# Adaption of the *WISC-IV* for Use in México: A Validity Study

Anthony D. Fina
The University of Iowa, USA
Pedro Sánchez-Escobedo
Universidad Autónoma de Yucatán, México
Liz Hollingworth
The University of Iowa, USA

**Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO, April 2010**

**Abstract**

This project seeks to provide evidence on the internal structure of the *Escala Wechsler de Inteligencia para Niños-IV* (*EWIN-IV*; Wechsler, 2007a) through a confirmatory factor analysis and intercorrelational study. Also provided is information on the adaption process and other sources of validity evidence in support of the *EWIN-IV* norms. The standardization data for the *EWIN-IV* were used for all analyses. The factor loadings and correlational patterns found on the *EWIN-IV* are comparable to those seen in the American versions of the test. The proposed factor and scoring structure of the *EWIN-IV* was supported.

*Keywords*: WISC-IV, EWIN-IV, validity, factor analysis, internal structure

# Adaption of the *WISC-IV* for Use in México

One of the most ineffective and dangerous practices in the measurement community is the adoption (as opposed to adaption) of instruments from one culture to another (Merenda, 2005); where little thought is given to adapting items, renorming, restandardizing the administration and scoring procedures, or ensuring the same structure of the construct being measured (Van de Vijver & Hambleton, 1996). It is the resulting test scores that serve as the basis for interpretations that are dangerous, because little attention is paid to the appropriateness of the instrument for the receiving culture. According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1999):

> When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested. (p. 99)

This single sentence encapsulates a great breadth of the adaption process; however, it states just a few of the many issues that needed to be considered in the Méxican adaption of the *WISC-IV*.

The Wechsler Intelligence Scale for Children - 4th Edition (*WISC-IV*; Wechsler, 2003a) is an individually administered clinical instrument designed to assess the cognitive ability of children aged 6 years through 16 years 11 months (Wechsler, 2003b). It is the most frequently used standardized test for assessing children's intelligence in the U.S. (Prifitera, Weiss, Saklofske, & Rolfhus, 2005). In 2005, The Psychological Corporation published the *WISC-IV Spanish* (Wechsler, 2005a) for use with populations of Spanish speaking American children acculturating to the United States. As part of the internationalization of the test, the *WISC-IV* (Wechsler, 2003a) was adapted for cultural fairness and piloted in México in 2005 with a sample

of participants to create norms for use with a Spanish-speaking Méxican population (Wechsler, 2007b).

This project seeks to provide additional validity evidence in support of the *Escala Wechsler de Inteligencia para Niños-IV* (*EWIN-IV*; Wechsler, 2007a), published in 2007 by Manual Moderno. (see Table 1 for a summary of the three versions of the *WISC-IV*.) This is needed because, without it, clinicians and practitioners in México are limited to using a test with insufficient evidence for validity decisions. Moreover, as an individual's score must be interpreted in light of a reference group's characteristics, the validity evidence collected in support of these norms supports the use of the *EWIN-IV* in México. In addition, we reflect upon issues put forth by the International Test Commission's *Guidelines for Test Adaptation* (2001) and the *Standards* (AERA et al., 1999). For example, in addition to test translation, the adaption process must consider other factors which can affect scores, including construct equivalence, test administration, item format, and the influence of speed on performance (Hambleton, 2005).

Table 1
Summary of the Three Versions of the *WISC-IV*

| Abbreviation | Country | Purpose |
|---|---|---|
| *WISC-IV* | USA | For use with the general population, ages 6 years to 16 years and 11 months. |
| *WISC-IV Spanish* | USA | For use with examinees aculturating to the United States. |
| *EWIN-IV* | México | For use in México with the Spanish speaking urban population. |

## Background

The rationale for test adaption is based on the belief that tests and their psychometric properties are influenced by culture, language, and the social conditions (Weiss, 2003). Prior to the *EWIN-IV*, the most recent adaption of a Wechsler test for use in México was the *Wechsler Adult Intelligence Scale-Third Edition* (Wechsler, 2001). Nevertheless, tradition itself is not

sufficient enough reason to adapt a test; one still must determine the degree to which the psychological processes considered are universal and the degree to which these processes are influenced by culture and captured by the adapted test (Georgas, 2003).

The 2003 revision of the *WISC* was necessitated in part to update the theoretical foundations on which the scales are based (Wechsler, 2003b). This revision represents one of the most significant revisions to date (Alfonso, Flanagan, & Radwan, 2005). The *WISC-IV* reflects an attempt to align it with modern theory, such as the Cattell-Horn-Carroll (CHC) theory (Keith, Fine, Taub, Reynolds, & Kranzler, 2006). See Alfonso et al. (2005) for a description of the CHC and a summary of how the CHC theory has impacted modern tests of cognitive ability. Therefore, any adaption of the *WISC-IV* needs to determine the appropriateness of the revised version for the receiving culture and the extent to which the measured processes are captured in the new version.

**Structure of the *WISC-IV***

The *WISC-IV* provides a measure of general intellectual functioning (FSIQ) and four index scores. The Verbal Comprehension Index (VCI) is composed of three core subtests (Similarities, Vocabulary, and Comprehension) and one supplemental subtest (Information). These subtests are designed to measure verbal abilities which utilize reasoning, comprehension, and conceptualization (Wechsler, 2003b). This index requires use of knowledge acquired from one's environment. The Perceptual Reasoning Index (PRI) is composed of three core subtests (Block Design, Picture Concepts, and Matrix Reasoning) and one supplemental subtest (Picture Completion). These subtests capture perceptual reasoning and organization and emphasize fluid reasoning abilities. The Working Memory Index (WMI) measures attention, concentration, and working memory. These tasks require examinees to temporally retain information in memory,

perform a task, and produce a result. The WMI consists of two core subtests (Digit Span and Letter-Number Sequence) and one supplemental subtest (Arithmetic). The Processing Speed Index (PSI) measures mental and graphomotor processing speed. It consists of two core subtests (Coding and Symbol Search) and one supplemental subtest (Cancellation). These subtests require examinees to quickly process visual information. (See the technical manual [Wechsler, 2003b] or WISC-IV *Clinical Use and Interpretation: Scientist-Practitioner Perspectives* [Prifitera, Saklofske, Weiss, & Rolfhus, 2004] for a comprehensive overview.)

Core subtests are required for composite scores. Supplemental subtests provide additional information about the cognitive functioning of an examinee. Substitution of a core subtest for a supplemental subtest is allowed if needed to derive an index score, but no more than one substitution on two separate indexes is allowed if the FSIQ is to be calculated.

