EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE
NCME national council on measurement in education

A Cross-Cultural, Comparative Study of the American, Spanish, and Mexican Versions of the WISC-IV

Journal:	*Educational Measurement: Issues and Practice*
Manuscript ID:	EMIP-SB-09-062
Manuscript Type:	Original Article
Keywords:	cross-cultural, adaptation, WISC-IV

Abstract

Test adaptation and use in different cultural contexts remain an issue in the field of educational measurement. This work compares three different standardized versions of the *Wechsler Intelligence Scale for Children - Fourth Edition* (*WISC-IV*): The original American version, the translated version with norms specifically derived for use with the American Hispanic population, and the Mexican version of the same test with its own norms, which was adapted for the Mexican population. Differences in the translations, the format of the test, and the interpretation of scores are reported and discussed. For example, if different norms yield different standardized scores from the same raw scores, then this could reflect actual differences in the population and may suggest possible biases from translation, adaption, and norm development. Methodological, technical, and cultural issues may influence the measurement of intelligence in a population with cultural and educational differences from the original target population.

Keywords: Cross-cultural, test adaptation, *WISC-IV*

A Cross-Cultural, Comparative Study of the American, Spanish, and Mexican Versions of the *WISC-IV*

The question regarding the appropriateness of the use of tests of intelligence and cognitive abilities developed in the United States to assess people from other countries was renewed in the debate between Suen and Greenspan (2008, 2009a, 2009b) and Sánchez-Escobedo and Hollingworth (2009). This controversy arose from challenges in the translation, adaptation, and norm development of an intelligence test (in this case, the *Wechsler Adult Intelligence Scale - Third Edition*) for use with the Mexican population. Suen and Greenspan argued against the use of Mexican norms in high stakes decision cases regarding Mexicans (i.e. the death penalty). Sánchez-Escobedo and Hollingworth claimed that despite some error and misgivings in the process of norm development, versions adapted to measure intelligence in other cultural contexts have the potential to provide useful information about the test taker and they should not be written off.

In this current work, we attempt to advance and extend the scrutiny of intelligence tests in bilingual or Spanish-speaking populations by examining differences in three versions of the *Wechsler Intelligence Scale for Children - Fourth Edition* (*WISC-IV*): the tests and norms created for use with English-speaking American children, the tests and norms created for Spanish-speaking children in American schools, and the tests and norms created for Spanish-speaking children in Mexico. This is an exploration of the challenges and demands for adapting tests beyond mere language translation and the effects of using norms developed in different cultural contexts to evaluate test takers.

The Wechsler Intelligence Scale for Children (WISC)

The *WISC* was first published by Wechsler in 1949. A revised edition was published in 1974 as the *WISC-R*. The third edition (*WISC-III*) was released in 1991. The current version, the *WISC-IV*, was published in 2003. The *WISC-IV*, like its predecessors, will likely be the most widely used children's intellectual ability assessment in the United States (Prifitera, Weiss, Saklofske, & Rolfhus, 2005).

The *WISC-IV* is the culmination of a 5 year research program in the U.S. (Psychological Corporation, 2003). It followed an iterative process, where each phase led to further refinements of the scale. Several pilot studies were administered prior to the National Tryout Stage. Then, all the accumulated evidence from the previous stages was reviewed and the research questions were reexamined. These data were used to further refine a standardization edition of the *WISC-IV*. The standardization stage included administering the test to a stratified sample of 2,200 children aged 6 through 16 years 11 months. This sample was stratified on the following key demographic variables: age, sex, race, parent education level, and geographic region. It contained equal numbers of males and females. The proportions of Whites, African Americans, Hispanics, Asians, and other racial groups were based on the corresponding age groups from the March 2000 census data of the U.S. population. Steps were in place to ensure quality examiners were recruited to administer the test.

Exclusion criteria for the *WISC-IV* standardization sample included the following: if a subject had been tested on any IQ test in the last 6 months, uncorrected visual impairment, uncorrected hearing loss, not fluent in English, primarily nonverbal or uncommunicative, upper

extremity disability that would affect motor performance, currently admitted to hospital, mental, or psychiatric facility, and currently taking medication that might depress performance (Psychological Corporation, 2003). Children with any of the following previously diagnosed physical conditions or illness were excluded as well: stroke, epilepsy, brain tumor, traumatic brain injury, brain surgery, encephalitis, and meningitis. However, to accurately represent the children attending school in the U.S., a representative proportion (about 5.7%) of children from special group studies were added to the norming sample.

