The Handbook of Chicana/o Psychology and Mental Health

Edited by

Roberto J. Velásquez
San Diego State University

Leticia M. Arellano
University of La Verne

Brian W. McNeill
Washington State University

LEA
LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2004 Mahwah, New Jersey London
Integrating a Cultural Perspective in Psychological Test Development

Steven R. López
University of California, Los Angeles

Amy Weisman
University of Miami

The purpose of this chapter is to review and evaluate critically the development of psychological tests for Mexican Americans and Latinos residing in the United States. To accomplish this objective, we first present an anthropologically informed definition of culture, one that respects the dynamic and social nature of culture (Lopez & Guarnaccia, 2000). The implication of this contemporary definition is that test developers, researchers, and examiners respect the changeable nature of culture and look to the social world to ascribe meaning to research findings and test results. We then review selected advances in test development as they relate to Latina/os in general and Mexican Americans in particular. Our main interest is to identify how culture is considered and to evaluate such considerations. We limit the scope of this chapter to the assessment of cognitive-intellectual functioning and psychodiagnosis because they represent central aspects of psychological assessment and are among the most widely studied. Based on the critical evaluation, we then recommend ways in which test development can better incorporate a cultural perspective.

AN ANTHROPOLOGICALLY INFORMED DEFINITION OF CULTURE

We draw on Kleinman’s (1995) theoretical notion of experience as the basis of our definition of culture. Experience is viewed as the “felt flow” of the intersubjective space between individuals (their minds and bodies) and their social world. In this theoretical space, the individual and his or her social world are interconnected and inseparable. For practical purposes this intersubjective medium is what is at stake or what matters for individuals and groups. Kleinman pointed out that “preservation of life, aspiration, prestige, and the like” (p. 97) is relevant for all, but it is that which
is at stake in peoples’ daily lives that is tied closely to culture. He argued that the cultural scientist’s central concern then is to interpret what matters for specific individuals in specific situations.

The notion that culture is associated with what is at stake for people in specific local worlds advances the cross-cultural psychology definition of culture as values, beliefs, and practices (Berry, Poortinga, Segall, & Dasen, 1992; Betancourt & López, 1993) in significant, interrelated ways. First, the scientist or examiner must look to the community of interest to determine what is at stake. Anthropological investigators do not typically apply existing measures of values, beliefs, or norms, nor do they infer what is at stake given the ethnic group under study. Instead, they observe the social world within specific settings to discern what matters. Second, researchers applying this anthropologically informed definition are not necessarily tied to studying specific ethnic or so-called cultural groups. Instead, researchers are free to determine what aspects of the community are relevant to identifying what is at stake; these could be other social categories such as age or gender, as well as particular experiences of a group of individuals with a shared history. Third, it is important to consider interconnected systems within social contexts—the historical, the political, and the economic circumstances in particular settings. Such considerations are likely to contribute to a rich analysis of the sociocultural context and avoid cultural glosses that reflect superficial analyses. For example, spirituality may be as much a function of individuals’ adjustment to longstanding economic and social difficulties as it is a set of values and beliefs passed from generation to generation. In all, the definition of culture associated with what is at stake for individuals and groups in local contexts represents an open construct that is capable of capturing culture’s richness and complexity.

Culture conceived as an open construct is not without limitations. Some may view this definition as too abstract and diffuse to be of value to standardized testing (see Segall, 1984, for a similar critique). We acknowledge the challenge in bringing interpretive anthropology and classical test theory into the same conceptual and applied space. However, we believe that there are ways in which psychological assessment, particularly test development, can move toward integrating a cultural perspective, particularly one that respects the rich and dynamic qualities of culture.

TEST DEVELOPMENT

The consideration of what is at stake for individuals in their local worlds is probably best addressed within the clinical domain as examiners collect and interpret data. Test developers, however, can take steps to facilitate the consideration of what matters most for specific groups. We examine the ways test developers take into account or fail to take into account ethnicity and culture in the development of selected cognitive-intellectual and psychodiagnostic tests used in the assessment of Mexican Americans and other Latina/o groups within the United States. (For related reviews, see Cervantes & Acosta, 1992; Dahlstrom, Lachar, & Dahlstrom, 1986; Geisinger, 1992; and McShane & Cook, 1985.) The current analysis is based largely on the review of test manuals and technical manuals of selected tests as they reflect the actual test development. The only exception is the Minnesota Multiphasic Personal-
<table>
<thead>
<tr>
<th>#</th>
<th>Year Published</th>
<th>Publisher</th>
<th>Country/ Locale of Norming Sample</th>
<th>% Latinos Norming Sample</th>
<th>Reliabilities Reported for Latinos</th>
<th>Validities Reported for Latinos</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Split-Half SEM</td>
<td>Temporal Stability</td>
</tr>
<tr>
<td>1</td>
<td>1989</td>
<td>NCS</td>
<td>U.S.</td>
<td>2.8%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>1997</td>
<td>NCS</td>
<td>U.S.</td>
<td>2.8%</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>in press</td>
<td>U. Minnesota</td>
<td>Puerto Rico, Miami, Los Angeles</td>
<td>100.0%</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2001</td>
<td>Mexico</td>
<td>Mexico</td>
<td>100.0%</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>Year Published</td>
<td>Publisher</td>
<td>% Latinos in Norming Sample</td>
<td>Reliabilities Reported for Latinos</td>
<td>Validities Reported for Latinos</td>
<td>Comments</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>----------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Country/Locale of Norming Sample</td>
<td>Split-Half SEM Stable</td>
<td>Temporal Stability</td>
</tr>
<tr>
<td><strong>Intelligence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAIS-III</td>
<td>1999</td>
<td>Psych Corp</td>
<td>U.S.</td>
<td>7.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WISC-III</td>
<td>1998</td>
<td>Psych Corp</td>
<td>U.S.</td>
<td>11.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIWN-R</td>
<td>1983</td>
<td>Psych Corp</td>
<td>U.S.</td>
<td>not reported</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WISC-RM</td>
<td>1984</td>
<td>Manual Moderno</td>
<td>Mexico City</td>
<td>100.0%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>EIWN-RPR</td>
<td>1992</td>
<td>Psych Corp</td>
<td>Puerto Rico</td>
<td>100.0%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>WAIS-III Español</td>
<td>2001</td>
<td>Manual Moderno</td>
<td>U.S.</td>
<td>7.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIWA</td>
<td>1967</td>
<td>Psych Corp</td>
<td>Puerto Rico</td>
<td>100.0%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Batería-R</td>
<td>1996</td>
<td>Riverside</td>
<td>U.S.</td>
<td>9.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonverbal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAMA</td>
<td>1997</td>
<td>NCS</td>
<td>U.S.</td>
<td>7.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNIT</td>
<td>1998</td>
<td>Riverside</td>
<td>U.S.</td>
<td>13.0%</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
and Hispanic examinees, and (c) the standardization sample of the WAIS-III plus an oversampling of 200 African American and Hispanic participants. Test items that were found to be biased or potentially biased by these subjective and objective analyses were reported as not included in the test. Reliability and validity data were not reported by ethnicity or race, however.