**The Test Adaption Process**

In 2008, the *WISC-IV* (Wechsler, 2003a) was adapted for use with the Méxican population. The adapted version, *EWIN-IV* (Wechsler, 2007a), was the primary instrument used. A questionnaire was also given to test takers to gather relevant background and demographic information. The test adaption team consisted largely of native speakers with previous experience in large-scale adaptions; many had training in psychometrics (P. Sanchez, personal communication, February 15, 2010). The test administrators were graduate or undergraduate students in psychology or education with previous course work in measurement and evaluation. Test administrators received extensive training on the administration and scoring procedures of the *EWIN-IV* to reduce the presence of method bias.

Efforts were made to reduce the most common types of bias during test translation (see Van de Vijver & Hambleton, 1996). For example, directions were translated into Spanish,

reviewed, and made simpler; some details were made more explicit. Other modifications consisted of the use of appropriate Spanish idioms and expressions in the directions and administration guide. These changes were made to increase comprehensibility for both the administrator and examinee. Items were reviewed for clarity and intent.

The most common change to subtests was reordering the items according to item difficulty. This was necessary because differences in item difficulty may be due to cultural differences. For example, in some parts of México, test takers were less familiar with an item depicting a bathtub in the Picture Completion subtest. This item was moved towards the end of the *EWIN-IV*. Other items were changed altogether. For instance, an item on the Information subtest asks about *London* on the *WISC-IV*, but was adapted to *New York* on the *EWIN-IV*. These changes do not invalidate the test because they are eliciting the same process. In this way, items were reviewed for clarity and intent.

Adapted items may not be identical, but they should elicit the same processes to ensure construct equivalence. If different constructs are measured, then construct bias may occur. In addition, successful adaption of an intelligence test requires construct equivalence between cultures. Although intelligence may be viewed differently within specific cultures, it is the similarities they share that support the adaption of the intelligence test to the target country. These similarities include common aspects between cultures and shared educational backgrounds. Support for universal cognitive processes as measured by the *WISC* comes from cross-cultural studies (Georgas, Weiss, R. Van de Vijver, & Saklofske, 2003). Expert panels, outside consultants, and practitioners were used to evaluate the proposed content of the *EWIN-IV* to maintain content coverage and relevance, especially for the new subtests (Wechsler, 2003b).

Test administration procedures are also considered when a test is adapted. For instance, whereas the test is administered in one session in the US, it was commonly observed that Méxican children became tired and unmotivated when all routines were tested in a single session; so it was suggested the test could be administered in two sessions, with a break from 20 minutes to 23 hours in between (Wechsler, 2007b). Many psychologists concurred that this modification was necessary to avoid biases in testing (P. Sanchez, personal communication, February 15, 2010).

During observations of clinical trials, a common failure among test administrators was to not complete the discrepancy analyses (P. Sanchez, personal communication, February 15, 2010). When investigated, many responded that it was a hassle, time consuming, and under-used in the schools. Thus, in the Méxican version, the discrepancy page was printed next to the summary page to facilitate transferring scores from one to the other, and significance levels were printed in the forms with a pre-established value of 0.05. These modifications were designed to foster the calculation of discrepancies and increase the probability of a more thorough analysis and reporting of all scores.

One major change to the procedures of administration was the change in length of the discontinue rule (P. Sanchez, personal communication, February 15, 2010). For Similarities, Vocabulary, and Comprehension, the discontinue rule is three incorrect items rather than five. It was determined during the standardization that the extra two attempts did not make a difference for a majority of the examinees. These examples demonstrate that beyond mere translation, adaptation comprises various other dimensions of the testing process, such as directions, pace, conditions, and scoring.

The adaption process was intended to enhance the ability of the *WISC* subtests to fairly test Méxican children. In addition to changes in the item order, as well as the items themselves, this includes the adaption of procedures and directions as well. The overarching goal of the adaption was to develop items that would elicit the same response processes, measure the same construct, and produce scores that are equivalent to the *WISC-IV*. The next section describes the methodology used to create the *EWIN-IV*.

## Methods

### Subjects

Méxican children and adolescents were sampled to represent 10 states in 5 different regions of the country and México City (Wechsler, 2007b). It was not intended that the sample be reflective of students from rural areas or whose native language is not Spanish, nor was it to be a proportional sample reflective of a national census; rather, it was stratified to capture the regional and cultural differences present in México (P. Sanchez, personal communication, February 15, 2010). This was necessary because a disproportionate number of people, nearly a third, live in México City or its surrounding metropolitan area. The following a priori considerations were taken into account prior to the sampling: school type, sex, and age group (Wechsler, 2007b). Consistent with the norming sample for the *WISC-IV*, there were several exclusions, including: the presence of an obvious physical or intellectual disability that could interfere with performance on the test, the presence of acute physical illness at the time of the test, if the student recently moved from rural area, and if Spanish was not their primary language. The data were collected from May 15 to November 15, 2005.

Table 2

Breakdown of the *EWIN-IV* Norm Sample by Age

| Age | Male | Female | Total |
|---|---|---|---|
| 6:0-6:11 | 76 | 47 | 123 |
| 7:0-7:11 | 68 | 52 | 120 |
| 8:0-8:11 | 58 | 51 | 109 |
| 9:0-9:11 | 57 | 60 | 117 |
| 10:0-10:11 | 60 | 54 | 114 |
| 11:0-11:11 | 58 | 53 | 111 |
| 12:0-12:11 | 59 | 43 | 102 |
| 13:0-13:11 | 45 | 49 | 94 |
| 14:0-14:11 | 60 | 40 | 100 |
| 15:0-15:11 | 67 | 52 | 119 |
| 16:0-16:11 | 72 | 53 | 125 |
| Total | 680 | 554 | 1234 |

Source: Weschsler (2007b)

Table 3

Breakdown of the *EWIN-IV* Norm Sample by State

| Region | State | Males | Females | State Total | Region Total |
|---|---|---|---|---|---|
| North Central | | | | | 270 |
| | Aguascalientes | 49 | 41 | 90 | |
| | Coahuila | 48 | 42 | 90 | |
| | San Luis Potosí | 46 | 44 | 90 | |
| North West | | | | | 184 |
| | Sinaloa | 23 | 21 | 44 | |
| | Sonora | 72 | 68 | 140 | |
| Central | | | | | 273 |
| | District Federal | 116 | 135 | 251 | |
| | Morelos | 10 | 12 | 22 | |
| West | | | | | 253 |
| | Colima | 39 | 0 | 39 | |
| | Jalisco | 111 | 50 | 161 | |
| | Michoacán | 35 | 18 | 53 | |
| SouthWest | | | | | 254 |
| | Campeche | 34 | 27 | 61 | |
| | Yucatán | 97 | 96 | 193 | |
| Total | | 680 | 554 | 1234 | 1234 |

Source: Wechsler (2007b)

The final sample consisted of 1,234 participants stratified on the previously listed criteria (see Tables 2 and 3 for a breakdown of the sample), of which 30 participants' data were removed due to serious violations of routine or test administration (Wechsler, 2007b). Additional analyses showed 1,150 students completed all core subtests. The most frequently incomplete subtest was Coding with 76 incompletes. It was not clear if this was due to clerical errors made by test administrators or a test taker's failure to finish the subtest.