There are 10 core subtests and 4 supplemental subtests in the *WISC-IV*. A 5th supplemental subtest, Word Reasoning, was removed for testing, norming, and reliability issues. The scaled scores for the 10 core tests sum to four different indices (the Verbal Comprehension Index, the Perceptual Reasoning Index, the Working Memory Index and the Processing Speed Index). The *WISC-IV* also provides a Full Scale IQ (FSIQ), which ranges from 40 to 160 points. Supplemental subtests can be given for a broader sampling of intellectual and cognitive abilities. They may also be used as substitutes for selected core tests. The typical administration takes between 65 and 80 minutes for most children.

There have been many translations and adaptations of the *WISC-IV*, and norms have been established for a number of countries and languages[1]. The *WISC-IV* reports an enhanced clinical utility for providing assessment of fluid reasoning, working memory, and processing speed (Psychological Corporation, 2003). Each successive version also claims to be less biased against minorities and females than previous editions, and each purports to make the administration

[1] For example, Spanish (U.S., Spain, and Mexico), French (France and Canada), German (Germany, Austria, and Switzerland), English (Canada, and United Kingdom), Welsh, Dutch, Japanese, and Chinese, among others.

more effective, since the test developers have considered input from practitioners and experts in the field.

The *WISC-IV Spanish* is a comprehensive adaptation of the *WISC-IV* and was developed for use with Spanish-speaking children who are learning English as a second language and are acculturating to the U.S. (Psychology Corporation, 2005). For the purpose of this discussion, the *WISC-IV Spanish* will be referred to as the *WISC-IV Hispanic* to avoid possible confusions between the language and the population targeted. The overarching goal of the Hispanic adaption was to develop an instrument equivalent to the *WISC-IV* with items that elicit the same response processes and measure the same construct. It was designed to be representative of Spanish-speaking children of diverse backgrounds living in the U.S. (Psychological Corporation, 2005, p. 52). This version was standardized with 851 subjects in a stratified sample comprised of children from various Hispanic origins living in the U.S., including Cuba, Puerto Rico, and Central and South America. Like the original *WISC-IV*, the sample drawn for the Hispanic version was stratified on age, sex, parent education level, geographic region of the U.S. However, this sample was also stratified by country of origin.

Test items were revised and some were re-worded to minimize cultural bias across multiple regions of origin. While the test items are presented in Spanish, children earn credit for answers in either Spanish or English. Results are comparable to all U.S. children the same age. Supplemental tables allow additional interpretations based on comparisons with all Hispanic children or by subgroups of the Hispanic population (e.g., Puerto Rican, Cuban). The *WISC-IV Hispanic* is appropriate to use when the child is Spanish–language dominant, is in his or her first

five years in the U.S. education system, or is referred for neuropsychological evaluation for educational diagnosis and services.

The Mexican version of the *WISC-IV* was published in 2007 (Sánchez-Escobedo, 2007). Like the *WISC-IV Hispanic*, it is a comprehensive adaption of the *WISC-IV*, and both the Hispanic and Mexican editions followed recommendations for translation, adaptation, and best practices put forth by the *International Guidelines for Test Use* (International Test Commission, 2001) and the *Standards for Psychological and Educational Testing* (AERA, APA, & NCME, 1999). The standardization sample for norm development consisted of 1,234 Mexican children in 11 age groups, with an average of 112 subjects per group. Participants were drawn from 12 of the 32 states in Mexico. Children with obvious physical or intellectual disabilities and those children whose first language is not Spanish (i.e., Mayan, Zapotecan, Nahuatl, etc.) were excluded from the sample. The sample was stratified on age, sex, and type of school (private or public).

