to have in moving immediately into college work (Samuda, 1998). However, as Samuda points out, these instruments and their outcomes say

nothing about the odds against which the student has had to struggle in developing those particular abilities or about the energy and determination the individual will put into college work. A student's ability to solve problems posed in a different language or sociocultural context may or may not be reflected in the scores, depending on how widely divergent the two cultures at issue may be. Equally important from a psychometric perspective is that the predictability of success in college as determined by grades and completion based on tests such as the Scholastic Assessment Test (SAT) is rarely above 0.2 to 0.3 (Zwick and Sklar, 2005). Nonetheless, the importance of the SAT over many decades has taken on a life of its own, as witnessed by the uproar when the president of the University of California, Richard Atkinson, a distinguished psychologist and a member of the National Academy of Science, challenged the utility of the SAT (see Cloud, 2001).

For a student whose schooling has been of lower quality or whose home and perhaps predominant community language is other than English, a mediocre score on a test may represent a triumph of ability, devotion to study, and persistence. The test scores are not designed to reflect these characteristics, not because these qualities themselves are unimportant but because testing has not yet produced ways to measure them. In educational assessment and selection practices, the student's holistic profile, including biographical record, demonstrated interest, and long-term perseverance as reflected in school grades, and especially as illuminated by the comments of those who know his or her history, are the indispensable bases for understanding the meaning of the scores resulting from standardized tests.

One other aspect that deserves consideration is that for members of any group about whom a negative academic stereotype exists, assessment situations make them vulnerable to stereotype threat: "the event of a negative stereotype about a group to which one belongs becoming self-relevant, usually as a plausible interpretation for something one is doing, for an experience one is having, or for a situation one is in, that has relevance to one's self-definition" (Steele, 1997, p. 616). Stereotype threat affects students who have gained skills and confidence in a particular domain for which their group is negatively stereotyped, such as school-identified African Americans or math-identified women. Steele and Aronson (1995) showed that inducing stereotype threat, by asking students to indicate their race before a test or by telling them that the test is a measure of intellectual ability, can undermine the test performance of African Americans, even though they have demonstrated their competency in numerous other ways prior to the test. Subsequent research extended these results to other groups, such as students from low socioeconomic backgrounds (Croizet and Claire, 1998) and female students in the domain of math (Spencer, Steele, & Quinn, 1999). Steele (1997) offers some general strategies to reduce stereotype threat, including teacher-student relationships that contribute to

discredit negative stereotypes, offering challenging school work rather than remediation, and stressing the expandability of intelligence in response to experience and training.

Testing environments such as Advanced Placement examinations, the SAT, and the Graduate Record Exams are characterized by the two conditions that contribute to stereotype threat: evaluative scrutiny and identity salience (Good, Aronson, and Inzlicht, 2003). Indeed, students without a doubt recognize that their performance in these tests can have important and long-term implications for their future. Also, they are often asked to indicate their gender and race/ethnicity prior to taking the test, and they take the test in the presence of White students and males. Students' group identity is therefore made salient in these test situations. As Natriello and Pallas (2001) point out, the threat of conforming to race-based stereotypes about academic ability may depress the academic performance of racial/ethnic minorities and other stigmatized groups. In this sense, high-stakes tests can potentially contribute to "exacerbate already substantial inequities in schooling outcomes" (p. 37).

High-stakes tests are prominent not only in the field of education but also in other arenas, such as mental health, forensics, and employment. Gray-Little and Kaplan (1998) point out that in clinical settings, misdiagnoses, whether based on interview or psychological testing, can have serious consequences for the client. Indeed, misdiagnosed clients are not likely to receive appropriate treatment, and if psychotropics are prescribed, such individuals are exposed to needless risks. Also according to Gray-Little and Kaplan, a review of the literature on the psychiatric diagnosis of ethnic minorities indicates that sometimes race and ethnicity are predictive of diagnosis independent of symptoms and that diagnosticians are subject to bias in clinical judgments. Furthermore, the expression of distress and psychopathology varies among racial and ethnic groups in the United States, and these variations may affect the interpretation of psychological tests and the accuracy of diagnosis.