***EWIN-IV***

When the *WISC-IV* was adapted for the Méxican population, the adapted version was renamed the *EWIN-IV*. Evidence of internal consistency was obtained using the split-half method. The reliability coefficient for each subtest is based on correlation between the total scores of the two half-tests and corrected for length using the Spearman-Brown formula. The split-half method is not appropriate for estimating the reliability of a measure for processing speed. The data collection method did not allow for coefficients of reliability to be determined for Coding, Symbol Search, and Cancellation. As a result, future investigations will need to determine the stability of these measures.

Table 4 presents the reliability estimates for the *EWIN-IV* subtests and composite scores for each age group. The reliability coefficients were calculated using the formula recommended by Nunnally and Burnstein (1994). The average reliability coefficients were calculated using the Fisher's *z* transformation. As Table 4 indicates, the overall range of the reliability coefficients was from 0.89 for Picture Concepts to 0.95 for Letter-Number Sequencing. On the WISC-*IV Spanish*[1], the overall coefficients were 0.75 for Coding, 0.74 for Symbol Search, and 0.82 for Cancellation (Wechsler, 2005b). The reliability coefficient was 0.82 for the PSI and 0.97 for the FSIQ on the *WISC-IV Spanish*. These were lower than those found in the *WISC-IV*. By age-

[1] These were reported in the following age groups: 6-7, 8-9, 10-11, 12-13, and 14-16.

group, the only reliability coefficient for a subtest which is of concern is Comprehension (0.66). This coefficient was noticeably lower compared with the other age groups and to the other versions. Overall, the reliability coefficients on the *EWIN-IV* were similar to those found on the *WISC-IV* and higher than those found on the *WISC-IV Spanish*. The composite scores were expected to have higher coefficients than subtests because they measured a broader sample of abilities. For the subtests that were reported, the *EWIN-IV* is reliable tool.

Table 4
Reliability Estimates for the *EWIN-IV* Using the Split-Half Method

| Subtest / Composite | Age | | | | | | | | | | | Overall Average $r_{xx}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| Block Design | .92 | .90 | .84 | .89 | .85 | .86 | .91 | .87 | .89 | .93 | .90 | .92 |
| Similarities | .95 | .92 | .90 | .89 | .92 | .85 | .93 | .91 | .85 | .93 | .93 | .94 |
| Digit Span | .90 | .87 | .84 | .82 | .82 | .81 | .90 | .83 | .83 | .92 | .85 | .90 |
| Picture Concepts | .90 | .83 | .85 | .86 | .82 | .83 | .85 | .80 | .86 | .82 | .89 | .89 |
| Vocabulary | .92 | .91 | .87 | .88 | .88 | .87 | .91 | .91 | .90 | .91 | .90 | .93 |
| Let-Num Seq. | .97 | .96 | .94 | .91 | .94 | .92 | .95 | .90 | .88 | .92 | .89 | .95 |
| Matrix Reasoning | .93 | .92 | .93 | .90 | .92 | .90 | .92 | .90 | .86 | .92 | .92 | .94 |
| Comprehension | .90 | .88 | .87 | .66 | .87 | .80 | .89 | .87 | .85 | .88 | .92 | .90 |
| Picture Completion | .92 | .91 | .90 | .89 | .91 | .92 | .91 | .87 | .88 | .89 | .88 | .93 |
| Information | .93 | .93 | .91 | .90 | .89 | .84 | .93 | .88 | .85 | .90 | .88 | .93 |
| Arithmetic | .96 | .89 | .89 | .86 | .90 | .87 | .92 | .90 | .87 | .93 | .88 | .93 |
| VCI | .97 | .96 | .95 | .93 | .95 | .93 | .96 | .96 | .94 | .96 | .97 | .97 |
| PRI | .96 | .95 | .94 | .94 | .94 | .93 | .95 | .93 | .92 | .95 | .95 | .96 |
| WMI | .97 | .97 | .94 | .93 | .94 | .93 | .96 | .93 | .92 | .96 | .93 | .96 |

Note: The overall average reliability coefficient was found using Fisher's *Z* transformation. The reliabity for the composite was found using the method recommended by Nunnally and Bernstein (1994).

Very much related to reliability is the standard error of measurement (SEM). This provides an estimate of the amount of error in an individual's observed test score. The smaller the SEM, the greater one's confidence is in the precision of the observed test score. For the subtests, the smallest overall SEM was 0.82 found on Letter-Number Sequencing and the biggest was found on Picture Concepts with a value of 1.18. In general, the SEMs for the *EWIN-IV* tended to be smaller when compared to the *WISC-IV* Spanish, but larger when compared to the *WISC-IV*. The SEMs for all the subtests and composites are reasonable in comparison to the *WISC-IV* and *WISC-IV Spanish* (see Table 5).