The Mexican adaption of the *WISC-IV* was necessary for several reasons. First, the *WISC* was not adapted for use in Mexico, so there was not an appropriate version to use in Mexico. Second, when the *WISC-R* emerged, it was found that when American norms were used, scores were roughly 15 points below the expected mean for the three main scales (Padilla, Roll, & Gomez, 1982). The consequently adapted version, the *WISC-RM* (Revised for Mexico), with norms adjusted for Mexican Children was widely used until 2007 (Esquivel, Heredia, & Lucio, 2007). However, practitioners discovered eventually that this test tended to overestimate the IQ. Many hypotheses were posited: the inclusion of many children with learning difficulties in the original sample, or procedures to compensate for the underestimation of IQ with the original

WISC-R, etc. Practitioners in Mexico called for a new and properly adapted version of the *WISC* test and norms specifically developed for Mexico.

In summary, all three versions of the *WISC-IV* under consideration consist of the same 10 core subtests, plus 4 supplemental subtests. In addition, all three were designed for use with children in the same age ranges: between 6 years and 16 years 11 months. In general, they require comparable administration time and all provide Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed, and Full Scale IQ scores. However, there are some substantive differences among the three versions. Table 1 summarizes the major features and differences of the three versions under scrutiny. What follows is an examination of the possible reasons for inconsistencies between the versions.

Differences in Forms

The three Record Forms were reviewed and qualitatively compared. The Hispanic form has a section for percentile rank adjustment depending upon parental level of education, U.S. educational experience, or both. On the Mexican version, the table to estimate the mean scores of the subtests was moved to the front page for stylistic purposes and to conserve space. On the Analysis page, the most salient difference is that the Mexican version uses a pre-established statistical significance level of $p \geq .05$ to estimate discrepancies and to facilitate scoring, since this is the common significance level used for interpretation. In the Mexican standardization process, consulted experts suggested that the inclusion of these figures would encourage screening for strengths and weaknesses. In addition, the Mexican version has slightly larger fonts and figures than the Hispanic and American versions.

The most significant difference among the subtests themselves is the ordering of items in each version based on the Index of Difficulty derived from the standardization process. In short, some items were easier (or harder) for Mexicans and Hispanic test takers than for American test takers. For example, in *Vocabulary*, the concept of *bicycle* was easier for Mexicans, ranking 10 on the American form and 12 on the Mexican form, whereas the term *brave* was easier for Mexicans ranking 10 for the Mexican form and 12 for the American one. The same order of items was found in subtests where the order is irrelevant to the test administration: such as *Digit Span, Coding, Letter-Number Sequence, Symbol Search,* and *Cancellation*.

The second important difference found was the number of items included in some routines. For example, in *Similarities*, the Hispanic version has 24 items, whereas the Mexican and American versions have 23 items. Likewise, in the *Comprehension* section, the American and Mexican versions have 21 items, whereas the Hispanic has 20. This different number of items may be due to ceiling effects (i.e. no participant responded to item 21 on the Hispanic version).

A third important difference becomes apparent when contrasting the Mexican and Hispanic versions with the American; some items have been replaced. The Hispanic and Mexican versions have different items in *Similarities*: For example, *pen/pencil, rubber/paper*, and *picture/statue* are found in the Mexican version, whereas *candle/light, guitar/drum*, and *ball/wheel* are only found in the Hispanic version. In *Vocabulary*, the Mexican version contains words such as *remedar*, *emigrar*, and *disparate* to replace the culturally inappropriate terms from the Hispanic version: words such as *garrulo*, *enmienda*, and *alemador*, which are quite unusual terms in Mexican Spanish. Likewise, in *Comprehension*, the Hispanic version has concepts such

as *doctors*, *newspapers*, and *monopoly*, whereas the Mexican version was edited to read *medics*, *news*, and *owner*. In *Letter-Number Sequence*, the Mexican version provides more practice items.

The Mexican version is completely written in Spanish, including directions to the test administrator. However, the Hispanic version has directions and the names of the subtests written in English. Thus, the test administrators for the Hispanic version need to be bilingual. Not surprisingly, differences were found between the Hispanic and Mexican version in the two verbal tests: *Similarities* and *Vocabulary*. Words like *plátano* (banana) are given multiple variants in the Hispanic version (e.g., *banana*, *guineo*) according to the language variation found in Hispanic children from different national origins.

