EDUCACIÓN Y CIENCIA
EDITORIAL
UADY
FACULTAD DE EDUCACIÓN
http://www.educacionyciencia.org

Retos en la medición de la inteligencia en México: Lecciones aprendidas de la estandarización de las escalas de Wechsler

Challenges Measuring Intelligence in Mexico: Lessons Learned from the Standardization of the Wechsler Scales

Pedro Sánchez Escobedo
Universidad Autónoma de Yucatán, México.

Información adicional sobre este manuscrito escribir a:
Pedro Sánchez Escobedo, psanchez@correo.uady.mx

Cómo citar este artículo:
Sánchez Escobedo, P. (2015). Challenges Measuring Intelligence in Mexico: Lessons Learned from the Standardization of the Wechsler Scales. *Educación y Ciencia, 4*(43), 65-79.

Resumen

Para evaluar la inteligencia en México, tanto en ambientes clínicos como en el sistema educativo mexicano, se utilizan ampliamente las tres escalas de Wechsler, en niños desde los tres años de edad hasta adultos mayores. En este artículo se analizan los procesos de adaptación y estabilización de estas escalas en México, y se reflexiona acerca de las dificultades y retos que implicaron estos procesos con la finalidad de identificar elementos claves que nos permitan mejorar, en el futuro, el proceso de adaptación de escalas estandarizadas en México. Las escalas de Wechsler son los instrumentos más utilizados para identificar discapacidad de aprendizaje, discapacidad intelectual, talento y sobredotación; por lo que la discusión del proceso de estandarización de adaptación en México es importante tanto desde el punto de vista psicométrico como desde el punto de vista práctico. Derivada de la experiencia de estandarizar estas escalas en México, este artículo pretende aportar a la discusión sobre las ventajas y dificultades de la adaptación de pruebas diseñadas en los Estados Unidos en México. Examina de manera crítica los procedimientos de recolección de datos, análisis de los mismos y la pertinencia de su uso. En particular, se identifican asuntos clave tanto técnicos como metodológicos y las implicaciones culturales de la medición de la inteligencia México, en un intento de facilitar las futuras adaptaciones y estandarización de ese tipo de pruebas.

Palabras clave: inteligencia, pruebas psicológicas, estandarización

Abstract

In both educational and clinical settings, the most frequently used instruments to measure intelligence in Mexico are the three Wechsler Scales (for children ages 3 years old through older adults). Because the Wechsler scales are the most important tools in screening for intellectual disability, learning difficulties, and giftedness; critical points for the standardization in Mexico of the three scales are discussed, such as differences in translation, formats, and appropriateness of items, and idiosyncrasies of test takers in Mexico. In addition, empirical evidence to help select which test to use when norms for a given age overlap is provided to guide the users. Derived from the experience in participating in the standardization to these scales in Mexico, this article contributes to the discussion on the advantages and limitations of using American made tests in other cultures. Beyond the psychometric properties of these tests, this article focuses on the process of test adaptation and norm development by critically examining the procedures of data collection and data analysis. Finally, methodological, technical and cultural issues of measurement of intelligence in Mexico are examined in an attempt to facilitate future adaptations and standardizations of intelligence tests developed in the United States and used in other countries.

Keywords: intelligence, psychological testing, standardization

Introduction

Standardization of a particular test in another context, refers to transformations that involve an adjustment of means and/or standard deviations of either individuals or groups, or both (Fischer, 2004).

Describing the process of adapting intelligence tests for new cultural contexts is important because the methods and procedures of adaptation can affect the reliability of the scores obtained and their interpretation in a specific setting.

The Wechsler tests have been adapted in many countries, for example, in the United Kingdom the adaptation of the WMS-III took the form of a validity and comparability study between the American norms and the scores of a representative sample of the UK population. This process is acceptable since the same language and many cultural conditions are similar to the context where this test was originally developed. Thus investigators needed only to provide sufficient information to allow well-informed use of the USA norms in the UK. However, in other English speaking countries, such as India, cultural differences, test familiarity and educational background of respondents were considered as confounding factors for the test results. Hence, investigators in this case carried out changes to items in six of the ten subtests and reported new Reliability Coefficients, and percentiles ranges based on Indian normative data (http://pearsonclinical.in/solutions/clinical-assessment-intervention/adult/memory/wms-iii-india/).