An important change in the WAIS-III is that the scale score for each subtest is based on the scores of a nonimpaired reference group by specific age group. In other words, to determine how a given individual's performance compares to others, the examiner compares that score to others from the same age group. This differs from the earlier Wechsler adult tests of intelligence for which only one age group—from 20 to 34—served as the reference group for all age groups. The problem with this single reference group is that older persons who may be functioning within the normal range for their age group appear to be impaired when compared to the 20- to 34-year-old reference group (Ryan, Paolo, & Brungardt, 1990). To avoid this problem, the scale scores in the WAIS-III now adjust for the age of the examinee. The importance of this change is that the test developers are directly considering for the first time in adult assessment a contextual factor—age and development—in determining a person's level of functioning. This advancement has implications for the consideration of other factors for which specific subgroup norms may be helpful.

The development of the Wechsler Intelligence Scale for Children (WISC) follows the same general pattern as that of its adult assessment counterpart. The original WISC manual (Wechsler, 1949) made no mention of race or ethnicity in the sampling design or standardization sample, whereas the WISC-III (Wechsler, 1991) addressed race and ethnicity in a very similar way to the WAIS-III with regard to item bias and sampling. Latina/os represented 11% of the standardization sample, which compares favorably to their percentage within the 1988 U.S. population of children ages 6 through 16—10.8% (Wechsler, 1991). In addition, only English-speaking children comprised the sample.

In summary, the recent English-language Wechsler tests of cognitive-intellectual abilities take into account ethnicity through their sampling and assessment of item bias. However, the tests' reliability and validity for Latina/os and for limited English-speaking individuals were not reported in the technical manuals.

Spanish-Language Wechsler Scales. There are a number of Spanish translations and adaptations of the Wechsler scales. The earliest translation that we know of is the Escala de Inteligencia Wechsler para Niños (EIWN; Wechsler, 1951), which was developed by Pablo Roca for the assessment of children on the island of Puerto Rico. Three studies with small convenience samples (Ns = 18, 41, and 69) were carried out to begin to examine test properties, primarily the level of difficulty of subtest items and the sample's mean level of functioning. Examiners who used this Spanish translation applied the U.S. English-language norms, as no norms were associated with the Spanish-language translation and adaptation (Herrans, 1985).

Several years later, at least three Spanish-language versions of the WISC-R (Wechsler, 1974) were developed: Escala Wechsler para Niños-Revisada (EIWN-R; Wechsler, 1983), Escala de Inteligencia Revisada para el Nivel Escolar (WISC-RM; Wechsler, 1984, adapted and normed in Mexico by Margarita Gómez-Palacio, Eligio Padilla, and Samuel Roll), and la Escala de Inteligencia Wechsler para Niños-Revisada en Puerto Rico (EIWN-RPR; Wechsler, 1992, adapted and normed by Laura Herrans and
Juana Rodriguez). The Psychological Corporation published both the EIWN-R and EIWN-RPR. They are both still listed and available for purchase on the Psychological Corporation's product catalog as of July 8, 2003. The WISC-RM is published and distributed by Editorial el Manual Moderno, a publishing company in Mexico dedicated, in part, to psychological tests and literature.

According to Herrans and Rodriguez (in Wechsler, 1992), the EIWN-R was a literal translation of the WISC-R carried out in Dade County, Florida, comprised largely of Cuban Americans. Except for the English-to-Spanish translation, there was no adaptation of the instrument for Spanish-speaking children and there were no norms developed for its use.² Herrans and Rodriguez reported that the only changes made to the instrument were changes in the scoring criteria used to identify acceptable responses, for example, in the vocabulary and comprehension subtests. Prewitt-Diaz and Rodriguez (1986), however, noted that two of the original WISC-R vocabulary words were deleted, thus, the content of the Spanish and English versions are nearly identical. Given that the norms were the same for the EIWN-R and the WISC-R, which limits their applicability to the assessment of Spanish-speaking children, we have chosen to focus on the two Spanish-language versions of the WISC-R that have Spanish-language norms—the WISC-RM and the EIWN-RPR.

Like the EIWN-R, the WISC-RM is also a Spanish-language translation in which few changes were made. Only the information, vocabulary, and comprehension subtests differ from the WISC-R and these changes included the deletion of nine, seven, and five items from each respective subtest and the inclusion of seven, six, and five new or significantly modified items. For other subtests, the content is the same, but slight alterations were made to the order of a few items. In addition to these changes, the WISC-RM was standardized on a sample of students from public schools in Mexico City during 1980 to 1981 (Padilla, Roll, & Gómez-Palacio, 1982). The sample was comprised of 1,100 students ages 6 through 16. Thus, the WISC-RM is a close literal translation of the WISC-R and it has its own norms.

With regard to the psychometric properties of the WISC-RM, some reliability estimates and only one limited test of validity were reported in the test manual.³ The reliability of the subtests was assessed using the split-half method with Spearman-Brown corrections. The average correlation across the 11 age ranges for each subtest ranged from .65 to .82. For the three IQ scales (verbal, performance, and full scale), their average reliability across the age ranges were .90, .89, and .94, respectively.⁴ The standard error of measurement (SEM) was also reported. Averaging across the age ranges, the SEM for the subtests ranged from 1.29 (block design) to 1.77 (comprehension). For the scales the standard errors of measurement were 4.64 (verbal), 5.06 (performance), and 3.79 (full scale).⁵ In terms of the test’s validity, the intercorrelations of the subtests and the three IQ scales were reported for each age group. In addition, the