Regarding protocols for test administration, the Mexican *WISC-IV* technical manual recommends that if the complete battery is to be administered, it should be done in two sessions: with a break of a minimum 30 minutes to 24 hours (Sánchez-Escobedo, 2007). This followed from observations made during the standardization process that some of the Mexican children became tired and distracted during the test administration. For the majority, it may be due to their lack of experience to testing routines such as this one.

Norm Comparison

In order to examine differences in features and possible implications of using one set of norms or another, a fictional raw score profile was created. Then raw scores were transformed using two hypothetical cases: a 7 and 16 year old. Obtained standardized score profiles provided a frame of reference to compare the three set of norms of the *WISC-IV*. The American version served as the baseline to compare the Hispanic and Mexican versions. Table 2 summarizes the raw scores and the standardized scores calculated using norms from each of the three versions of

the *WISC-IV* under scrutiny, from which the composite scores were calculated and compared (see Table 3). Figures 1 and 2 provide plots of the results and reveal similar patterns in the profiles of subtest scores and composites scale scores across the three versions.

When compared against American norms, both the Hispanic and the Mexican versions tend to underestimate the FSIQ of high aptitude 7 year olds, whereas they tend to overestimate the performance of low aptitude 16 year olds. In almost every case, the Mexican norms tend to differ more from the American norms than the Hispanic norms do. It can be observed that similar patterns of the indexes tend to be the same across the three versions. One substantive difference is the Hispanic norms tend to overestimate Working Memory when compared with the other two. Likewise, Mexican norms tend to overestimate Perceptual Reasoning and Verbal Comprehension.

Discussion

In this section we provide some clues to help us understand the differences among the three forms. Most importantly, we attempt to provide some guidelines in making decisions regarding what test to use, how to deal with differences, and what to expect in future development of norms and adaptation to other contexts for psychological test batteries. Administration criteria seem to be clear; the American *WISC-IV* should be used when cultural immersion and English language competency is warranted. English Language Learners in the U.S. should be assessed with the Hispanic form and Mexican children should be assessed with the Mexican form.

During the translation and adaption of a test, it is important to make sure that the test is understandable to the test takers, the directions are easy to comprehend, and the items are

ordered on an appropriate scale of difficulty. In general, translation of verbal routines seems to be appropriate, and the differences between the Hispanic and Mexican version language translations are minor. Given that the Mexican children are usually less exposed to large-scale testing than American students, and to reduce possible sources of bias, it is appropriate that the Mexican version provides additional practice exercises prior to some of the subtests (Geisinger, 1994). This reduces the effect familiarity with the item format might have on the score.

Further research is needed to study the effects of using different norms for a given version of the test across all the ages. For example, what happens when the American version is interpreted with Hispanic or Mexican norms? In addition, because scores for special populations require the use of American norms, it casts doubt on the accuracy of the scores and the interpretations that can be made. One interesting question to consider in future research is "would the estimation of an average score, derived from the use of the three set of norms provide a better estimate of the child's intelligence?"

It is not surprising to see differences in results using the different set of norms, since the norms were developed for use with different populations. These inconsistencies may in fact be due to meaningful differences in the characteristics of the population taking the test. For example, to understand how culturally different American and Mexican public schools are from one another, consider this: of the 87% of Mexican students attending the public Mexican educational system, 53% of these students started their formal education in first grade, and 90% of them only attend school part time for the primary levels; In addition, in Mexico there is a dropout rate of 22% in 7th grade and only 8% of Mexico's population above the age of 18 has a

bachelor's degree or its equivalent (Santibañez, Vernez, & Razquin, 2005; INEGI, 2009; INEE, 2009).

Different set of norms for a test are appropriate when it is administered to qualitatively differential populations. In this case, the cultural backgrounds, educational experiences, and testing experiences are vastly different across the three groups. Few Mexican public schools provide services such as hot meals and transportation. Furthermore, special education and psychological and counseling support is partial and inconsistent, and schools suffer from rotating teachers and scarce resources. For example, whereas the Mexican government invested an average of $1,350 US per student, the United States invested an average of $11,293 in 2005 (U.S. Department of Education, 2009; INEGI, 2009). These figures in and of themselves are reason enough to warrant different adaptions of a test. Different norms are not unexpected because, beyond language, the target populations have greatly different educational and social backgrounds, which will impact what is being measured.