Khaleefa (2006) compared the WISC-III adaptations for Japan and Sudan and reported high levels of reliability for both countries and the same factor structure. The only consistent difference was better performance in visuospatial tasks for Sudanese children, whereas Japanese children performed better in verbal tasks.

After revising the standardization of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III) in sixteen countries, Georgas, Weiss, Van de Vijver, and Saklofske (2003) suggested that in spite of differences across cultures, there were more similarities than differences among the translated forms. They observed that in every country this test has shown the same factor structure and similar psychometric properties. The authors warn, however, that the interpretation and use of the test in each country should take into consideration the particularities of each culture.

The purpose of this paper is to reflect upon the process, results, and experiences of the adaptation and standardization of the three major Wechsler scales used to measure intelligence in educational, working, and clinical settings in Mexico. It is intended to identify advantages and limitations of the use of these scales, to pinpoint needs and challenges, and to facilitate future adaptations of new versions and standardizations or norms of these tests.

The Wechsler Scales in Mexico

Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III)

The Mexican Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) is the most recently published test (2011). Its development was planned in two phases: the first was considered a national trial and comprised a sample of 1,801 children

from 21 of the 32 states in Mexico. This first attempt helped to identify the ambiguous translation of certain items and to correct some artifacts in the response protocols.

The actual standardization phase included 829 children from four major regions of the country, clustered in nine age related groups. Exclusion criteria included children with disabilities, obvious chronic disease and to those children whose birth language is not Spanish (for example, Maya, Nauhatl, etc.). Data for this article were collected from the national trial.

The WPPSI-III claims to place less emphasis on acquired knowledge than the other Wechsler tests and features shorter, more game-like activities that hold the attention of children as young as 2-1/2 years. Simplified instructions and scoring procedures enhance the ease of administration for examiners. Both children and examiners benefit from the thoughtful, carefully constructed revisions implemented to build a highly respected, reliable test that completely reflects what users wanted for preschool children. The range of testers is from 2 years 6 months to 7 year old.

Younger children take fewer subtests that aim to measure verbal comprehension and perceptual organization abilities. Older children take a greater number of subtests designed to measure verbal comprehension, perceptual organization, and processing speed abilities and the use of queries and prompts is generally unrestricted (http://www.pearsonassessments.com).

The Wechsler Intelligence Scale for Children – 4th Edition (WISC-IV)

The Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV) was published in 2003 in the United States and in 2007 in Mexico. This is the battery most widely used to assess IQ in both countries (Prifitera, Weiss, Saklofske, & Rolfhus, 2005; Sánchez-Escobedo, 2007).

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 were excluded from the sample. The sample was stratified by age, controlled for gender and type of school (private or public).

In Mexico, it was particularly important to standardize this test because the WISC-III had not been published or used in Mexico, and previously derived Mexican norms tended to overestimate Mexican children's IQ when compared with American norms (Padilla, Roll & Gomez Palacio, 1982; Esquivel, Heredia & Lucio, 2007).

Sánchez-Escobedo & Hollingworth (2012) examined the psychometric characteristics of this test, and through a confirmatory factor analysis and inter-correlational studies provided information on the WISC-IV. Factor loadings and correlational patterns were found to be comparable to those seen in the American versions of the test.

The Wechsler Adult Intelligence Scale - Third Edition (WAIS-III)

In 2001, a preliminary version of the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) was adapted for use in Mexico to adjust norms solely for the group of reference. Then, a more extensive standardization process was carried out in 2003 to establish norms by age bands. In the preliminary phase, a translated version was

administered to a group of 287 persons between the ages of 16 and 70 years old controlling for gender and previous school experience, in an attempt to have a conventional sample, balanced for gender and educational level. In 2004, newly revised norms were published to adjust some discrepancies reported by users of this test.