Table 5
Standard Errors of Measurement for the *EWIN-IV's* Subtests and Composites

| Subtest / Composite | Age | | | | | | | | | | | Overall Average SEM[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| Block Design | .85 | .93 | 1.19 | 1.01 | 1.17 | 1.12 | .90 | 1.06 | 1.00 | .82 | .95 | 1.01 |
| Similarities | .70 | .82 | .94 | .98 | .85 | 1.16 | .82 | .89 | 1.18 | .82 | .77 | .91 |
| Digit Span | .93 | 1.07 | 1.22 | 1.26 | 1.27 | 1.29 | .95 | 1.24 | 1.23 | .87 | 1.15 | 1.14 |
| Picture Concepts | .95 | 1.22 | 1.16 | 1.14 | 1.26 | 1.22 | 1.16 | 1.36 | 1.14 | 1.28 | 1.02 | 1.18 |
| Vocabulary | .85 | .90 | 1.08 | 1.04 | 1.05 | 1.07 | .90 | .92 | .97 | .91 | .93 | .97 |
| Let-Num Seq. | .56 | .58 | .75 | .91 | .71 | .87 | .67 | .93 | 1.05 | .85 | .99 | .82 |
| Matrix Reasoning | .82 | .84 | .80 | .93 | .85 | .94 | .84 | .95 | 1.12 | .83 | .84 | .89 |
| Comprehension | .96 | 1.05 | 1.07 | 1.74 | 1.08 | 1.35 | .98 | 1.07 | 1.14 | 1.05 | .86 | 1.15 |
| Picture Completion | .82 | .92 | .97 | 1.00 | .92 | .84 | .92 | 1.06 | 1.02 | .98 | 1.02 | .95 |
| Information | .81 | .81 | .91 | .94 | 1.00 | 1.20 | .79 | 1.04 | 1.16 | .95 | 1.05 | .98 |
| Arithmetic | .61 | 1.00 | .99 | 1.12 | .97 | 1.06 | .87 | .95 | 1.09 | .79 | 1.02 | .96 |
| VCI | 2.57 | 2.90 | 3.37 | 4.04 | 3.19 | 3.97 | 2.92 | 3.11 | 3.66 | 2.93 | 2.66 | 3.25 |
| PRI | 2.90 | 3.43 | 3.78 | 3.69 | 3.82 | 4.10 | 3.46 | 4.05 | 4.22 | 3.15 | 3.22 | 3.64 |
| WMI | 2.74 | 2.74 | 3.60 | 3.96 | 3.63 | 3.95 | 2.93 | 4.01 | 4.12 | 3.05 | 3.92 | 3.55 |

Note: The reliability coefficents shown in Table 3 and the population standard deviations (i.e. 3 for the subtests) were used to compute the standard errors of measurement.

[a] The average SEMs were calculated by averageing the sum of the squared SEMs for each group and obtaining the square root of the result.

**Construct validity evidence for the *EWIN-IV*.** Validity is a unitary concept (AERA et al., 1999, p. 11). It is the degree to which evidence and theory supports the stated purposes of a test. Test validation then is for supporting test score interpretations and justifying test use. All possible sources of validity evidence are subsumed under construct validity. Construct validity comes from the integration of any possible sources of evidence that come to bear on the interpretation of test scores (Messick, 1989). Sources of validity evidence include the description of the adaption process, characteristics of the sample, and evidence of response processes, to name a few. While all sources of validity evidence are essential for the valid interpretation of a test score, the remainder of this article will focus on just one: internal structure.

Validity evidence based on the internal structure reveals "…the degree to which the relationships among test items and test components confirm to the construct on which the proposed test score interpretations are based" (AERA et al., 1999, p. 13). One source of evidence for this comes from intercorrelational studies. These studies examine the degree to which data

support a priori hypotheses about the pattern of relationships among parts of the test. Factor analytic studies provide evidence of internal structure as well. Confirmatory factor analyses (CFA) offer insight into the internal structure of assessment instruments and the possible latent abilities contributing to the observed responses. The multitrait-multimethod correlation matrix and the investigation into the factor structure of the *EWIN-IV* are described below.

***Multitrait-multimethod correlation matrices*.** Over 50 years ago Campbell and Fiske (1959) advanced a theoretical methodology for interpreting the correlations seen in a multitrait-multimethod matrix. Briefly, one would expect two measures of the same trait to be more highly correlated than two measures of different traits. These differences among correlations lend support of evidence for both convergent and discriminate validity.

There were several a priori hypotheses made about the multitrait-multimethod matrix for the current study. First, all subtests will show some degree of correlations with each other because all subtests are assumed to be, to some degree, measuring a general intelligence factor (i.e., *g*). Second, it was predicted that subtests will correlate most highly and frequently with the subtests within their own index, and these correlations will tend to be higher than correlations with subtests corresponding to other index scales. Third, based on previous studies, some subtests have higher *g* loadings than others. For example on the *WISC-IV,* Sattler and Dumont (2004) found that Block Design, Similarities, Vocabulary, Comprehension, Matrix Reasoning, Information, and Arithmetic had high *g* loadings. Based on this evidence, it was predicted that, regardless of scale membership, subtests with high *g* loadings will correlate highly with each other. Also, the subtests with high *g* loadings from the same index scale will tend to be more highly correlated than with subtests with high *g* loadings in other index scales. Fourth, previous research indicates a pattern of split loading with Picture Completion and the Verbal

Comprehension and Perceptual Reasoning scales (Wechsler, 2003b). So it was expected that Picture Completion would correlate highly with subtests on both scales.

***Confirmatory factor analysis*.** Past research supports that the *WISC-IV* measures for cognitive domains: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed (Sattler & Dumont, 2004; Wechsler, 2003b). It was expected that the *EWIN-IV* would produce a similar factor structure to the *WISC-IV*.

PROC CALIS, a SAS procedure, was used to conduct the CFA. The model fitting procedure was maximum likelihood, and the default settings were not altered. When estimating the CFA model parameters, several loadings were set equal to one (see Figure 1). Specifically, these constrained loadings included the loadings between the error variances to the observed variables, one loading from each first-order factor to one observed variable, and the loading from the second-order factor to one of the first-order factors. To maintain consistency with the factor loadings for the *WISC-IV*, standardized loadings for the *EWIN-IV* were reported.

For the current study, the Comparative Fit Index (CFI) was examined, which is little affected by sample size (Fan, Thompson, & Wang, 1999). The CFI indicates the percentage of covariance observed in the data that can be reproduced by a given model. Good fit is indicated by values greater than or equal to 0.95 (Hu & Bentler, 1999). The standardized root mean square residual (SRMSR) is the absolute value of the covariance residuals and was also examined in this study. Values less than 0.06 indicates good model fit (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA) represents the magnitude of discrepancy per degree of freedom (Jöreskog & Sörbom, 1993) and should fall below 0.06 (Hu & Bentler, 1999). Another useful index is the Non-Normed Fit Index (NNFI), or Tucker-Lewis Index. It is one of the indices less affected by sample size when the sample is large (Hu & Bentler, 1998). Values

greater than or equal to 0.95 indicate good model fit, especially if the rules of thumb for SRMSR and RMSEA are also met (Hu & Bentler, 1999). Therefore, the four above-mentioned indices were used to determine the adequacy of model fit.