²Limited assessments of the EIWN-R’s reliability and validity have been conducted (Prewitt-Diaz & Rodriguez, 1986; Rodriguez & Prewitt-Diaz, 1990).
³A factor analysis has since been conducted on a sample of 300 Mexican children from Baja California (Rousey, 1990).
⁴For the three IQ scales (verbal, performance, and full scale), reliability coefficients were computed according to a formula based on a composite of several tests (Gullford, 1954).
⁵Note that the standard error of measurement is inversely related to the reliability coefficient; that is, the greater the reliability, the lower the standard error of measurement. Furthermore, the differences between the subtests and the scales are a function of different standard deviations (3 for the subscales and 15 for the scales). Thus, the lower scores of the subtests do not mean that they are necessarily more reliable.
average intercorrelations of the subtests and the three IQ scales were reported across all age groups. A general pattern of the average intercorrelations indicated that the verbal subtests correlated more highly with the overall verbal subtest score (.45 to .71) than with the overall performance subtest score (.31 to .52). A more attenuated pattern was observed in the relations between the overall performance subtest score with each of the performance subtests (.34 and .58) and each of the verbal subtests (.38 to .49).\(^6\)

The third Spanish-language version, the EIWN-RPR, is like the WISC-RM in that it was translated and normed. In this case, the standardization sample was based on 2,200 Puerto Rican island children whose ages ranged from 6 through 16 and who were enrolled in both public and private schools from 1986 to 1987 and 1989 to 1990. In addition to the norming, changes were made to the test content in all subtests except digit span, object assembly, coding, and mazes, which were identical to those of the WISC-R. According to Herrans and Rodriguez (1992), changes were made to ensure that the test content was congruent with the life experiences of Puerto Rican children. Of the subtests that were modified, the percentage of changes ranged from 15% of the picture completion items to 47% of the similarities items (Wechsler, 1992). More than one third of the subtest items were changed for only three subtests. Thus, the EIWN-RPR is the version of the WISC-R that was most adapted for use with a Spanish-language group. Moreover, it has specific norms for Puerto Rican children.

The psychometric properties of the EIWN-RPR were assessed using the same tests as those for its Mexican counterpart, plus tests of temporal stability and concurrent validity. With regard to internal consistency, the average correlation across the 11 age ranges for each subtest ranged from .62 to .84. For the three IQ scales (verbal, performance, and full scale), the average reliability across the age ranges was .92, .88, and .94, respectively. Averaging across the age ranges, the standard error of measurement for the subtests ranged from 1.20 (block design) to 1.94 (object assembly). For the scales, the standard errors of measurement were 4.26 (verbal), 5.23 (performance), and 3.77 (full scale). To assess the test's temporal stability, the test was administered a second time within a 2-week period to a subsample (n = 150) of five age groups.\(^7\) The corrected stability correlation coefficients ranged from .53 (mazes) to .91 (vocabulary) for the subtests and were .91, .77, and .90 for the three main scales of verbal, performance, and full scale.

To assess the EIWN-RPR's validity, the intercorrelations of the subtests and the three IQ scales were first reported for each age group and then for the average intercorrelations across all age groups. The verbal subtests correlated more highly with the overall verbal subtest score (.51 to .73) than with the overall performance subtest score (.39 to .46). A less differentiated pattern was observed in the relations between the overall performance subtest score with each of the performance subtests (.23 and .57) and each of the verbal subtests (.25 to .43). In addition to examining the inter-

---

\(^6\) When correlations were computed for subtests with scales that included that particular subtest (e.g., vocabulary with verbal IQ), then the correlation was corrected following procedures outlined in McNemar (1955). In addition, for the correlational analyses, the verbal and the performance scale scores did not include the digit span or mazes, respectively.

\(^7\) The test-retest stability coefficients were corrected for the variability of the standardization sample, which provides an estimate of the score stability within the population.
correlations, small groups ranging from 31 to 48 subjects were administered the EIWN-RPR and other intelligence tests, including the Raven Progressive Matrices, the Test Rápido de Barranquilla (BARSIT), and the Escala de Inteligencia Wechsler para Adultos (EIWA). The participants' total scores on each of these tests were significantly associated with their total score on the EIWN-RPR: BARSIT, $r = .56$; Raven's $r = .58$; EIWA, $r = .84$. In addition, a significant correlation ($r = .50$) was observed between the Full Scale IQ of the EIWN-RPR and the academic record of nearly all the examinees from the standardization sample (1,956 of 2,200).

With regard to assessing Spanish-speaking adults, we know of two test versions, the Escala de Inteligencia para Adultos, or WAIS-III Español (Wechsler, 2001), and the Escala de Inteligencia Wechsler para Adultos (EIWA; Wechsler, 1967). Like the EIWN-R, the WAIS-III Español is a literal translation of the WAIS-III with no new norms or reported psychometric properties. It is published and distributed by Editorial el Manual Moderno. The EIWA, on the other hand, is similar to the EIWN-RPR, as it has its own norms based on an adaptation and standardization carried out on the island of Puerto Rico. The Psychological Corporation publishes and continues to distribute this test as it is the only Spanish-language intelligence test that was normed on a specific Latina/o community (country) of origin outside the mainland, with potential relevance to Latina/os residing in the United States.

The reliability of the EIWA was estimated primarily by calculating split-half correlations⁸ and the standard errors of measurement for three age groups 16 to 19, 25 to 34, and 45 to 54 taken from the standardization sample. For the 25 to 34 age group ($N = 224$), for example, the subtests' reliabilities ranged from .66 (digit span) to .95 (vocabulary) and the three scales were .97 (verbal), .96 (performance), and .98 (full scale). Considering the same age group, the standard errors of measurement for the subtests ranged from .66 (vocabulary) to 1.71 (digit span) and for the scales they were 2.64 (verbal), 2.85 (performance), and 2.01 (full scale).

Intercorrelations of the subtests and the scales comprised the only assessment of the EIWA's validity. Again, data were reported for the three age groups used in the reliability assessment. Based on the 25 to 34 age group, for example, the correlations of the six verbal subtests with the verbal scale ranged from .72 to .85 and from .68 to .78 with the performance scale, which was made up of the five subtests. In contrast, the associations of the performance subtests with the performance scale ranged from .75 to .82 and from .64 to .77 for the verbal scale. No tests of concurrent or predictive validity were reported.

In summary, there are two types of Spanish-language versions of the Wechsler tests. One type includes a Spanish translation with minimal adaptation and the use of the U.S. English language norms, whereas the second type includes a Spanish translation with more significant adaptation and new Spanish-language norms based on examinees from Puerto Rico or Mexico. The Spanish-language normed tests report the most comprehensive assessment of the test's psychometric properties for Spanish-speaking individuals. However, these data are based on children and adults living outside the United States mainland. No data were reported regarding the tests'⁸

---

⁸The exception to calculating the subtests' reliability via split-half correlation pertained to the digit span, which was assessed by correlating the scores from the forward and reverse recall procedures, and to the digit symbol, which was assessed by correlating the scores of the first and second administration (test-retest).
reliability and validity in the assessment of Spanish-speaking children and adults residing on the mainland.