Results from this first effort set the foundation for the standardization process, adapting and improving the previous version of the test and administering the battery to 970 Mexicans, from six different regions of the country. Previous experience helped to discard about 3% of tested individuals due to severe violations of the protocol of administration. The standardization sample controlled for gender, region of the country, and age. Due to sample size, norms were developed for 10 age bands, in contrast to the 14 bands in the American test.

In 2014, a new version of the test The WASIS-IV was published. This new version that claims to consider new demographic and clinical tendencies will gradually replace the older version of this scale for adults in Mexico.

Table 1 depicts and compares some of the major features of each of the Mexican Wechsler scales analyzed in this study. In this section, the target demographic for each of the Wechsler scales is described and a brief history of the scales' adaptation for use in Mexico is provided.

Table 1. Comparative between the tests

		WIPSSI-IV	WISC-IV	WAIS-III
Sample Size	Pilot	1,801	-	287
	Final	829	1,234	970
Year	Pilot	2009	-	2001
	Final	2011	2007	2003
# Subtests	Verbal	5	5	7
	Non verbal	5	4	7
Age	Bands	9	17	13
	Ranges	2:6 -7:03	6-16:11	16-≥70

General comments of the Mexican Wechsler scales

Translation and management of items

According to Hambleton (1996), "the term 'adaptation' rather than 'translation' was preferred by the test adaptation guidelines committee because the former term is broader and seemed to more accurately reflect the process of preparing a test or instrument for use in a second language or culture" (p. 5). Adaptation refers to the translation of items and the verification that the way it was translated is meaningful and familiar to the test taker. The process of test adaptation also involves ordering items according to the degree of difficulty to a specific population.

In the case of the Mexican scales, adaptation was first performed by professional translators; the majority of items were simply translated. Further adaptation was then implemented after analyzing responses in the standardization processes. For instance, extra care was taken to adapt questions for idiomatic expressions, to adapt to information relevant to the Mexican geography and history; Mexican children were asked who Benito Juarez was (a Mexican hero), instead of 'Abraham Lincoln'. In other cases, additional positive and correct responses were considered. For instance, the item "*colony*" translated

as "*colonia*" in Spanish could be answered correct when the respondent asserted "perfume or fragrance", instead of the intended translation of "settlement in a distant territory".

However ranking of items was different than the original American versions since items were ordered according to the Facility index derived from the responses in the standardization samples. This decision was based considering the need of fairness in the discontinuation criteria pre-established for the tests. By ordering items accordingly to Mexican indexes, respondents are less likely to meet the discontinuation criteria due to the original test structure. Of course, scales such as *Digit Span, Coding, Letter-Number Sequence, Symbol Search,* and *Cancellation* remained unchanged.

Sampling methods

In Mexico, none of these tests used a randomly stratified sampling method due to many constraints: for example, budget restrictions, access issues in some school districts and regions, and the lack of an extended research network or acceptance of intelligence testing in the country. Hence, there is always the question about sampling methods and the degree to which the participants represent the intended population. More data should help assess this matter. For example, in every case, more than 50% of the sample size of the original American standardization procedures was reached.

For the standardization of the three tests, an effort was made to achieve some sort of balance of representation in the sampling. For example, there were attempts to balance for previous education, gender, and regional cultures of the country.

Norms derived from the standardization process of the Wechsler scales are representative of the population included in the sample, that is, for children and adolescents in the school system without obvious physical or mental disabilities. The adult scales are useful for functioning adults with various degrees of education and work experience. Rather than asking how representative the sample is for the Mexican population, the examiner must consider whether the subject to be tested resembles the characteristics of the Mexican standardization sample. Clear exclusion criteria were established for the three tests: they are not to be administered to people with obvious physical or mental disabilities, to people from rural communities, or to people whose native language is not Spanish. Thus, none of the tests have norms for special populations.

Test administrators

Test administrators in the United States are usually experienced psychologists or school counselors. In contrast, in Mexico advanced psychology students who were specifically trained for the purposes of the pilot testing administered most of the tests. Despite the training, results evidenced in some cases demonstrated a lack of care, a breach of pre-established protocols, or a lack of abidance to pre-established sampling criteria. In the future, testing conditions, scoring and data analysis should become a minimal source of error.

