Please use the following citation when referencing this work:

McGill, R. J., Ward, T. J., & Canivez, G. L. (2020). Use of translated and adapted versions of the WISC-V: Caveat emptor. *School Psychology International*. Advance online publication. doi: 10.1177/0143034320903790

Use of Translated and Adapted Versions of the WISC–V: Caveat Emptor

Ryan J. McGill

Thomas J. Ward

William & Mary

Gary L. Canivez

Eastern Illinois University

Author note

Correspondence concerning this article should be addressed to Ryan J. McGill, William & Mary School of Education, P.O. Box 8795, Williamsburg, VA 23188. E-Mail: rmcgill@wm.edu

Abstract

The Wechsler Intelligence Scale for Children (WISC) is the most widely used intelligence test in the world. Now in its fifth edition, the WISC–V has been translated and adapted for use in nearly a dozen countries. Despite its popularity, numerous concerns have been raised about some of the procedures used to develop and validate translated and adapted versions of the test around the world. The purpose of this article is to survey the most salient of those methodological and statistical limitations. In particular, empirical data are presented that call into question the equating procedures used to validate the WISC–V Spanish, suggesting cautious use of that instrument. It is believed that the issues raised in the present article will be instructive for school psychologists engaged in the clinical assessment of intelligence with the WISC–V Spanish and with other translated and adapted versions around the world.

Keywords: WISC-V, cross-national assessment, evidence-based assessment, IQ, test validation

Use of Translated and Adapted Versions of the WISC–V: Caveat Emptor

The Wechsler Intelligence Scale for Children (WISC) was originally developed by David Wechsler as a downward extension of the Wechsler-Bellevue Intelligence Scale (Boake, 2002). Since its publication, the WISC has been re-normed several times and has been subjected to numerous modifications and refinements. Now in its fifth edition, the WISC is regarded as one of the most popular and commercially successful ability measures of its era.

Surveys have consistently indicated that the WISC is the most commonly administered intelligence test to children in the United States. Benson and colleagues (2019) found that the WISC–V (Wechsler, 2014a) was the most frequently utilized intelligence test and the second most used assessment measure overall with an average of 3.49 uses per month among nationally surveyed school psychologists Furthermore, the WISC–V averaged more uses per month than the three nearest intelligence tests combined.

Although comparable estimates of the assessment practices of international school or educational psychologists are not presently available, Four country-specific versions of the WISC (Mexico, Japan, United States, and The Netherlands) were cited by authors surveying local assessment practices in nine countries in a recent special issue of the *International Journal of School and Educational Psychology* (Kranzler, 2016). Not surprisingly, numerous adaptations of the WISC have been developed by the test publisher in North America, Europe, Asia, and Australia over the course of the last 30 years. As a result, the WISC is presently considered to be the most widely used intelligence test in the world (van de Vijver et al., 2019).

Regardless, substantive questions about the construct validity and the procedures used to validate the U.S. and international adaptations of the WISC–V have been raised by independent researchers (e.g., Beaujean, 2016; Canivez, Watkins, & McGill, 2019) within the empirical

literature. Whereas many of these investigations have found similar issues to those encountered by scholars who have attempted to replicate the posited factor structure of the U.S. version of the WISC–V (e.g., Canivez, Dombrowski, & Watkins, 2018; Canivez, Watkins & Dombrowski, 2016, 2017; Dombrowski, Canivez, Watkins, & Beaujean, 2015), other studies have identified unique validation concerns for several of the international WISC–V versions that are presently being used by school and educational psychologists practicing internationally and in cross-cultural contexts in the United States.

Purpose

The goal of the present article is to outline significant limitations regarding the validation procedures used to develop and claim empirical support for translated and adapted versions of the WISC–V in several countries and cross-cultural contexts. As the U.S. WISC–V serves as a reference instrument for translated and adapted versions of the WISC–V worldwide, we review its background and development. In particular, concerns have been raised in the professional literature about the structural validity of the measure since its publication. As will be demonstrated, these debates have important implications for international versions of the test given the linkages between those instruments and the U.S. WISC–V in many test technical manuals. The issues raised in the present article are instructive for school and educational psychologists engaged in the clinical assessment of intelligence around the world as they determine when or whether to adopt a particular adapted version of the WISC–V or similar psychological test where similar validation procedures have been reported.