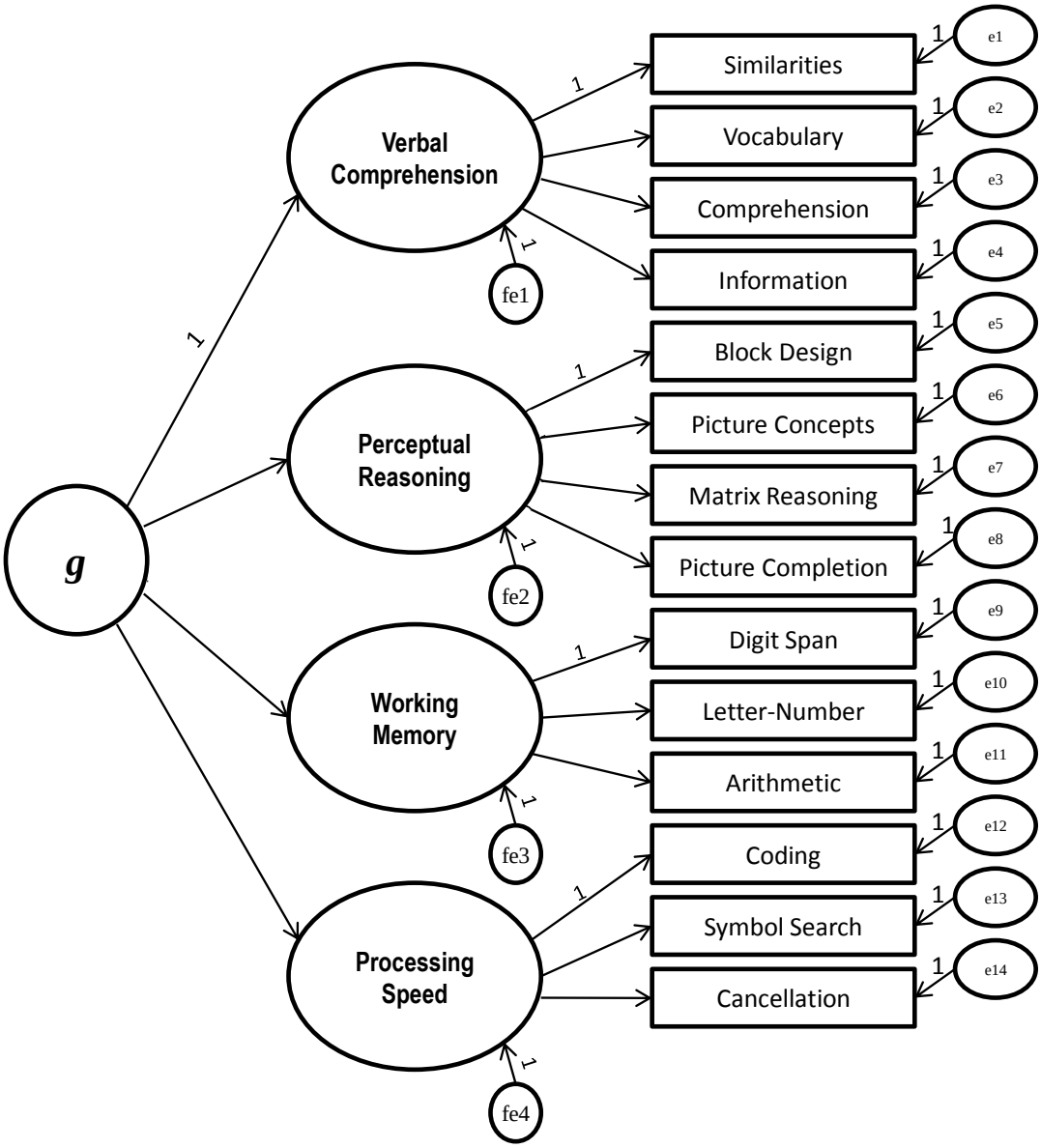


*Figure 1*. The factor and scoring structure used in the *WISC-IV*. Note: due to norming and reliability issues, the Word Reasoning subtest was removed from the *EWIN-IV* structure.

## Results

The results supporting the validity of the *EWIN-IV* are discussed below. First, the multitrait-multimethod correlation matrix is described in detail. Second, the results of the CFA are reported. In the discussion section, comparisons between the results for the *EWIN-IV* and results in previous research regarding the *WISC-IV* are made.

### Multitrait-Multimethod Correlation Matrix

The intercorrelations of the core and supplemental subtests and the sums of the scaled scores were calculated: Specifically, the (a) corrected correlation coefficients between the sum of scaled scores and scale scores and (b) uncorrected coefficients with the scale score included in the sum were determined (see Table 6). All the correlations between the subtests were significant. This was not surprising given the size of the sample. The lowest correlation for a subtest can be found between Coding and Picture Concepts ($r = 0.18$). Coding also had several low correlations with other subtests, including Similarities, Digit Span, Letter-Number Sequencing, and Picture Completion, $r = 0.21$ for all. The highest correlation between subtests was with Vocabulary and Comprehension ($r = 0.76$).

The Verbal Comprehension subtests also correlated highly with some Working Memory subtests, as both indices have auditory comprehension demands. Past research has suggested that Picture Concepts is related to verbal abilities (Wechsler, 2003b). This was supported by the correlation between Picture Concepts and Vocabulary and Information both exceeding 0.5.

As expected, there were moderate to large correlations between the Verbal Comprehension and Perceptual Reasoning subtests due to their mutual high *g* loadings. The moderate correlations of Perceptual Reasoning subtests with the Working Memory subtests suggest working memory's role in fluid reasoning tasks. Working Memory subtests correlated

highly with other Working Memory subtests (see Table 6). Correlations of 0.54 or greater were found between Digit Span, Letter-Number Sequencing, and Arithmetic. Working memory also correlates highly with several other subtests, namely Vocabulary ($r$ = 0.52-0.60), Information ($r$ = 0.52-0.66), and Matrix Reasoning ($r$ = 0.51-0.61). This is not unexpected considering the fluid reasoning, auditory comprehension, and cognitive flexibility demands of the subtests.

The Processing Speed subtests correlate most highly with each other. The correlations ranged from 0.34 to 0.47. Moderate correlations existed with other subtests. For example, the correlation between Symbol Search and Block Design (r = 0.43) might be related to the visual-perceptual and motor abilities required by both tasks. Based on correlations, Cancelation seemed to be the least *g*-loaded subtest (no correlation exceeded 0.49).