**Woodcock Johnson Psycho-Educational Battery.** The Woodcock Johnson Psycho-Educational Battery—Revised (WJ-R) is described as a "wide-range, comprehensive set of individually administered tests for measuring cognitive abilities, scholastic aptitudes, and achievement" (Woodcock & Mather, 1989/1990, p. 1).\(^9\) The WJ-R is comprised of 21 subtests to measure cognitive ability (7 subtests in the standard battery and 14 supplemental subtests) and 18 subtests to measure achievement (9 subtests in the standard battery and 5 subtests and 4 derivative measures for the supplementary battery). In addition, two subtests from the cognitive battery and two subtests from the achievement battery are used to comprise a separate language survey (Woodcock & Muñoz-Sandoval, 1993). Together, the WJ-R battery provides a comprehensive assessment of cognitive ability and achievement skills and serves as the foundation on which its Spanish-language version, the Batería-R, is based.

The cognitive abilities tests of the Batería-R are referred to as the "Prueba de habilidad cognitiva—Revisada" (Batería-R COG, Woodcock & Muñoz-Sandoval, 1996b) and the achievement tests are referred to as the "Pruebas de aprovechamiento—Revisada" (Batería-R ACH, Woodcock & Muñoz, 1996a). All subtests in the Batería-R were adapted from the parallel English-language subtests in the WJ-R. Some subtests have identical item content and task requirements for the two language versions (e.g., spatial relations). Other tests with significant verbal content have different item content in the two language forms (e.g., oral vocabulary). The authors of the Batería-R based the Spanish-language norms on the U.S. English-speaking normative sample (Woodcock & Muñoz-Sandoval, 1996b). They argued that drawing a set of norms from Spanish-speaking individuals residing in the United States would be difficult because this community is in flux and because professionals would not likely accept such a group as a representative norming sample for assessment purposes (Woodcock & Muñoz-Sandoval, 1996b, p. 23). Moreover, the authors noted that U.S. English-speaking individuals are "identifiable," "relatively stable over time," and represent the "standard reference for norms on tests" in the United States.

Because the norms of the Batería-R were based on the U.S. English-speaking norms, it is important to consider the English-language standardization sample, a U.S. national sample of 6,359 persons, ages 24 months to 95 years. It used a stratified sampling design that controlled for four community variables (e.g., census region and community size) and six examinee characteristics (e.g., race, Hispanic origin, and educational level). The examiner's manual indicates that subjects were randomly selected within this stratified sampling design, although the manner in which this was carried out was not delineated. A total of 9.3% of the standardization sample was Hispanic, which compares favorably to the group's representation across all ages in the 1980 U.S. census (6.4%). Individual subject weighting was used in the data analysis so that the distribution of the subsamples as they relate to the 10 sampling variables approximated the 1980 U.S. census data.

---

\(^9\)The Woodcock Johnson-III (WJ-III) is now available, but because its Spanish-language companion has yet to be published, we do not consider the WJ-III in this chapter.
For those Spanish-language subtests with content identical to their English-language counterparts, no new norms were derived. The U.S. English-language norms were applied without modification. For those tests in which the subtest content differed, equated U.S. norms were developed that are very similar, though not identical, to the U.S. English-language norms. An important step in developing the equated norms was to calibrate the Spanish-language items with largely monolingual Spanish-speaking individuals. The calibration analyses addressed the assessment of item difficulty rather than the development of specific Spanish-language norms. This sample was comprised of approximately 2,000 native Spanish-speaking persons who primarily resided in Mexico and the United States; others lived in Costa Rica, Peru, Puerto Rico, and Spain. Those from the United States reported having been born in 21 different Latin American countries, including the United States and Spain. Thus, for the calibration analyses, test developers sampled monolingual Spanish-speaking individuals from a large number of Spanish-speaking countries and the United States.

The reliability assessment of the Batería-R was based on the calibration sample data. The test’s internal consistency reliability coefficients (split-half correlations corrected for length by the Spearman-Brown formula) and standard error of measurement were reported. The median internal consistency coefficients across a number of ages and for selected subtests ranged from .83 (incomplete words) to .92 (verbal analogies). The median standard error of measurement for the same selected ages and subtests ranged from 4.9 (verbal analogies) to 8.1 (memory for words).

Two limited concurrent validity studies were reported in the supplemental manual examining the relationship between the oral language cluster measures from the Batería-R to other language measures. One study of about 70 kindergarten students designated as limited English proficient (LEP) found correlations ranging from .44 to .85. The second study with about 120 LEP second graders found a wider range of associations, from .15 to .88. The authors did not indicate why they chose to focus specifically on language measures rather than other measures of cognitive-intelligence abilities. Based on this limited assessment of reliability and validity, the authors reported that “the reliability and validity characteristics of the Batería-R COG meet basic technical requirements for both individual placement and programming decisions” (Woodcock & Muñoz-Sandoval, 1996b, p. 29).

In summary, the Batería-R COG distinguishes itself by being a Spanish-language version of a cognitive-intelligence performance test using equated norms based on the U.S. English-language norms. There were no reports of the test’s validity in the assessment of cognitive-intellectual functioning.

**Nonverbal Tests of Cognitive Ability.** The General Ability Measure for Adults (GAMA; Naglieri & Bards, 1997) is a standardized, norm-referenced test developed to measure intellectual abilities of adults using nonverbal stimuli. Some researchers have argued that tests requiring English-language skills and knowledge learned in school are limited in assessment ability (e.g., Prewett, 1995), particularly among

---

10The supplemental manual reported that the developers tested the associations between the Batería-R oral-language test and the Woodcock-Munoz Language Survey–Spanish oral-language test in these two studies. However, given that the two oral-language tests are noted as having common content, we did not include them in the range of correlations.
persons with diverse educational, cultural, and linguistic backgrounds. One of the objectives of this test then is to address that concern by using abstract designs as the test's content. The test is brief (66 total items), can be administered individually or in groups, and is comprised of four subtests: matching, analogies, sequences, and construction. The test contains a general set of written instructions that are to be read aloud for each subtest as well as written instructions for each of the 66 items. The reading level is noted to be at the third-grade level. A Spanish-language version of the test is available.