Background and Development of the WISC–V

The WISC–V includes 10 “primary subtests” which contribute to the measurement of five “primary” index scores (Verbal Comprehension, Fluid Reasoning, Visual-

Spatial, Working Memory, and Processing Speed) and a global FSIQ composite[1]. Additionally, there are six "secondary subtests" but they do not contribute to the measurement of the primary indexes or FSIQ composite[2]. One of the major goals of the WISC–V revision plan was to redesign the instrument so that it provided for better measurement of posited broad abilities from the Cattell-Horn-Carroll model (CHC; Schneider & McGrew, 2018) as well as other neurocognitive constructs (Wechsler, 2014b). To that end, two subtests (Picture Completion and Word Reasoning) from the WISC–IV were eliminated and three new subtests were added. Two of these (Figure Weights and Visual Puzzles) were adapted from the WAIS–IV. To better comport with CHC theory, the former Perceptual Reasoning Index was split into separate Fluid Reasoning and Visual-Spatial Indexes resulting in a total of five factor-based index scores to compliment the global FSIQ; however, it is suggested in the Technical Manual and accompanying interpretive resources (e.g., Flanagan & Alfonso, 2017) that the index scores should serve as the primary point of interpretation for the test.

It is worth noting that the decision to move from a four-factor to a five-factor structure was presaged in a special issue of the *Journal of Psychoeducational Assessment* devoted to Wechsler Theory and practice in 2013. Weiss, Keith, Zhu, & Chen (2013b) reexamined the WISC–IV normative data using confirmatory factor analysis (CFA) and furnished evidence to suggest that an alternative five-factor model (similar in many respects to what was later modeled in the WISC–V Technical Manual) could be used to interpret those data. Whereas both the four-

[1] In contrast to previous editions of the WISC where all 10 primary subtests contributed to the measurement of the FSIQ (i.e., general intelligence), for the WISC–V, the FSIQ is a linear combination of seven of the primary measures.

[2] Various combinations of the secondary subtests contribute to the measurement of "ancillary" and "complimentary" index scores. However, these scores are not derived from the CFAs reported in the Technical Manual.

and five-factor model fit statistics were virtually indistinguishable, they argued that the five-factor model could be used to guide the future evolution of the test. However, in a critique of that article, Canivez and Kush (2013) challenged those results. They argued that the various five-factor models that were explored provided no more than a miniscule improvement in model fit and that there was evidence of potential model misspecification in the form of a second-order path coefficient between Fluid Reasoning and *g* of 1.0, indicating that the two dimensions were mathematically redundant[3]. Additionally, they raised concern about the number of complex parameters (e.g., specification of an intermediary Quantitative Reasoning variable and several subtest cross-loadings) that were required to achieve optimal fit. Based on these observations, Canivez and Kush (2013) concluded "We strongly believe that the substantial theoretical, methodological, and practical limitations greatly limit any interpretations of the results, particularly those suggesting utility of the findings for practitioners" (p. 166). In a rejoinder, Weiss and colleagues (2013a) suggested that Canivez and Kush's commentary served to steer the conversation away from the primary discussion of whether the WISC–IV measured four or five first-order abilities. They concluded, "Applying the spectacles of history, we note that several independent research teams following different lines of inquiry are converging on a model of intelligence that includes at least five of the same main domains" (p. 241).

WISC–V Structural Validity

Whereas extensive psychometric information is presented in the WISC–V Technical Manual to support a five-factor measurement model, a number of methodological concerns were again raised in an independent review by Canivez and Watkins (2016). These include, (a) the

[3] A path loading of 1.0 is technically permissible in CFA. Estimates exceeding 1.0 are considered out of bounds estimates.

use of Weighted Least Squares (WLS) estimation without justification[4], (b) the reported degrees of freedom in some of the models did not appear to match the measurement model that was reportedly examined, (c) retention of a complex model that has Arithmetic loading on three different factors, and (d) path coefficients between *g* and Fluid Reasoning that again approach or equaled unity suggesting that the WISC–V is likely overfactored (Brown, 2015).

Later, in a bit of psychometric detective work, Beaujean (2016) discovered that the discrepancy between the degrees of freedom reported in the Technical Manual and the structural model that was presented graphically was likely the result of employing an undisclosed constraint on the model. To be fair, there is nothing wrong with the constraint that was employed (effects coding) per se; however, this lack of disclosure should concern practitioners in an era that has been marred by a replication crisis in scientific psychology (Simmons, Nelson, & Simonsohn, 2011).