Differences in Record Forms and administration

Response protocols follow the general design of the original tests. The three Record Forms were reviewed and qualitatively compared. All three tests are completely written in Spanish.

In the case of the WISC-IV, 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. Likewise, the Mexican version uses larger fonts and figures than the American versions.

In the Mexican version, Word Reasoning, was removed for testing, norming, and reliability issues, because it made no sense to most Mexican children. For instance, questions such as "goes out at night" lead to responses like "my mom, when dad is on a trip", when it was supposed to be a lead for "Moon".

Another major difference was that in the case of Mexico, the manual allows for a break anywhere from 20 minutes up to 24 hours, to prevent fatigue of the children. This was derived from the observations during the standardization process that many of the Mexican children in the pilot group grew tired and distracted, due to their lack of previous exposure to testing routines such as this one. Formal recognition of the detrimental effects on the scores due to the lack of exposure and practice of these types of tests in Mexican children was an important advance.

Lessons Learned

Lessons learned led investigators to check and revise protocols to search for errors, inconsistencies, and even simulated responses in some rare cases. Over time, more rigid controls and supervision of data were necessary to avoid the problems reported in earlier phases of the test standardization process in Mexico. Data spreadsheets were audited up to three times to prevent missing or erroneous data, before analysis. Errors due to scoring decreased with every other standardization process beginning with the WAIS-III adaptation process in 1997, to the WIPSSI-III standardization in 2011.

In all three processes, the method of linear standard scores was used to calibrate norms and traditional statistical methods were used to analyze data, to correct errors, and to adjust norms. Scores in every test allowed for the establishment of scaled scores for the 10 core tests to calculate the four different indices: Verbal Comprehension, Perceptual Reasoning, Working Memory and the Processing Speed; and the total IQ. Table 2 depicts the difference among the three tests subscales.

Different statistical procedures were used to adjust means when atypical deviations from the expected mean of 10 was observed in a given sub-scale in a particular age band, attributing such deviation to sampling, scoring, or random errors. In future standardizations, investigators should ponder the effects of adjusting American norms with the Mexican derived data, rather than developing norms from scratch. For this, methods of mean and linear equating should be considered.

Table 2. Comparison of subscales across tests

Sub-tests	WPPSI-III		WISC-IV		WAIS-III	
	Type	# items	Type	# items	Type	# items
Block Design	E	0 - 40	E	0 - 68	E	0 - 68
Similarities	S	0 - 46	E	0 - 44	E	0 - 33
Information	E	0 - 34	S	0 - 33	E	0 - 28
Picture completion	S	0 - 32	S	0 - 38	E	0 - 25
Vocabulary	E	0 - 43	E	0 - 68	E	0 - 66
Matrix reasoning	E	0 - 29	E	0 - 35	E	0 - 26
Comprehension	S	0 - 38	E	0 - 42	E	0 - 33
Symbol search	s	0 - 50	E	0 - 60	E	0 - 60
Picture concepts	E	0 - 50	E	0 - 28	-	-
Coding	E	1 - 65	E	1 - 65	-	-
Letter-number	-	-	E	0 - 30	E	0 – 21
Digit span	-	-	E	0 - 32	E	0 – 30
Arithmetic	-	-	S	0 - 34	E	0 – 22
Word reasoning	E	0 - 28	S	0 - 24	-	-
Digit-symbol coding	-	-	-	-	E	0 - 133
Picture arrangement	-	-	-	-	E	0 – 22
Object assembly	-	-	-	-	E	0 – 52
Cancellation	-	-	S	0 - 136	-	-
Picture naming	S	0 - 30	-	-	-	-

Legend: S = supplementary; E = essential.

Which test to use?

One common question for practitioners is which test to use when two versions are available for the same age band. For instance, a seven -year-old child suspected of a learning disability can be tested with either the WIPPSI-III or the WISC-IV. In theory, the former test is easier and designed for younger children, and the upper bands of norms would be used to assess performance. The latter conveys more difficult demands, so the lower bands of norms are to be used. However the two instruments are different, containing different items and tasks. In the next section, we provide Mexican practitioners with the empirical information to guide decisions regarding which test to use.