Table 6
The Intercorrelation Matrix Used for the Multitrait-Multimethod Matrix

| Subtest / Composite | BD | SI | DS | PCn | CD | VC | LN | MR | CO | SS | PCm | CA | IN | AR | VCI | PRI | WMI | PSI | FSIQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BD | | | | | | | | | | | | | | | | .63 | | | .69 |
| SI | .58 | | | | | | | | | | | | | | .74 | | | | .73 |
| DS | .48 | .46 | | | | | | | | | | | | | | | .53 | | .61 |
| PCn | .50 | .51 | .45 | | | | | | | | | | | | | .60 | | | .61 |
| CD | .24 | .21 | .21 | .18 | | | | | | | | | | | | | | .40 | .32 |
| VC | .60 | .72 | .52 | .51 | .25 | | | | | | | | | | .81 | | | | .78 |
| LN | .53 | .51 | .53 | .43 | .21 | .57 | | | | | | | | | | | .53 | | .64 |
| MR | .62 | .60 | .51 | .58 | .23 | .61 | .49 | | | | | | | | | .70 | | | .71 |
| CO | .51 | .67 | .48 | .46 | .22 | .76 | .54 | .52 | | | | | | | .90 | | | | .78 |
| SS | .43 | .37 | .37 | .35 | .40 | .42 | .36 | .39 | .39 | | | | | | | | | .40 | .60 |
| PCm | .61 | .55 | .42 | .55 | .21 | .57 | .50 | .59 | .52 | .40 | | | | | | | | | |
| CA | .45 | .41 | .35 | .38 | .34 | .43 | .38 | .39 | .45 | .47 | .43 | | | | | | | | |
| IN | .63 | .72 | .52 | .53 | .27 | .73 | .57 | .64 | .65 | .45 | .59 | .49 | | | | | | | |
| AR | .57 | .53 | .54 | .51 | .25 | .60 | .55 | .57 | .50 | .40 | .53 | .35 | .66 | | | | | | |
| VCI | .62 | .88 | .54 | .55 | .25 | .92 | .60 | .64 | .90 | .43 | .61 | .48 | .78 | .60 | | | | | |
| PRI | .84 | .67 | .57 | .83 | .25 | .68 | .57 | .87 | .59 | .46 | .69 | .48 | .71 | .65 | .71 | | | | |
| WMI | .58 | .55 | .84 | .50 | .24 | .63 | .90 | .56 | .59 | .42 | .53 | .42 | .63 | .62 | .65 | .65 | | | |
| PSI | .38 | .34 | .33 | .32 | .85 | .39 | .34 | .37 | .36 | .82 | .35 | .47 | .41 | .38 | .40 | .42 | .38 | | |
| FSIQ | .76 | .80 | .68 | .70 | .45 | .84 | .72 | .77 | .78 | .62 | .68 | .56 | .80 | .69 | .90 | .88 | .80 | .63 | |
| Mean | 9.7 | 9.0 | 9.0 | 9.9 | 9.8 | 9.3 | 9.5 | 8.8 | 9.5 | 9.3 | 9.4 | 9.9 | 9.6 | 10.1 | 27.7 | 28.4 | 18.4 | 19.1 | 93.8 |
| SD | 3.5 | 4.0 | 2.9 | 3.6 | 3.9 | 4.0 | 3.7 | 3.4 | 3.9 | 3.5 | 3.9 | 3.8 | 4.0 | 3.7 | 10.8 | 8.9 | 5.8 | 6.2 | 26.0 |

Note: The uncorrected correlation coefficients appear below the diagonal, the corrected coefficients for each subtest with its composite appears above the diagonal and to the right. BD=Block Design; SI=Similarities; DS=Digit Span; PCn=Picture Concepts; CD=Coding; LN= Letter-Number Sequencing; MR=Matrix Reasoning; CO=Comprehension; SS= Symbol Search; PCm=Picture Completion; CA=Cancelation; IN=Information; AR=Arithmetic; VCI=Verbal Comprehension Index; PRI=Perceptual Reasoning Index; WMI=Working Memory Index; PSI=Processing Speed Index; FSIQ=Full Scale Score.

**Confirmatory Factor Analysis**

The structure for the CFA for the current study was identical to the structure of the *WISC-IV*, as proposed by the test author (Wechsler, 2003b). Subtests load onto one of four broad cognitive abilities, which load onto a second-order general intelligence factor (i.e., *g*). The CFA used the correlation matrix based on 1,150 examinees who completed every subtest (see Table 7). As seen in the examination of the correlational matrix earlier, the intercorrelations were as expected. The patterns of correlations were also in line with what was observed in the *WISC-IV* and the *WISC-IV Spanish* (see Wechsler, 2003b, 2005b).

Table 7
Correlation Matrix Used in the CFA and the Squared Multiple Correlation Resulting from the CFA.

| Subtest | BD | SI | DS | PCn | CD | VC | LN | MR | CO | SS | PCm | CA | IN | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BD | | | | | | | | | | | | | | .61 |
| SI | .58 | | | | | | | | | | | | | .69 |
| DS | .47 | .46 | | | | | | | | | | | | .48 |
| PCn | .50 | .52 | .45 | | | | | | | | | | | .48 |
| CD | .24 | .21 | .21 | .18 | | | | | | | | | | .24 |
| VC | .59 | .72 | .52 | .51 | .24 | | | | | | | | | .77 |
| LN | .53 | .52 | .52 | .42 | .21 | .56 | | | | | | | | .53 |
| MR | .62 | .59 | .51 | .58 | .23 | .61 | .48 | | | | | | | .64 |
| CO | .50 | .67 | .47 | .45 | .22 | .75 | .53 | .51 | | | | | | .64 |
| SS | .40 | .36 | .36 | .36 | .40 | .41 | .36 | .39 | .39 | | | | | .48 |
| PCm | .59 | .55 | .41 | .54 | .21 | .57 | .49 | .58 | .52 | .39 | | | | .56 |
| CA | .43 | .41 | .34 | .38 | .34 | .43 | .39 | .39 | .46 | .45 | .42 | | | .48 |
| IN | .63 | .72 | .52 | .53 | .27 | .74 | .58 | .64 | .66 | .43 | .59 | .47 | | .75 |
| AR | .56 | .53 | .54 | .49 | .25 | .59 | .54 | .57 | .49 | .39 | .52 | .34 | .66 | .60 |

Note: Correlations are based on examinees who completed all subtests (N=1150). BD=Block Design; SI=Similarities; DS=Digit Span; PCn=Picture Concepts; CD=Coding; LN= Letter-Number Sequencing; MR=Matrix Reasoning; CO=Comprehension; SS= Symbol Search; PCm=Picture Completion; CA=Cancelation; IN=Information; AR=Arithmetic; VCI=Verbal Comprehension Index; PRI=Perceptual Reasoning Index; WMI=Working Memory Index; PSI=Processing Speed Index; FSIQ=Full Scale Score.