A second major source of variation in the norms has to do with the test itself and the previous experience students have had with similar tests. It is well known that a previous administration of the *WISC* may explain an increase between 7 and 10 points in the score of the second administration. Likewise, a lack of exposure to testing in the Mexican educational system may explain many of the observed differences in norms.

Hispanic test takers in the United States may have experienced Mexican schooling, but they are in transition into a more comprehensive American school system by the time they take the *WISC-IV Hispanic*. These students often receive assistance, meals, transportation, English language support, and have undergone previous testing for placement purposes. It is therefore

reasonable to expect that test takers in these three circumstances would perform differently. As Obgu (1994) asserted, IQ tests are constructed to measure a specific aspect of intelligence and "the cognitive skills tapped by these tests are those that Western cultures emphasize in their formal schooling" (p.369). If looked upon in this light, the inconsistencies in the score profiles can be expected.

Conclusions

The original *WISC-IV* have been adapted due to different semantic variations in the target populations. This is appropriate and necessary to increase the level of understanding and to foster successful test administration. It is therefore not surprising that the language differs between the Hispanic and Mexican versions. In general, the format variations serve to facilitate the administration and scoring of the test. This in turn may decrease the biases and obstacles involved in cross-cultural intelligence testing.

Regarding norms, the differences in standardized scores are a result of different sets of norms and are expected. For example, the Mexican norms tend to estimate higher IQ than the Hispanic and American norms for children with relative low competence, whereas Mexican norms tend to underestimate IQ of high competence children. Indeed, the long standing debate about fairness in cross-cultural intelligence testing is revived by the issues described in this comparison of different versions of the same test. In particular, it is imperative that test administrators select the appropriate edition of both the tests and the norms for the target population to ensure the scores are interpreted in their cultural context. Teachers should be aware of the limitations and boundaries of intelligence tests such as the *WISC-IV*. In this case, teachers must consider the effects of using a test created in one population and then administered in

another, and they also must consider if the defined construct is present in the target population and how well it is measured by the adapted test. As Garcia-Coll and Magnuson (1999) assert, "basic psychological and behavioral constructs might not mean in one culture what they mean in another" (p. 10).

Tests adaptation and norm development in different cultures seem to be a renewed field of interest in educational psychology. It is apparent that there are various interesting avenues of research that can be undertaken in the future to address the questions raised in this study.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Esquivel, F., Heredia, C., & Lucio, E. (1999). *Psicodiagnostico Clinico del niño* (2 edición). Mexico DF: El Manual Moderno.

García-Coll, C. & Magnuson, K. (1999). Cultural influences on child development: Are we ready for a paradigm shift? In A. Masten, (Ed.) *Cultural processes in child development: The Minnesota symposium of child psychology.* Vol. 29, pp. 1-24.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaption issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 303-312.

Instituto Nacional de Evaluación Educativa INEE. (2009). Retrieved from http://www.inee.edu.mx/

Instituto Nacional de Geografía, Estadistica e Informatica INEGI. (2009). Retrieved from http://www.inegi.org.mx/inegi/default.aspx

International Testing Commission. (2001). International guidelines for test use. *International Journal of Testing, 7,* 91-106.

Ogbu, J. U. (1994). From cultural differences to differences in a cultural frame of reference. In P.

Greenfield, M., & Cocking, R. R. (Eds.). *Cross-cultural roots of minority child development* (pp. 365-391). Hillsdale: NJ: Lawrence Erlbaum Associates.

Padilla, E. R., Roll. S., & Gómez, P .M. (1981). Ejecución del WISC-R en adolescentes Mexicanos. *Intermaerican Journal of Psychology 16*(2), 122-128.

Prifitera, A., Weiss, L. G., Saklofske, D. H., & Rolfhus, E. (2005). The WISC-IV in the clinical assessment context. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives.* (pp. 33-71). San Diego, CA: Academic Press.

Psychological Corporation, The. (2003). *WISC-IV Techical and interpretive manual.* San Antonio, TX: Harcourt Assessment.