From available data, raw scores were converted into standardized scores using norms derived for each test (WIPSSI, WISC or WAIS) and their overall mean calculated. Considering a theoretical mean of 10 for the population pre-established for each scale, differences in the means from the samples were inspected. Table 3 illustrates the ways overlapping age bands were used for this analysis.

Contrary to the general idea of this test being easier for older children, the use of the WIPSSI-III generally provided, in general lower scores, proving to be more difficult than the WISC-IV. Hence, when faced with the choice of the two tests, clinicians should be aware that children assessed with the WIPSSI-III will show lower scores, and thus will be more likely to be within ranges of mental disability and learning disorders, and less likely

to achieve standards for giftedness. Data would support advantages in the use of the WISC-IV when both tests are available.

Regarding the use of the WISC or the WAIS with 16-year-old children, there seems to be the same degree of bias in different directions: the WISC underestimates abilities whilst the WAIS overestimates them. Of course, further insight into which test to use will require administering both tests to the same subjects and compare the results. This proves to be an interesting venue for future research.

Table 3. Comparison of Mean of standardized scores in the sample per subscale in overlapping ages

Age 6	M-WIPSSI-III		DIF	M-WISC-IV		DIF
	M	SD		M	SD	
CD	7.88	(3.05)	-2.12	11.33	(3.61)	+1.33
VC	6.86	(2.95)	-3.14	12.50	(4.73)	+2.50
SM	6.64	(3.32)	-3.36	9.40	(4.97)	-0.60
MX	7.65	(2.48)	-2.35	11.16	(3.83)	+1.16
IF	7.45	(2.26)	-2.55	10.90	(4.14)	+0.90
SY	8.1	(3)	-1.90	10.1	(2.99)	+0.10
CM	6.87	(2.80)	-3.13	10.69	(3.91)	+0.69
Trend			-2.65			+0.86
Age 7	M-WIPSSI-III			M-WISC-IV		
CD	8.27	(2.52)	-1.73	10.92	(4.32)	+0.92
VC	6.27	(3.07)	-3.73	11.15	(4.27)	+1.15
SM	7.53	(3.14)	-2.47	9.94	(4.54)	-0.06
MX	6.55	(2.47)	-3.45	10.55	(4.01)	+0.55
IF	6.63	(2.57)	-3.37	10.37	(4.17)	+0.37
SY	8.1	(3)	-1.90	10.1	(2.99)	+0.10
CM	8.27	(2.52)	-1.73	10.92	(4.32)	+0.92
Trend			-2.63			+0.56
Age 16	M-WISC-IV			M-WAIS-III		
CD	9.07	(3.54)	-0.93	10.22	(3.20)	+0.22
VC	7	(3.98)	-3.00	16.69	(4.56)	+6.69
SM	8.59	(4.05)	-1.41	11.37	(0.65)	+1.37
MX	8.24	(3.28)	-1.76	10.20	(3.19)	+0.20
IF	9.07	(3.54)	-0.93	10.22	(3.20)	+0.22
SY	8.1	(4.3)	-1.90	10.1	(2.97)	+0.10
CM	8	(4.24)	-2.00	10.18	(3.02)	+0.18
Trend			-1.70			+1.28

DIFF: Difference form the theoretical mean of 10.

Additionally, Sanchez and Hollingworth (2010) have previously reported potential differences in score interpretation, depending on the norms used to create standardized profiles. The authors compared American, Hispanic, and Mexican norms in the WISC-IV, and reported that American norms tend to underestimate the IQ 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 Hispanic norms, and Mexican norms tend to overestimate Perceptual Reasoning and Verbal Comprehension when compared to American norms.

Although differences in standardized scores are expected when using different sets of norms, this continues to be a major issue regarding the external validity of the test.

Indeed, deciding the appropriateness of which set of norms or test is best to use needs further empirical data, derived from clinical and educational studies that depict how these batteries behave when confronting populations with known characteristics thoroughly documented by other clinical, psychometric, neurologic and image data. Table 3, compares mean standard scorers in overlapping ages.