The test was standardized on a sample of 2,360 people from 18 to 96 years of age. The sample included nine age groups stratified by race or ethnicity, gender, five educational levels, and four geographical regions. The racial and ethnic groups were African American, American Indian, Asian or Pacific Islander, Hispanic, and White. The stratification by the main demographic variables approximated the 1990 U.S. census. This generally appeared to be the case with regard to the age and educational level of Hispanics who comprised 7.8% of the sample and who in 1990 comprised 7.9% of residents 18 years and older in the United States (U.S. Census, 2001a). In addition to being represented in the standardization sample, race and ethnicity were also considered in the composition of the tryout sample of 604 persons. The manual reports that differential item functioning (DIF) for Whites and non-Whites was assessed based on the tryout data and, where possible, the items with a significant DIF were eliminated. Three such items were identified as having a significant differential item function for Whites and non-Whites. With regard to language background, nearly 7% reported that they spoke more than one language. Thus, it appears that the sample was over 93% English speaking only, with the remaining sample having some unknown level of proficiency in another language.

Overall, the test authors report a range of reliability and validity assessments, although none were conducted specifically with Hispanics. Given that this is a nonverbal test of intelligence it is most important to consider how it relates to other tests of intellectual ability that include verbal content. In one study, the GAMA was administered concurrently with the WAIS-R and the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990) and found to have a similar pattern of mean scores and significant correlations with these other tests' full IQ scores (.75, .70), verbal scores (.65, .54), and performance scores (.74, .72). These data support the validity of the GAMA as a general measure of intellectual ability. The sample of this study was comprised of 194 individuals of whom 88% were identified as White; no Hispanics were noted as being included in this sample.

The Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) represents a recently developed, individually administered nonverbal measure for children and adolescents from ages 5 through 17. Like the GAMA, the UNIT is principally designed to address issues of fairness in the cognitive-intellectual abilities of youth, particularly those "who may be disadvantaged by traditional verbal and language-loaded measures" (Bracken & McCallum, p. 1). Among these groups are children and adolescents from diverse cultural, linguistic, educational, and hearing-impairment backgrounds. In addition, the UNIT is intended to measure a broad range of abilities, not relying on a single ability dimension (e.g., problem solving). Another key characteristic of the UNIT is that it is completely nonverbal, from test content to test administration.
The test is comprised of six subtests, three measures of memory and three measures of reasoning. In addition, the instrument is designed to tap symbolic processes, those thought to be associated with language, and nonsymbolic processes, those thought to be less associated with language. The test can be administered as an abbreviated or screening battery (two subtests), a standard battery for placement decision making (four subtests), or an extended battery for additional diagnostic information (six subtests).

The UNIT is norm referenced and has been standardized with a sample of 2,100 children and adolescents. An additional 1,765 youth participated in studies that assessed reliability, validity, and fairness. The sample was stratified according to the usual demographics including race and ethnicity, sex, geographical region, community setting (rural, urban, or suburban), and parental educational attainment (four levels). With regard to Hispanics, 13% comprised the normative sample, which is consistent with the census reports for 1995 for children in the 5 to 17 age range (13.1%, U.S. Census Bureau, 1996). In addition to these usual demographic variables, the test designers included in the standardization sample students receiving special education services, such as services for learning disabilities (5.6%), bilingual education (1.8%), and English as a Second Language (ESL; 2.0%).

Given the main objective of the UNIT, to provide a fair assessment of persons from diverse backgrounds, considerable attention was given to fairness. The authors first presented the steps they took to reduce any potential cultural bias. Included among the steps was the subjective review by ethnic minority psychologists and hearing-impaired individuals for items that were potentially biased or for any materials that might be perceived as offensive. Another step was the assessment of differential item functioning of all items within and between ethnic groups, level of English-language proficiency, and other key sample characteristics. In analyses with Hispanics, only one item was identified as possibly having a different meaning for Hispanics as opposed to the overall normative sample. However, after careful review of the data and the content, bias experts and the authors concluded that the evidence did not support the hypothesis that the item was biased and it was therefore retained.

To further assess the test’s fairness, the reliability and validity of these tests were examined separately for specific demographic groups, including Hispanics (Ns = 120 to 156). In terms of the subtests’ reliability for Hispanics, correlations using the split-half method with Spearman Brown corrections ranged from .75 (mazes) to .94 (cube design). The internal consistency of the scales for the standard battery, for example, ranged from .91 (symbolic) to .95 (full scale). The standard error of measurement for the subtests ranged from .57 (cube design) to 1.12 (object memory) and for the scales of the standard battery from 2.46 (nonsymbolic scale) to 4.06 (symbolic scale). The age range and source of the sample for the reliability assessment were not reported.

In terms of validity, the authors reported the correlation of the UNIT with the Bateria-R (Broad Cognitive Ability Early Development) for two native Spanish-speaking samples, from bilingual education classrooms (n = 27) and ESL classrooms (n = 26). For the bilingual classroom students, the correlations of the scale scores for the standard battery of the UNIT and the cognitive ability scores of the Bateria-R ranged from .10 (nonsymbolic scale) to .38 (symbolic scale). The same correlations for the ESL sample were noticeably lower, from .04 (reasoning scale) to .23 (memory scale). Moreover, confirmatory factor analyses for Hispanics drawn from the
standardization sample \( (n = 215) \) showed that the models tested had an adequate fit to the data.

In addition to carrying out tests of reliability and validity, the authors of the UNIT tested whether the scale means differed for three pairs of groups, all relevant in the assessment of Hispanics. These comparisons included matched samples of Hispanics and non-Hispanics, Bilingual or ESL students and English-speaking examinees, and Ecuadorians and a subsample from the standardization sample. For the standard battery, effect sizes of the scales' mean differences for these comparisons ranged as follows: Hispanics and non-Hispanics (.04 to .20), bilingual/ESL and English speaking (.06 to .36), and the Ecuadorian and standardization subsample (.04 to .73).

In summary, the selected nonverbal tests of cognitive-intellectual functioning range from considering ethnicity much like the English-language Wechsler tests, with attention to sampling and tests of bias, as was the case with the GAMA, to examining the test's technical properties specifically for U.S. Latina/os, as reported in the UNIT's test manual. In addition, for the UNIT, studies of the test's reliability and validity were reported for limited-English-speaking children in the United States and for a non-U.S. sample of Latina/o children from Ecuador.