Psychological Corporation, The. (2005). *WISC-IV Spanish Manual.* San Antonio, TX: Harcourt Assessment.

Sánchez-Escobedo, P. (2007). *Validación y normas para México de Escala Weschler de Inteligencia para Niños IV*. México: El Manual Moderno.

Santibañez, L., Vernez, G., & Razquin, P. (2005). *Education in Mexico: Challenges and opportunities.* Santa Monica, CA: The Rand Corporation.

Suen, H., & Greenspan, S. (2008). Linguistic sensitivity does not require one to use grossly deficient norms: Why U.S. norms should be used with the Mexican WAIS-III in capital cases. *Psychology in Intellectual and Developmental Disabilities*. Official publication of Division 33, American Psychological Association. Available online at: http://www.apa.org/divisions/div33/docs%5C34-1.pdf

U.S. Department of Education, National Center for Educational Statistics, Institute of Education Sciences. (2009). *The Condition of Education.* (NCES Publication NO. 2009081). Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/pubs2009/2009081_1.pdf. For summary of school expenditures see Table A-34-1 at http://nces.ed.gov/programs/coe/2009/section4/table-tot-1.asp.

Table 1

Comparison of the 3 versions of the *WISC-IV*

	WISC-IV version		
	USA	Spanish	Mexican
Sample size	2,200	851	1,234
Simple type	Random stratified	Conventional stratified	Conventional stratified
Administration time (minutes)	65-80	70-90	60-90
Number of norm groups	33	33	20
Publication date	2003	2004	2007
Authors	Rolfhus & Zhu	Harris & Williams	Sánchez et al.
Publishers	Harcourt Assessment /PsychCorp	Harcourt Assessment /PsychCorp	Manual Moderno
Number of sessions	1 Session	1 Session	2 sessions if 14 subtests

Table 2

Raw to Scale Score Conversions for 3 Versions of the WISC-IV

		Age 7			Age 16		
Subtest	Raw score	English	Spanish	Mexican	English	Spanish	Mexican
Similarities	22	6	5	7	16	14	16
Vocabulary	34	5	5	6	16	16	16
Comprehension	21	5	5	7	14	15	14
Information	16	5	6	6	15	16	15
Block Design	34	6	6	7	15	15	16
Picture Concepts	14	4	4	6	12	12	12
Matrix Reasoning	17	5	5	6	13	13	14
Picture Completion	19	3	3	5	11	11	11
Digit Span	16	8	10	8	14	16	14
Let-Num Seq.	15	5	6	6	13	14	12
Arithmetic	17	4	4	6	12	13	13
Coding	32	1	1	2	7	7	6
Symbol Search	23	5	5	8	11	11	11
Cancellation	60	3	3	5	11	11	11

Table 3

Composite Scores for 7 and 16 Year Olds

Scale	Sum of scaled scores	Age and version of WISC-IV					
		7			16		
		English	Spanish	Mexican	English	Spanish	Mexican
Verbal Comprehension	22	124	130	132	73	71	81
Perceptual Reasoning	34	121	121	125	69	69	77
Working Memory	21	120	129	116	80	88	83
Processing Speed	16	94	94	91	62	62	73
FSIQ	22	128	125	123	57	66	65