Changes with age

By consider the three tests altogether, it was possible to explore how test adaptation behaves across the life span in the Mexican population. A comparison of variations in standardized scores across age was performed using 3 scales that persist across the life span. Figure 1 illustrates how the means of transformed scores for cubes design, vocabulary, comprehension and digit span, derived from the raw scores from the standardization samples, depart from the theoretical mean of 10 across the life span. It can be observed that there is a random variation from the expected theoretical mean of ten, either above or below the mean, with a mean difference of .6. Most likely, this minor variation is due to measurement error, sampling issues, or other testing artifacts, and they have little significance in practice, since variations seems to occur without a clearly identified trend.

However, clear underestimation of IQ above 60 years old is evident in this graphic. Thus, the future standardization of the next WAIS should pay attention to the development of norms for age bands above 70 years old, considering that currently, 9.7% of the Mexican population is older than 60 and the consequent need for better norms for age bands after 65 years old (UNAM, 2014).

Comparison by laterality and gender

Given that gender and laterality differences have been considered important in interpretation of scores, these variables were examined across the three tests. Table 4 compares means across the tests by gender. It can be seen that no significant gender differences were found across the standardization processes; this is in contrast with the expected slight superiority of females on verbal abilities reported in the extensive meta-analysis of Hyde and Marcia (1998).

Table 5 likewise shows that laterality was not a factor of significant statistical differences, except for vocabulary and matrix reasoning favoring left-handed people. Such differences may not be of practical significance and it could be assumed consistent with the arguments of Nettle (2003) arguing no practical differences in hand laterality and cognitive ability.

Table 4. Comparison of subscales by gender

Sub-tests	**WPPSI-III**			**WISC-IV**			**WAIS-III**		
Gender	M	F		M	F		M	F	
	X (SD)	X (SD)	t p	X (SD)	X (SD)	t p	X (SD)	X (SD)	t p
Block Design	23.06 (7.40)	22.85 (6.65)	.33 .73	31.39 (14.32)	29.67 (14.27)	2.11 .035	37.12 (26.88)	31.92 (12.79)	3.9 .001*
Similarities	15.75 (10.04)	15.88 (9.95)	-.15 .87	16.92 (9.46)	16.59 (9.35)	.62 .53	19.21 (12.27)	17.63 (6.68)	2.52 .012*
Vocabulary	16.58 (8.02)	16.92 (8.15)	-4.7 .63	30.66 (10.92)	29.75 (11.29)	1.44 .15	36.38 (13.05)	34.19 (12.84)	2.62 .009
Matrix reasoning	12.81 (5.40)	12.55 (4.97)	.56 .57	16.95 (6.04)	16.90 (6.26)	.11 .90	15.28 (10.77)	13.86 (9.17)	2.21 .027
Comprehension	14.23 (8.11)	14.59 (8.22)	-.51 .60	20.17 (7.64)	20.06 (7.41)	.26 .79	17.46 (7.21)	17.66 (7.34)	1.69 .09
Picture completion	17.18 (7.68)	16.97 (6.55)	.34 .72	22.22 (7.41)	21.85 (7.02)	.88 .37	18.02 (4.86)	17.45 (11.70)	.97 .32

Table 5. Comparison of subscales by Laterality

	WISC-IV			WAIS-III		
	L	R		L	R	
Sub-tests	X (SD)	X (SD)	t [p]	X (SD)	X (SD)	t [p]
Block Design	33.32 (15.83)	30.51 (14.28)	1.26 .20	35.96 (12.99)	34.35 (21.22)	.53 .59
Similarities	18.46 (9.60)	16.63 (9.40)	1.25 .21	19.01 (6.09)	18.37 (9.93)	.45 .64
Vocabulary	33.67 (12.51)	30.03 11.08	2.10 .036	35.27 (12.98)	35.26 (12.93)	.005 .99
Matrix reasoning	18.37 (7.10)	16.86 (6.09)	1.57 .11	21.25 (31.27)	14.18 (7.02)	4.96 .001*
Comprehension	21.16 (8.54)	20.01 (7.48)	.98 .32	17.88 (7.10)	17.01 (7.28)	.83 .40
Symbol search	23.06 (8.23)	22.7 (8.77)	.20 .84	28.11 (10.46)	25.69 (11.16)	1.51 .13
Picture completion	23.37 (7.31)	21.91 (7.26)	1.28 .19	18.39 (4.35	17.71 (9.35)	.51 .60