Psychodiagnostic Measures

**MMPI and MMPI-II.** The second edition of the Minnesota Multiphasic Personality Inventory (MMPI-II; Hathaway & McKinley, 1991) is the most widely used objective psychodiagnostic measure in the world (Greene, 1991) and among U.S. Latina/os (Velasquez, Ayala, & Mendoza, 1998). This self-report instrument contains 567 items that comprise 10 clinical scales and 3 basic validity scales. Many additional scales are also routinely scored to supplement the clinical scales, including scales to address areas such as substance abuse (MacAndrew Alcoholism Scale) and response consistency (e.g., Variable Response Inconsistency). The normative sample of this instrument is large \( (n = 1,138 \text{ men}; n = 1,462 \text{ women}) \) and was selected from seven geographic regions within the United States. Overall demographic characteristics of the sample parallel the 1980 census data in areas such as age, gender, and some ethnic group representation (Whites, Blacks, and American Indians). However, the normative sample deviates from the 1980 census (U.S. Census Bureau, 2001b) in the underrepresentation of Hispanics (2.8% in sample, 6.4% in U.S.) and persons with a high school education or less. In addition, there was an overrepresentation of persons with at least some college education.

Other than reporting the ethnic breakdown of the sample, the only other analyses of Hispanics reported in the test manual were included in a table of the mean raw scores from the normative sample for Hispanics and Caucasians by gender (see Appendix H, tables H-1 and H-2). The data indicated that Hispanic men's \( (n = 35) \) average raw score was slightly higher than that of Caucasian men \( (n = 933) \) on two of the three traditional validity scales and 7 of 10 clinical scales (Graham, 2000). Hispanic women \( (n = 38) \) in the normative sample were found to score higher than their Caucasian counterparts \( (n = 1,184) \) on all but one validity scale and two clinical scales. The meaningfulness of these group differences is limited by many factors, including the small sample size of Hispanics, the lack of statistical analyses, and in general small mean differences. In fact for men, none of the scale mean differences
were greater than the standard deviation for any of the scales for either Hispanics or Whites.\footnote{See Velásquez et al. (1996) for a review of the available studies of the MMPI-2 with Latina/os.}

To our knowledge, the MMPI was first translated into Spanish in 1949 for use in Cuba by Bernal and Fernandez (cited in Butcher, Cabiya, Lucio Maquero, & Velasquez, in press). Since then, several Spanish-language translations have been developed, including the Núñez translation for Mexico and the García and Azán translation in 1984. García and Azán (1993) also translated the MMPI-2, which has been widely used in the U.S. and in Puerto Rico. Until recently, there were no norms developed for the Spanish-language MMPI-2. Butcher, Azán-Chaviano, Cabiya, and Scott (in press) normed the Spanish-language version of the MMPI-2 with a sample from three U.S. cities, Los Angeles (N = 205 with 165 valid-usable profiles), Miami (N = 200 with 165 valid-usable profiles), and San Juan (N = 200 with 151 valid-usable profiles). Study participants were primarily of Mexican, Cuban, and Puerto Rican descent but persons from other Latin American countries were included in the sample if they met criteria. Participants were selected based on the following guidelines: (a) 18 years or older, (b) able to read Spanish at a 6th-grade level or higher, and (c) of "normal" mental health (i.e., not mental health clients). An additional criterion was that no site was to include more than 10% college or graduate students in the sample. The average level of education for the Hispanic sample was 13.56 years for males and 13.41 years for females. With respect to gender, 52.9% of the sample was female. Most of the sample (52%) was married or never married (35.1%) and few participants were divorced (7.8%), widowed (3%), or separated (2.1%). The authors reported adequate test-retest reliability for all of the clinical scales, ranging from .61 to .91. A manual supplement will soon be available from the University of Minnesota Press with the U.S. Hispanic set of norms.

In addition to the U.S. Hispanic set of norms, Lucio and colleagues developed a Spanish-language translation (Lucio, Reyes-Lagunes, & Scott, 1994) and a set of norms for Mexico (Lucio, Ampudia, Durán, León, & Butcher, 2001). In their initial line of investigation they compared the responses of Mexico City college students from la Universidad Nacional Autonoma de Mexico (UNAM; N = 2,174) to those of U.S. college students (N = 1,312). The authors found few significant differences (SD > .5) between the two samples on the validity and clinical scales (Lucio et al., 1994). In a second study examining the validity, clinical, and content scales, the investigators found that the responses of UNAM college students (N = 2,246) differed greatly from a diagnostically heterogeneous sample of patients (N = 233) from three psychiatric hospitals in Mexico City (Lucio, Palacios, Durán, & Butcher, 1999). In addition, they carried out a factor analysis of the patients' responses to the validity and clinical scales separately for men and women.\footnote{In two other studies, the authors report on the internal consistency and temporal stability of the Spanish-language MMPI-2 with Mexican college student samples (Ampudia, Durán, & Lucio, 1995; Lucio, Pérez-Parías, & Ampudia, 1997). We could not evaluate these findings as we were unable to obtain copies of these two Spanish-language research reports.}

In an effort to extend the norms beyond a college student sample, Lucio et al. (2001) developed adult norms based on a sample of 1,744 (860 men and 884 women) from different parts of Mexico that represent the "economically active urban population" (p. 1461). The standardization sample included persons between the ages of 19 and 80
years (M = 31). In addition, over one third of the sample had at least a college-level education (38% for men and 44% for women). Lucio et al. reported that the internal consistency of the validity and clinical scales ranged from .62 to .87 for males and .59 to .86 for females. In addition, they examined the differences between the Mexican and U.S. norms. Significant nationality differences (effect sizes greater than .40) were observed for three scales for men (L, Hs, Mf) and seven scales for women (L, F, Hs, D, Mf, Sc, Ma). The authors concluded that, although there is considerable similarity between the norms of the two countries, it is “preferable to use the Mexican norms” in the assessment of Mexicans because such norms take into account cultural factors.

Millon Clinical Multiaxial Inventory—III. The MCMI-III is another widely used self-report inventory that focuses on the assessment of personality disorders associated with the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1994) as well as other clinical syndromes related to Millon’s theory of personality. This instrument contains 175 true-false items that provide scores on a variety of scales. Fourteen scales measure clinical personality patterns (Axis II), three scales measure severe personality pathology (Axis I), seven scales measure clinical syndrome (Axis I), and three scales measure severe clinical syndrome scales (Axis I). There are also three modifying indices and a validity scale. The three modifying indices—disclosure, desirability, and debasement—assess response tendencies that are associated with particular personality patterns or Axis I conditions. This instrument is intended for use with adults who can read at the 8th-grade level or higher and for those who are seeking mental health services. A strength of the MCMI-III is its strong theoretical framework, but few validation studies have been conducted to verify the accuracy of the theoretical deductions.