In the area of forensic assessment, decisions based on testing instruments can have a large impact on matters such as legal responsibility and personal freedom. The consideration of sociocultural factors when interpreting the results of forensic tests is therefore crucial. As an example, in the context of urban Black families, a biological parent's deficits in resources or skills may be compensated for by the availability of an extended kinship network very much invested in caring for the child. As Gray-Little and Kaplan (1998) point out, a lack of cultural awareness in this case could potentially lead to the child's being removed without cause from a functional and nurturing family.

Employment tests are frequently used for purposes of selection and promotion. As Sireci and Geisinger (1998) point out, "The decisions made on the basis of scores derived on employment tests often mean the difference between work and unemployment or between upward mobility and stagnation"

(p. 105). Latinos are becoming more prominent in the workforce due to their youthfulness relative to the aging majority population, higher fertility rates, and increased immigration levels (Toossi, 2004). According to Toossi, Asians are also expected to have a large increase in the labor force participation rate over the next few years, reflecting especially the increasing number of Asian women entering the workforce. These demographic trends highlight the need to consider seriously issues inherent in the psychological assessment and evaluation of Latinos and Asian Americans in order to ensure their educational and occupational opportunity and success. In so doing, the social well-being of all Americans will be fostered.

RECOMMENDATIONS FOR NONBIASED ASSESSMENT PRACTICES

There are several ways in which tests may be biased. For example, the very content or language of test items may be biased in the sense that they give unfair advantage to one group over another. Several methods for detecting differential item functioning (DIF) across contrasting groups of test takers have been devised. As Kristjansson, Aylesworth, McDowell, and Zumbo (2005) point out, systematic application of DIF detection methods followed by expert review contributes to validity and equity in testing. These should therefore be key steps in test development and validation. Another type of bias, called *construct bias*, occurs when the construct being measured does not have the same meaning across groups. The standard method to identify construct bias is to examine the factor structure of the test under consideration with different subgroups. The presence of items that have different meanings and relationships with other items for different groups signals construct bias (Skiba, Knesting, and Bush, 2002). To prevent construct bias, extensive knowledge of the different populations with whom a test is to be used is crucial during test development and test adaptation or translation.

Bias can also result from incidental features such as mode of test administration or even examiner personality factors that favor one group of examinees over another. This type of bias, commonly referred to as *method bias*, can be prevented, or at least ameliorated, through extensive training of test administrators (van de Vijver and Phalet, 2004). Finally, bias may occur due to inappropriate application, which results in the identification of one set of applicants over others. This type of bias is present when tests that are not valid for tracking or promotion are nonetheless used for those purposes or when students do not have the opportunity to learn the knowledge and skills that the high school exit exam measures. It is important to remember that use of a test is appropriate only to the extent to which the inferences derived from the test scores and the actions that follow are appropriate (Heubert, 2001).

One of the most publicized approaches to nonbiased assessment was the use of differential norms, such as in the System of Multicultural Pluralistic Assessment (SOMPA; Mercer, 1979). In this system, a child's score on the Wechsler Intelligence Scale for Children (WISC) is "corrected" based on acculturation and other sociocultural variables. Typically scores of European American children are shifted downward, while the scores of Mexican American and African American children are shifted upward (van de Vijver and Phalet, 2004). Although SOMPA constitutes a valuable attempt to deal with the issue of acculturation in assessment, it has been criticized for its poor performance in predicting future academic performance (Sandoval and Durán, 1998; van de Vijver and Phalet, 2004). Additional approaches to address acculturation in multicultural assessment, including cutoff scores, acculturation as moderator, standardization or centering, item response theory, method factors, and "person-fit tradition," are discussed in detail by van de Vijver and Phalet. We agree with these researchers that given the increasingly multicultural nature of our society and the pervasive influence of acculturation on behavior, the need to consider the extent of acculturation in assessment situations can no longer be ignored.

Having the competence necessary for becoming culturally sensitive in assessment procedures is not an easy task. Psychologists who employ tests as part of their professional responsibility may be unaware of how their ethnic and cultural experiences and position in mainstream society influence their selection of particular tests and the interpretations they derive from psychological instruments. Although the situation is improving, many psychologists are not trained in nonbiased assessment, and as a result, they know little about procedures for evaluating students from diverse backgrounds. Geisinger and Carlson (1998) contend that to assess members of diverse groups, it is necessary to understand psychometric concepts such as test bias and test fairness. They add that during their training to use psychological assessment, students must read test manuals carefully and determine whether and how a test may be properly used with members of different groups. This is especially critical if the person or group to be tested is not similar to the group on which the test was normed. Future psychologists should also be encouraged to consult with others when assessing an individual from a cultural group with which they have limited or no experience.