Discussion

Advantages of adapting the Wechsler Scales

During the translation and/or adaption of any test, it is important to make sure that the test is more 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 American and Mexican versions regarding language competencies are minor. On executive routines, there was less influence from erroneous contextual and cultural factors, because many pictures were adapted to portray Mexican children or situations.

Disadvantages of adapting the Wechsler Scales

Given that Mexicans are usually less exposed to standardized tests it is usually wise to provide more time and expanded explanations before testing.

In addition, testing conditions should be carefully revised because in many cases testing was performed in a classroom, in a laboratory, or even outdoors. It is also important to provide additional practice exercises prior to some of the subtests and to make sure that subjects understand directions and procedures (Geisinger, 1994).

Testers should be also trained and supervised. Practice and mastery in handling materials, providing directions and interacting with the test taker are conditions that should not be taken for granted.

The use of appropriate norms

Maybe the most crucial question in adapting and standardizing tests pertains to the appropriateness of the newly developed norms to judge the tester´s performance. Further research is needed to study the effects of using different norms for a given version of the test. For example, what happens when the American version is interpreted with Hispanic or Mexican norms? The Mexican form requires the American norms for special populations and suggests the comparison between sets of norms in case of doubt.

Different standard scores are naturally different depending on which test is used. Therefore, the key issue here is to decide, considering the characteristics of the test taker, which instrument would produce the most reliable information to compare against other data, whether it was psychometric or qualitative in nature, in order to contribute to a decision about the case. The fact that test results are only one element of the decision making equation is of paramount importance to bear in mind, not only to judge the case under study, but to assess the virtues and limitations of a given test used in the process.

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. Different scores may be associated 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 children attending the public Mexican educational system, 53% of these children started their formal education in first grade, and 90% of them attend school on part time basis (Santibañez, Vernez, and Razquin, 2005; INEGI, 2009; INEE, 2009). Furthermore, for immigrants, it is difficult to measure the degree of acculturation and thus decide whether to test in English or Spanish, and use Mexican, American or Hispanic norms.

Conclusions

In the adaptation of the Wechsler scales to Mexico, translations from the original versions are in general adequate, and language adaptations to different semantic variations according to the target population are appropriate. In these scales, major differences were found related to the number of items, its order of presentation, and the replacement of some items for others. In general, all format variations were designed to increase understanding of directions, facilitate administration and scoring of the test. In sum, test adaptation may reduce biases due to cultural differences, practice, familiarity with standardized tests and other factors affecting Mexicans performance on the test and increase fairness in intelligence assessment.

Regarding norms, as expected, differences in standardized scores derive when different sets of norms are used. A same raw score may yield different standardized scores depending upon the set of norms used. Further research is needed to elicit information to help us decide which scale is better and which set of norms are more appropriate to use.

Future research needs to be conducted using the two overlapping age scales with the same person, and should be evaluated with other clinical, social, and educational criteria,

thus indicating which of the two provides a better index of intellectual functioning under specific circumstances and in accordance with different decision making needs. Perhaps the estimation of an average score, derived from the use of the 3 set of norms will provide a better estimate of the person's intelligence. Indeed, the longtime debate around cultural fairness of intelligence testing is revived by the results shown in this work.

The debate on limitations and boundaries of popular intelligence tests used abroad is revived when considering options between the psychometric strengths of these scales and the challenges in using them in different cultural contexts. 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).

The interest of scholars in the advantages and limitations of using adapted versions of American tests and the challenges of developing norms in other cultural contexts continues to be a source of disagreement. For example, the debate between Suen and Greenspan (2008) and Sánchez-Escobedo and Hollingworth (2009) regarding the use of either the Mexican or the American versions of Wechsler Adult Intelligence Scale (WAIS-III) in a death penalty case, renewed the discussion over what test to use and how to interpret the results when cross-cultural issues are present.