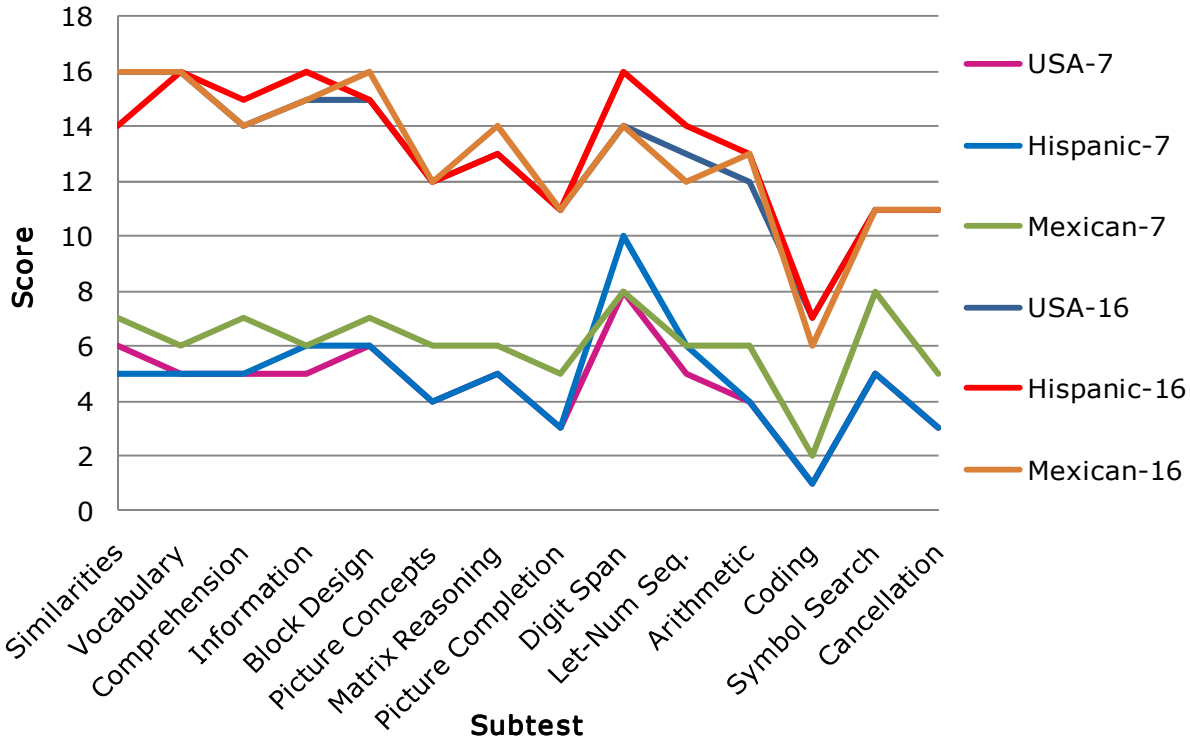

Figure 1. A comparison of the subtest scaled scores from 3 sets of norms. The top 3 lines show the score pattern for a hypothetical 16 year old and the bottom 3 lines show the score pattern for a hypothetical 7 year old.

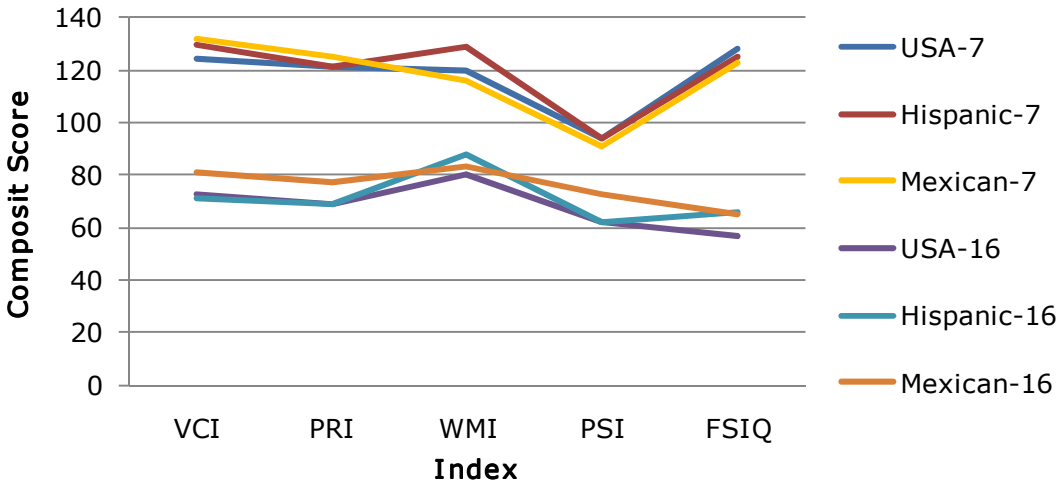

Figure 2. A comparison of the composite scores from 3 sets of norms. The top 3 lines represent the pattern for a hypothetical 7 year old and the bottom 3 lines represent a pattern for a hypothetical 16 year old.