The proposed model for *EWIN-IV* fit the data well (SRMR = 0.029, RMSEA = 0.055, NNFI = 0.995). All the fit indices met the guidelines recommended by Hu and Bentler (1998, 1999). The standardized loadings from the first-order factors to the observed variables were relatively large, indicating that the factors explain a great deal of the observed variables' covariance structure. In fact, the range of loadings was 0.45 to 0.86 (see Figure 2). The error terms ranged from 0.48 on Vocabulary to 0.87 on Coding. The error term for the first-order

factors were all below 0.40, except for Processing Speed ( $\lambda$ = 0.65). The loadings for the second-order factor were generally high ( $\lambda$ = 0.93 or greater). Processing Speed had the lowest ( $\lambda$ = 0.77), which is as expected.

Finally, the squared multiple correlation ($R^2$) between the observed variables and the latent variables had a wide range, from 0.24 to 0.77 (see Table 7). This is the percentage of variance explained by the latent variable.

## Discussion and Conclusion

This investigation was primarily focused on determining the extent to which the observed variables (e.g., subtests) were related to the latent variables on the *EWIN-IV*, as specified by the scoring structure of the *WISC-IV*. Interestingly, the *WISC-IV* (see Keith et al, 2006) and the *EWIN-IV* (see Figure 2) had very similar loading patterns. Even with the Word Reasoning subtest removed from the *EWIN-IV*, there were similarities between the factor loading patterns of the *WISC-IV* and *EWIN-IV*.

The loadings for the *WISC-IV* are as follows: for Similarities, Vocabulary, Comprehension, Information, and Word Reasoning, loadings were 0.82, 0.89, 0.74, 0.82, and 0.74, respectively; for Block Design, Picture Concepts, Matrix Reasoning, and Picture Completion, loadings were 0.73, 0.61, 0.71, and 0.66, respectively; for Digit Span, Letter-Number Sequencing, and Arithmetic, loadings were 0.62, 0.66, and 0.80, respectively; for Coding, Symbol Search, and Cancellation, loadings were 0.68, 0.80, and 0.45, respectively; for the Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed second-order factors, loadings were 0.88, 0.91, 0.94, and 0.66, respectively (see Keith et al. [2006] for a description of the CFA).

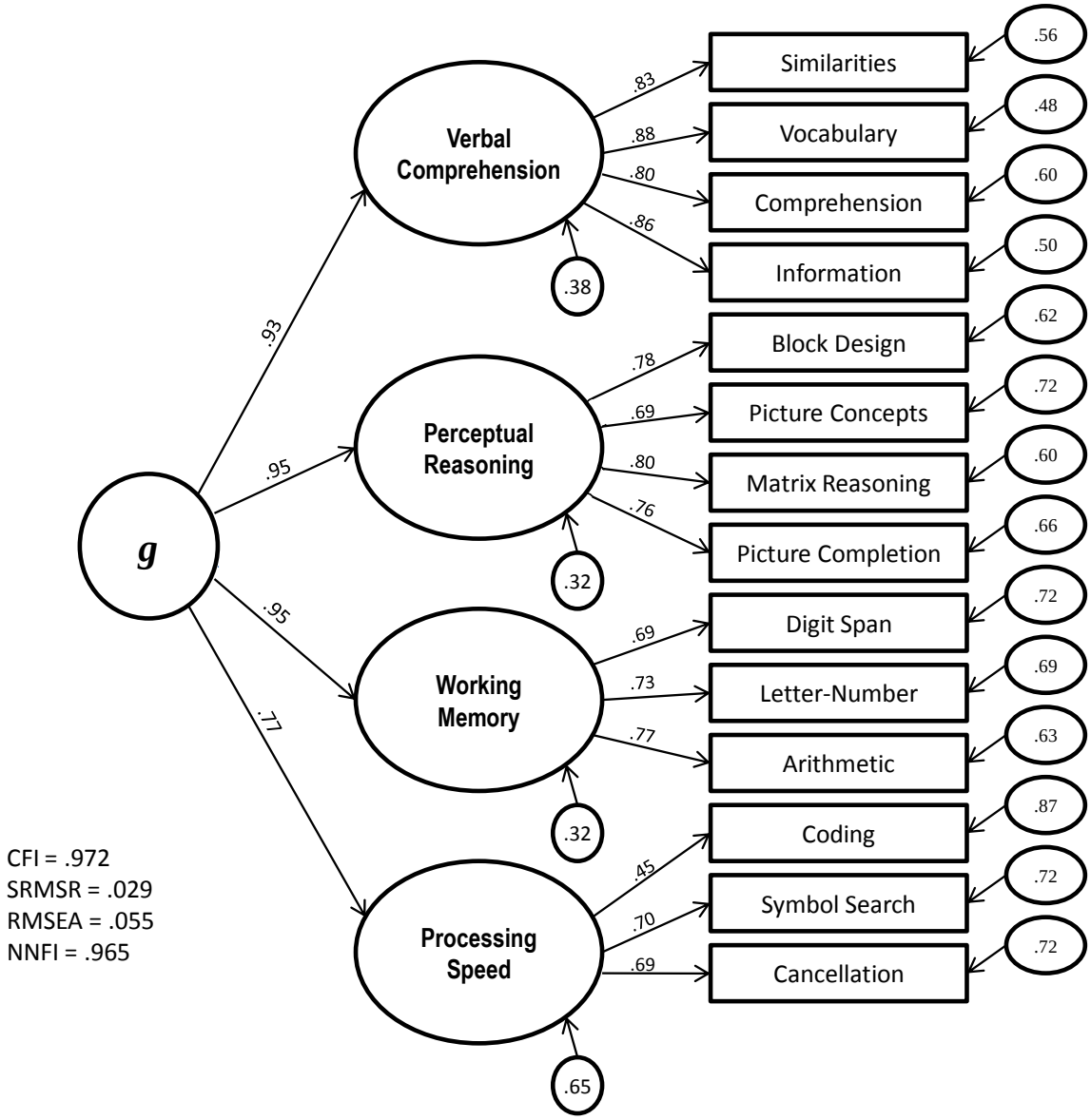


*Figure 2*. Factor loadings and fit statistics for the *EWIN-IV*.