In sum, the use of intelligence tests, despite cultural and contextual influences, are necessary for practitioners in need of empirical evidence to support a number of high-stakes decisions; however, the results of these tests should be interpreted with caution. Test adaptation across cultures is a renewed field of interest in school psychology that offers various interesting lanes for future research.

Referencias

American Educational Research Association, A. P. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Child, D. (1977). *Psychology and Teacher* (2nd ed.). Holt, Reinehart & Winston.L td.: London.

Esquivel, F., Heredia, C., & Lucio, E. (1999). . *Psicodiagnóstico Clínico del niño* (2da ed.). Mexico DF: El Manual Moderno.

Fina, A., Sánchez-Escobedo, P., & Hollinworth, L. (2012). Annotations on Mexico's WISC-IV: A validity study. *Applied Neuropsychology: Child* , *1*, 6-17.

Fischer, R. (2004). Standardization to accouent for croos-cultural response bias. *Journal of Cross Cultural Psychology* , 263-267.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment. Hoboken, NJ:John Wiley & Sons.*

García Coll, C., & Magnuson, K. (1999). Cultural influences on child development: Are we ready for a paradigm shift? . In A. Maste (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.*

Georgas, J., Weiss, L., Van de Vijver, F., & Saklofske, D. (2003). *Culture and Children's intelligence: Cross cultrual analysisi of the WISC-III.* San Diego: Elsevier.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton., P. F. Merenda, C. D, & Spielberger (Eds.). *Adapting educational*

and psychological tests for cross-cultural assessment (pp. 3-38). Mahway, NJ: Lawrence Erlbaum Associates.

Hyde, J. S., & Linn, M. C. (1998). Gender differences in verbal ability: A meta-analysis. *Psychoilogical Bulletin , 104* (1), 53-69.

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

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

Khaleefa, O. (2006). Adaptation of the WISC-III in Sudan and Japan: A cross cultural study. *Arapsynet.e.journal* , 149-154.

Nettle, D. (2003). Hand Laterality and cognitive ability: A mulpile regression approach. *Brain and Cognition* , 17-22.

Ogbu, J. U. (1994). From cultural differences to differences in a cultural frame of reference. In M. Greenfield, R. R, & Cocking (Ed.), *Cross-cultural roots of minority child development* (p. 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 D. H. Saklofske, G. L., & Weiss (Ed.), *WISC.IV clinical use and interpretation Scientist-practitiones perspectives* (pp. 33-71). San Diego CA: Academic Press.

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

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

Reschly, D. J. (1981). Psichologial Testing in Educational Classification and Placement. In *American Psychologist* (Vol. 36, pp. 1094-1102). Iowa State University.

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

Sanchez, P. (2001). Capítulo 6: Validación Preliminar para México. In D. Tulsky, & J. Zhu, *Manual Técnico del WAIS-III.* México DF: Manual Moderno.

Sanchez, P., Canton, B., & Sevilla, D. (1999). *Compendio de Educación Especial.* Méico DF: 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 *in capital cases. Psychology in Intellectual and Developmental Disabilities. Official publication of Division 33, American Psychological Association.* http://www.apa.org/divisions/div33/docs%5Cd33news-current.pdf.

Tulsky, D., & Zhu, J. (2001). *Escala Wechsler de Inteligencia para adultos-III.* Mexico DF: El manual Moderno.

Tulsky, D., & Zhu, J. (2003). *Escala Wechsler de Inteligencia para adultos-III.* Mexico DF: El manual Moderno.

U.S. Department of Education, N. C. (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-34-1at http://nces.ed.gov/programs/coe/2009/section4/table-tot-1.asp

UNAM. (2014). *Para el 2050 mas de la cuarta parte de la población sera vieja en México.* Unidad León: Escuela Nacional de Estudios Superiores.

Weiss, L. (2003). Culture and children's intelligence: Cross-cultural analysis of the WISC-III . In J. Weiss, L. Van de Vijver, Saklofske, & D. (Ed.), *Georgas* (p. 50). United States: San Diego, CA: Academic Press.