Like the standardization sample of the MMPI-2, Hispanics were not well represented in the standardization sample of the MCMI-III: they made up only 2.8% of the development sample and 2% of the cross-validation sample of the MCMI-III (Millon, 1997), compared to 9.2% of the 18 and older U.S. 1995 population (U.S. Census, 1998). No breakdown was given in terms of subgroups (e.g., Cuban Hispanics vs. Puerto Rican Hispanics), due to the fact that the overall number was modest (T. Millon, personal communication, February 6, 2002). Millon (personal communication, February 6, 2002) also conveyed that, although the test developers had intended to collect separate minority group norms, this idea was dropped in the MCMI-III because several scholars reacted negatively, suggesting that “group separations” such as White and Black with such a small sample could produce pernicious and false comparisons. The MCMI-III is available commercially in both English and Spanish. However, no published data are yet available on differences between Hispanics and Whites on the MCMI-III (Strack, 1999).

EVALUATION

Perhaps the single critical question with regard to norm-referenced tests is: What is the appropriate norm for Mexican Americans? Given our definition of culture, which

---

13 The full names of the noted scales are: L = Lie; F = Frequency; Hs = Hypochondriasis; D = Depression; Mf = Masculinity-Femininity; Sc = Schizophrenia; Ma = Hypomania.
acknowledges its shifting and changing nature and which emphasizes the importance of specific local contexts, multiple norms are essential for the psychological assessment of persons of Mexican origin living in the United States. Among the most important contributions by test developers in the assessment of Mexican Americans is the development of Spanish-language tests based on standardization samples from Mexico. Spanish language tests developed in other locales (e.g., Puerto Rico) have also been helpful. Those persons of Mexican origin who are likely to be at the greatest risk for error in psychological assessment are those whose predominant language is Spanish or who are recent immigrants to the United States (Figueroa, 1989). Having Spanish-language tests normed within a Spanish-speaking locale adds greatly to the available English-language tests and nonverbal tests that can be used in assessing Mexican Americans and other Latina/os.

The EIWN-RPR is, in our view, one of the best Spanish-language normed tests developed thus far. Thoughtful adjustments were made to the test content, scoring, and administration to take into account the specific background of Puerto Rican island children. In addition, the adjustments do not result in significant deviations from the original WISC-R. It is not clear, however, if the test can be effectively used with mainland Puerto Ricans and other Latina/o children in the United States. The careful translation and adaptation suggests that it is a promising measure for other Latina/o children, however.

The Mexican adaptation of the WISC-R is limited by having only sampled children from public school settings. As a result, it may tend to overestimate children’s level of functioning (Fletcher, 1989). The very recent developments of U.S. Hispanic norms and Mexican norms for the Spanish-language MMPI-II are also encouraging. However, the investigation of the MMPI-II scale’s relations with other clinical measures is needed to examine further its construct validity.

A second development in testing that has important implications for the assessment of Mexican Americans is the further development of nonverbal tests of cognitive-intellectual functioning with contemporary norms. Like the Spanish-language versions of tests, these tests can contribute to assessing those Latina/o adults and children most at risk for error in their assessment. The UNIT is exemplary in the multiple ways the authors considered ethnicity and related examinee variables in the test’s development. For example, of the tests reviewed in this chapter for which U.S. national norms were obtained, the UNIT was the only test that reported reliability and validity data for Latina/o and limited English-proficient persons. Nonverbal tests can be most beneficial in the assessment of limited English-speaking Latina/os, but also in other circumstances (e.g., with deaf individuals).

In addition to the strides made in developing Spanish-language normed tests and nonverbal tests, there have been key developments in how tests are standardized in the United States. For example, most test developers now include in their standardization sample a representative proportion of the major racial and ethnic groups. The selected cognitive-intellectual tests for the most part met this standard, whereas the psychodiagnostic tests fell short. A second, relatively new development is the assessment of potentially biased test items. Consultants from racial and ethnic minority groups were asked to carry out subjective reviews to identify potentially biased or insensitive items. In addition, objective statistical analyses, specifically, analyses of differential item functioning, are now part of test development protocols. In the best
of circumstances, persons from the major racial and ethnic groups are oversampled and then analyses based on item response theory are carried out for the specific groups. With regard to Latina/os, this type of analysis was only carried out in the development of the UNIT. The other cognitive-intellectual tests grouped the ethnic minority groups together to assess DIF. The test developers of the psychodiagnostic tests did not report such analysis, although studies have since been published to begin examining the DIF of the MMPI (Waller, Thompson, & Wenk, 2000).

Other developments include the analyses and reporting of test reliability and validity not only by the entire sample or selected subsamples (e.g., age cohorts), but also by racial and ethnic groups. Reliability and validity analyses by race and ethnicity, including separate analyses of Latina/os, were reported only in the UNIT. The reliability analysis for the Bateria-R, however, was limited to the Spanish-language calibration sample and the validity assessment only concerned language skills.

Tests that are developed in Spanish but then equated with U.S. English-language norms, as is the case for the Bateria-R, have some strengths and limitations. The strength of this approach is that the examiner can compare individuals' level of functioning in Spanish to that of their English-language peers in the United States. In other words, given equated norms, a certain level of functioning in Spanish corresponds to a certain level of functioning in English. Another advantage is that the examiner can administer both language versions of the test to assess how a given individual performs in both languages using a nearly identical metric. It may be that the individual has a higher or lower level of functioning in one language than in another. Using the same metric, one can discern the relative strengths or weaknesses in one language compared to the other. When using distinct norms for each language version of the same or similar test, it is difficult to know the relative level of functioning given that the metric may change with the two distinct sets of norms (one from Puerto Rico and the other from the United States, for example). Thus, having one metric apply to two language versions of a given test has clear advantages.

The limitation, however, is significant; it is not clear how a given Spanish-speaking person is functioning relative to others from that country or place of origin. For example, if a Spanish-language-dominant person performs in English at a level consistent with mild mental retardation, then the examiner may want to assess the person’s level of functioning in Spanish as one way to ascertain the role of language and possible sociocultural factors in the evaluation. It may be that the low level of functioning is the result of limited English-language skills. If the person performs within the mild mental retardation range in Spanish compared to others having been raised in Mexico, for example, then this would suggest that linguistic and sociocultural factors do not likely contribute to the identified level of functioning. The examinee is functioning within the mild mental retardation range in both English and Spanish, compared to a U.S. sample and a Mexican sample as well. However, if the Spanish-language assessment reveals that the person is performing within the average range in her native language compared to others raised in a similar environment, then it is more likely that language and sociocultural factors contributed to the English language-based assessment that suggested mild mental retardation. Assessing the person’s level of functioning in English and Spanish and comparing that functioning to U.S. and Mexican norms provides a unique vantage point to examine whether linguistic and sociocultural factors play a role. If the Spanish-language test is equated to approximate English-
language norms, as is the case with the Bateria-R, then the Spanish-language test will not be able to address the question of how language and other possible sociocultural factors may play a role in the examinee’s observed functioning.