In order for test examiners to increase their cross-cultural assessment competency, they must become knowledgeable and comfortable with the traditional customs and communicative styles of many individuals who do not represent the prototypical middle-class person on whom most assessment instruments are based. We recommend that test users involve minority community members

in selecting instruments to be used in a school, employment venue, placement center, and so forth. This practice increases the minority community's trust and rapport regarding testing practices and results in more appropriate assessment measures, practices, and decision making. Ultimately complex judgments concerning appropriate and equitable test use can best be made by users who are familiar with those being assessed and the environment in which the test is administered (Lam, 1993).

Research emphasis is usually placed on a comparative approach that uses similar measures to compare groups of people who differ in culture, language, or social class. We believe that a paradigm shift is required whereby the study of a specific group is valued for its own sake and need not be compared to another group, especially if the comparison is likely to be biased. Instruments that are biased and favor a particular group should not be used to evaluate differences between culturally distinct groups of people. Furthermore, instruments must also be appropriate for accurately assessing changes in learning or behavior that are due to a treatment or educational program. However, if assessment devices are inappropriate in a pretest context, they will also be poor measures of postintervention learning or behavior changes.

Test makers and users need to be aware of how test performance is influenced by inequality in educational and economic opportunity, parents' educational attainment, cultural orientation, language spoken at home, proficiency in English, socialization experiences, occupational status and income of wage earners, and level of motivation to do well. When sufficient information is given beforehand about possible confounding variables in deciding to test a particular individual or group, a more informed decision can be made about the suitability of the test to be used.

CONCLUSION

It is important to sensitize professionals to discriminatory practices while broadening assessment methods. In advocating for a systems approach that is culturally sensitive, it is crucial that we redouble our efforts to increase the pool of qualified minority psychologists who are trained in psychometric theory and test construction. Furthermore, we need to train individuals who are expert in psychological assessment to assume leadership positions in the field. There are too few psychologists with the expertise necessary to advance the discussion of culturally sensitive assessment beyond that which has prevailed for the past three decades. In this new millennium, we look forward to assessment practices that better reflect the cultural face of America.

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.
- Artiles, A. J., Rueda, R., Salazar, J. J., & Higuera, I. (2002). English language learner representation in special education in California urban school districts. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. 117–136). Cambridge, MA: Harvard Education Press.
- Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy. *Remedial and Special Education, 25*(3), 141–152.
- Cabassa, L. J. (2003). Measuring acculturation: Where we are and where we need to go. *Hispanic Journal of Behavioral Sciences, 25*(2), 127–146.
- Cloud, J. (2001). Should SATs matter? *Time Magazine*. Retrieved June 27, 2006, from <http://www.time.com/time/nation/article/0,8599,101321-1,00.html>.
- Croizet, J., & Claire, T. (1998). Extending the concept of stereotype and threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*(6), 588–594.
- Cuéllar, I. (1998). Cross-cultural clinical psychological assessment of Hispanic Americans. *Journal of Personality Assessment, 70*(1), 71–86.
- Cuéllar, I. (2000). Acculturation as a moderator of personality and psychosocial assessment. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 113–129). Mahwah, NJ: Erlbaum.
- Darling-Hammond, L. (2004a). Inequality and the right to learn: Access to qualified teachers in California's public schools. *Teachers College Record, 106*(10), 1936–1966.
- Darling-Hammond, L. (2004b). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047–1085.
- Ferri, B. A., & Connor, D. J. (2005). In the shadow of *Brown*: Special education and overrepresentation of students of color. *Remedial and Special Education, 26*(2), 93–100.
- Frisby, C. L. (1998). Culture and cultural differences. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 3–16). Washington, DC: American Psychological Association.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.
- Geisinger, K. F. (2003). Testing students with limited English proficiency. In J. E. Wall & G. H. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 147–159). Greensboro, NC: CAPS Press.
- Geisinger, K. F., & Carlson, J. F. (1998). Training psychologist to assess members of a diverse society. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, &