For the Verbal Comprehension, Perceptual Reasoning, and Working Memory factors, the biggest loading difference between *WISC-IV* and *EWIN-IV* was found on Picture Completion (0.66 vs. 0.76, respectively), a supplemental subtest. The remaining subtests' loadings were all consistent, usually differing by 0.06 or less for any given subtest. However, the CFA did show one inconsistency. The pattern of loadings for the Processing Speed factor to the observed variables was different. Specifically, the *EWIN-IV* structure indicated that the factor loading for Coding was low ( $\lambda = 0.45$) while the factor loading for Cancellation was high ( $\lambda = 0.69$). This was the opposite trend seen on the *WISC-IV* (see Keith et al., 2006), with Coding and Cancellation loadings of 0.68 and 0.45, respectively. Although one cannot say the factor structures are identical, they do lend support for a successful adaption of the *WISC-IV* for use in México.

Furthermore, the patterns of correlations seen on the *EWIN-IV* were similar to the patterns seen in the *WISC-IV* and the *WISC-IV Spanish* (Wechsler, 2003b, 2005b). For instance, the subtests comprising the Verbal Comprehension scale correlated most highly with each other and with Picture Concepts, Letter-Number Sequencing, Matrix Reasoning, Picture Completion, and Arithmetic. The high correlations between the Verbal Comprehension subtests and Perceptual Reasoning subtests might suggest performance on the Perceptual Reasoning subtests are verbally mediated (Wechsler, 2003b).

Additional future analyses are also warranted. First, researchers should investigate if the factor structure of the *EWIN-IV* is indeed stable across age levels. This is needed to determine the extent to which the matrices are invariant across age-groups. The matrix used in this study is, in a sense, an unweighted correlation matrix because the matrix used was not weighted across age-groups. Future studies testing more complicated hypotheses could benefit from this; by

splitting the data into even/odd age groups, Keith et al. (2006) were able to develop a model and test it with a separate sample of test takers.

It would also be interesting to investigate the extent to which the structure of the implicit CHC theory found in the *WISC-IV* (Keith et a., 2006) is applicable to the *EWIN-IV*. Several studies have demonstrated that the CHC model provides a better structure than does the four-factor scoring procedure examined in this study (e.g., Chen, Keith, Chen & Chang, 2009; Keith el at., 2006).

In summary, this work was designed to validate the use of the *EWIN-IV* as a test of cognitive ability in Méxican children ages 6 years to 16 years and 11 months. The very fact that "*all* tests of intelligence and cognitive ability reflect culture" (Ortiz & Ochoa, 2005, p. 154), and because "culture dictates which responses are right and which are wrong on tests of intelligence and cognitive ability…" (Ortiz & Ochoa, 2005, p. 155), necessitates the adaption and renorming of the *WISC-IV* for the Méxican population. According to Ortiz and Ochoa (2005), "at the most fundamental level, standardized norm-referenced tests are designed to provide information that allows comparison of individual performance against the performance of a group of individuals with similar characteristics when all other relevant factors are controlled" (p. 158). This basic premise is at the heart of one of the most common practices plaguing the measurement community, and it must be confronted: the misinterpretations of scores resulting from the application of an instrument not appropriate for the receiving culture (Merenda, 2005).

This study offers a promising start to establishing a collection of validity evidence supporting the *EWIN-IV*, a successful adaptation of the *WISC-IV* for the Méxican population. This information and evidence of validity, which is required by the *Standards for Psychological Testing*, is needed by the measurement community and will assist in the valid interpretation of

*EWIN-IV* test scores for practitioners in México. By providing this information, these results serve to inform professionals across multiple disciplines such as educational counseling, psychometrics, and teaching. Furthermore, practitioners and cross-cultural researchers in the U.S. will have another tool at their disposal to measure intelligence in Méxican children.

References

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. p. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. (2nd ed., pp. 185-202). New York, NY: Guilford Press.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminate validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* pp. 81-105.

Chen, H., Keith, T., Chen, Y., & Chang, B. (2009). What does the WISC-IV measure? Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences 2009, 54*(3), pp. 85-108.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on Structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), pp. 56-83.

Georgas, J. (2003). Cross-cultural psychology, intelligence, and cognitive processes. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.*), Culture and children's intelligence* (pp. 23-27). San Diego, CA: Academic Press.

Georgas, J., Weiss, L. G., R. van de Vijver, F. J., & Saklofske, D. H. (Eds.). (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. San Diego, CA: Academic Press.

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.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to under parameterized model misspecification. *Psychological Methods, 3*, 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

International Test Commission. (2001). *International Test Commission guidelines for test adaptation*. London: Author.

Jöreskog, K. & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Hillsdale, NJ: Erlbaum.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fourth edition: What does it measure? *School Psychology Review, 35*(1), pp.108-127.

Merenda, P. F. (2005). Cross-cultural adaption of educational and psychological testing. In R. k. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 321-341). Mahway, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education and Macmillan.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory third edition.* New York, NY: McGraw-Hill.

Ortiz, S. O., & Ochoa, S. H. (2005). Conceptual measurement and methodological issues in cognitive assessment of culturally and linguistically diverse individuals. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students: A practical guide* (pp. 153-167). New York, NY: Guilford Publications.

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.

Sattler, J. M., & Dumont, R. (2004). *Assessment of children: WISC-IV and WPPSE-III supplement.* San Diego, CA: Jerome M. Sattler Publisher.

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-99.

Wechsler, D. (2001). *Escala Wechsler de Inteligencia para adultos – third edition*. México, DF: Manual Moderno.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children - fourth edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003b). *Wechsler Intelligence Scale for Children - fourth edition: Technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2005a). *Wechsler Intelligence Scale for Children - fourth edition Spanish*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2005b). *WISC-IV Spanish manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2007a). *Escala Wechsler de Inteligencia para Niños-IV*. México, DF: Manual Moderno.

Wechsler, D. (2007b). *Escala Wechsler de Inteligencia para Niños-IV manual técnico*. México, DF: Manual Moderno.

Weiss, F. (2003). United States. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Cultural and children's intelligence: Cross-cultural analysis of the WISC-III* (pp. 41-59). Burlington, MA: Academic Press.