The assumption underlying the equated norms is that the U.S. English-language norm is the most relevant norm in carrying out cognitive-intellectual assessment. An anthropologically informed perspective would argue for the availability of multiple norms to advance an understanding of a given individual relative to specific local contexts. This is not unlike the decision to apply age-appropriate norms for the WAIS-III rather than one set of norms across the adult ages, as was done in prior versions of the WAIS.

The most striking limitation of tests for Mexican Americans and other Latinos residing in the United States concerns the development of norm-referenced Spanish-language tests (U.S. Department of Health and Human Services, 2001, p. 145). The available Spanish-language tests of cognitive-intellectual ability of adults, for example, are limited by either significantly outdated norms or limited reports of reliability and validity. Among the most frequently used options for psychologists who administer Spanish-language verbal tests of cognitive intellectual functioning to adults residing in the United States include the EJWA with norms that are over 35 years old, the WAIS-III Espanol with no reported reliability or validity for Spanish-speaking individuals, and the Bateria-R COG with very limited assessments of its reliability and validity. Up-to-date, psychometrically valid instruments in Spanish are clearly needed.

FUTURE DIRECTIONS

To build on the reported developments and to address the noted limitations, we recommend that test developers consider a number of issues (see also Suzuki & Valencia, 1997). We believe that the general idea of the Bateria-R as a family of tests built around a basic test provides a fruitful design for the development of tests for Mexican Americans and other Latina/os residing in the United States. Such a test would have the English-language version with U.S. norms as well as the Spanish-language version with Mexico, Puerto Rico, or both sets of norms. With both English- and Spanish-language sets of norms available from specific countries or locales of origin, an examiner could then administer a cognitive ability test and examine how that child or adult functions relative to U.S., Mexican, or other Spanish-speaking children or adults. Many alternatives would be available to the examiner, providing a rich assessment protocol.

Let us consider such a test in the assessment of a 13-year-old Mexican immigrant youth who has lived in the United States for 5 years and is having academic difficulties. Let us assume that she is Spanish dominant but also speaks English. The examiner could evaluate her in Spanish first. Her performance could then be compared to that of Mexican children her age. This comparison is useful because it provides a metric that is close to her sociocultural background.

In addition to comparing her functioning with Spanish-speaking Mexicans, this hypothetical test could also be used to assess her functioning with English-speaking youth from the United States. The English-language version could be administered
as well. The relative functioning of this adolescent girl in the two languages compared to the two normative samples would provide a wealth of information. Questions of whether her academic difficulties are a function of language, culture, or low cognitive ability can be better addressed by having data from both linguistic and cultural perspectives. For example, if she was found to have some deficits in English and not in Spanish then this would suggest that her academic difficulties may be related to her transitioning to English. On the other hand, if she was found to be low functioning in a given area in both English and Spanish, then this would suggest a deficit in that specific area.

The potential contribution of a multiple-normed test is great, but there are two interrelated challenges in developing such an instrument. One concerns the degree of overlap in test content. Should the two language versions of the same subtests have identical content whenever possible? For example, should the content be the same or different for short-term recall of digits? The advantage of similar or identical content is that a similar metric can be developed. The disadvantage is related to the second challenge, which concerns order effects. If the content is similar or identical in the two language versions, then the level of functioning as assessed with whichever test is administered second is likely to be elevated due to order effects. On the other hand, if the content is different, then it is less clear that the two language versions reflect the same metric. These challenges are not insurmountable. For example, a large item pool could be developed, equated, and drawn from if more than one test version is administered.

In addition to developing Spanish- and English-language test versions, multiple Spanish-language versions could be adapted and standardized with other Spanish-language populations (e.g., Mexicans and Puerto Ricans), if so desired. This could be applied to tests of psychodiagnostics as well as of cognitive-intellectual functioning. Multiple norms could be developed for nonverbal tests as well. The main idea is to have more than one norm for a given test. From an anthropological point of view, this test prototype attempts to understand a given person’s functioning within his or her given context. For some individuals, this may mean the U.S. English-language norm, for others the Mexican Spanish-language norm, and still for others both sets of norms will be useful in trying to understand their behavior within their specific setting or settings.

In addition to developing a multiple-language and multiple-normed test, test developers can take steps to improve the way they consider ethnicity and culture in the development of single-norm tests, particularly in developing English-language or nonverbal tests. It would be particularly helpful if test developers oversampled Latina/os and other major ethnic groups. This would not only help in carrying out subsequent DIF analyses for specific ethnic groups, but would also enable test developers to examine within-group variance. Analyses by ethnicity, especially of English-speaking-only Latina/os, provide little opportunity to examine the role of language and immigration, among other potentially important variables. The tests that are shown to be reliable and valid with Spanish-dominant and immigrant subgroups are the tests that are more likely to be used with those subgroups that are the most difficult to assess.

For tests with multiple language versions and multiple norms, it is most important that the psychometric properties be assessed for all versions of the test. One can
not assume that the reliability and validity of the original English-language version applies to Spanish-language versions, even if they have been carefully translated, calibrated, or even equated. Also, Spanish-language tests and their norms should be revised periodically to maintain their relevance in our changing world (Lopez & Taussig, 1991). In all, the highest standards of test development should be applied to all tests, regardless of whether they are the original English-language versions or adapted Spanish-language versions.

For many Mexican Americans in the United States who were born in the United States and whose dominant language is English, available psychological tests can be applied effectively. In fact, the lack of systematic test bias with this segment of Mexican Americans (see Sattler, 2001, for a review) suggests that traditional assessment approaches are likely to be valid. For Spanish-dominant Mexican Americans, particularly those who are recent immigrants, U.S. English-language standardized tests have significant limitations. In this chapter, we found advances in a number of areas of selected cognitive ability and psychodiagnostic tests. However, there is still much to accomplish before we have an adequate set of assessment tools that consider the complex and dynamic nature of culture. The development of multiple-normed tests, norms that reflect U.S. English-language norms and Mexican Spanish-language norms, is of greatest value in assessing persons of Mexican origin, especially those who are Spanish-language dominant and are recent immigrants. Moreover, the availability of well-established nonverbal tests can be of great help (Figueroa, 1990). Given the rapidly increasing number of Mexican Americans throughout the United States (U.S. Census, 2001a), it is imperative that the proper assessment tools be developed and made available to address the complexities of language, culture, and psychological functioning.

REFERENCES


Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. Psychological Methods, 5, 125–146.