- J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 375–386). Washington, DC: American Psychological Association.
- Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24*, 645–662.
- Gopaul-McNicol, S., & Armour-Thomas, E. (2002). *Assessment and culture: Psychological tests with minority populations*. Orlando, FL: Academic Press.
- Gray-Little, B., & Kaplan, D. A. (1998). Interpretation of psychological tests in clinical and forensic evaluations. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 141–178). Washington, DC: American Psychological Association.
- Gregory, S., & Lee, S. (1986). Psychoeducational assessment of racial and ethnic minority groups: Professional implications. *Journal of Counseling Psychology and Development, 64*(10), 635–637.
- Heubert, J. P. (2001). High-stakes testing and civil rights: Standards of appropriate test use and a strategy for enforcing them. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 179–194). New York: Century Foundation Press.
- Karoly, L. A., & Bigelow, J. H. (2005). *The economics of investing in universal preschool education in California*. Santa Monica, CA: RAND Corporation.
- Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935–953.
- Lam, T.C.M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179–191.
- Losen, D. J., & Orfield, G. (2002). Introduction. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. xv–xxxvii). Cambridge, MA: Harvard Education Press.
- Loveless, T. (1999). *The tracking wars: State reform meets school policy*. Washington, DC: Brookings Institution Press.
- Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41*(1), 115–157.
- Mensh, E., & Mensh, H. (1991). *The IQ mythology: Class, race, gender, and inequality*. Carbondale: Southern Illinois University Press.
- Mercer, J. R. (1979). *Technical manual: System of Multicultural Pluralistic Assessment (SOMPA)*. New York: Psychological Corporation.
- Mickelson, R. A. (2003). When are racial disparities in education the result of racial discrimination? A social science perspective. *Teachers College Record, 105*(6), 1052–1086.

- Natriello, G., & Pallas, A. M. (2001). The development and impact of high-stakes testing. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (p. 38). New York: Century Foundation Press.
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.
- Oakes, J. (2000). Grouping and tracking. In A. E. Kazdin (Ed.), *Encyclopedia of psychology*. Washington, DC: American Psychological Association.
- Padilla, A. M. (1988). Early psychological assessment of Mexican-American children. *Journal of the History of the Behavioral Sciences, 24*, 113–115.
- Padilla, A. M. (1992). Reflections on testing: Emerging trends and new possibilities. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 271–284). Washington, DC: American Psychological Association.
- Samuda, R. J. (Ed.). (1998). *Psychological testing of American minorities: Issues and consequences* (2nd ed.). Thousand Oaks, CA: Sage.
- Sanchez, G. I. (1932a). Group differences in Spanish-speaking children: A critical review. *Journal of Applied Psychology, 40*, 223–231.
- Sanchez, G. I. (1932b). Scores of Spanish-speaking children on repeated tests. *Journal of Genetic Psychology, 40*, 223–231.
- Sandoval, J., & Durán, R. (1998). Language. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 181–211). Washington, DC: American Psychological Association.
- Scheuneman, J. D., & Oakland, T. (1998). High-stakes testing in education. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 77–103). Washington, DC: American Psychological Association.
- Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 111–121). Mahwah, NJ: Erlbaum.
- Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 105–140). Washington, DC: American Psychological Association.
- Skiba, R. J., Knesting, K., & Bush, L. D. (2002). Culturally competent assessment: More than nonbiased tests. *Journal of Child and Family Studies, 11*(1), 61–78.
- Spencer, S., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4–28.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629.

- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797–811.
- Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education*, 12(3), 275–287.
- Thurlow, M. L., Lazarus, S. S., Thompson, S. J., & Morse, A. B. (2005). State policies on assessment participation for students with disabilities. *Journal of Special Education*, 38(4), 232–240.
- Toossi, M. (2004). Labor force projections to 2012: The graying of the U.S. workforce. *Monthly Labor Review*, 127(2), 37–57.
- van de Vijver, F.J.R., & Phaet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review*, 53(2), 215–236.
- Williams, T. S. (1983). Some issues in the standardized testing of minority students. *Journal of Education*, 165, 192–208.
- Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42, 439–464.