Subsequent independent WISC–V structural research has provided inconsistent support for the posited five-factor measurement model. A series of exploratory factor analyses (EFA) using a variety of EFA methods (Canivez, Dombrowski, & Watkins, 2018; Canivez, Watkins, & Dombrowski, 2016; Dombrowski et al., 2019; Dombrowski, Watkins, & Canivez, 2018; Dombrowski, Watkins, Canivez, & Beaujean, 2015) and CFAs (Canivez et al., 2020; Canivez, Watkins, & Dombrowski, 2017) have found that a four-factor model consistent with previous Wechsler Theory best explains performance on the WISC–V in normative and clinical samples.

The results from independent research also suggest that an alternative bifactor measurement model may best explain the structuring of WISC–V variables and not the indirect

[4] Maximum Likelihood estimation is usually the default method used for intelligence test data, WLS is commonly used to examine ordinal level data.

hierarchical model preferred by the test publisher. Briefly, in a bifactor model, general intelligence and the group-specific factors have simultaneous direct influences on the measured variables (i.e., subtests). Conversely, in the indirect hierarchical model, general intelligence has an indirect effect on the measured variables (MVs) through the group factors which have direct effects on the MVs (Beaujean, 2015). For example, Canivez, Watkins, and Dombrowski (2017) subjected the WISC–V normative data to best practice CFA procedures and found that a four-factor bifactor model consistent with previous Wechsler theory best fit the data. Of note, the publisher preferred five-factor hierarchical model contained evidence of model misspecification in the form of a Heywood case for Fluid Reasoning and was not considered a tenable explanation for the data. In essence, each of these studies represents a potential replication failure for the WISC–V measurement model as posited by the test publisher (Carroll, 1995). Further, even when theoretically consistent first-order dimensions can be located, they do not appear to account for meaningful variance beyond general intelligence (with the exception of Processing Speed), which may degrade the potential clinical utility of the primary indices.

Reynolds and Keith (2017)[5] challenged these results with an alternative CFA model in which a five-factor model was found to best fit the WISC–V normative data. However, that model included a new parameter in the form of a correlation between Fluid Reasoning and Visual-Spatial (.77). It was argued that this parameter was justified based on the shared content between the tests representing those factors which implied the presence of a "Nonverbal" mediating factor. Subsequent CFA investigations (Fenollar-Cortés & Watkins, 2019; Watkins,

[5] Meyer and Reynolds (2018) also recently used multidimensional scaling (MDS) to examine the WISC–V structure and concluded that the results supported a five-factor structure. However, even a casual inspection of the resulting MDS maps reveals that many of the Fluid Reasoning and Visual Spatial measures are located virtually equidistant to each other and thus, it could be plausibly argued that the two dimensions are indistinguishable.

Dombrowski, & Canivez, 2018) of international versions of the WISC–V have compared the model fit afforded by the alternative structure posited by Reynolds and Keith (2017) in comparison to rival models. Findings from these studies indicate that the four- and five-factor models are not statistically nor meaningfully different. In such circumstances, methodologists have long recommended that the more parsimonious structure is preferred (Brown, 2015).

In sum, these results raise significant questions about the theoretical structure for the WISC–V. Since its publication, a number of rival measurement models have been posited in the assessment literature for the instrument, and questions remain about whether it is measuring a five-factor model, a four-factor model, an alternative five-factor model, or a model that has yet to be identified. These issues are not merely the simple musings of psychometric researchers as they have bearing on whether certain WISC–V scores can or should be confidently interpreted by practitioners in clinical practice. As will be discussed later, these results may also have bearing for interpreting many of the international versions of the WISC–V given the suggested linkages between those instruments and the validity information provided for the U.S. WISC–V in their respective technical manuals.

International Adaptations of the WISC–V

In terms of structure and content, international versions of the WISC–V largely mirror the U.S. version (Weschler, 2014b). Each of the major adaptations employs the same five-factor interpretive structure. However, there are some differences in the configuration of subtests and index scores. For example, the WISC–V Spain (Wechsler, 2015), French WISC–V (Wechsler, 2016b), and German WISC–V (Wechsler, 2017) do not contain the Picture Concepts subtest nor the complementary subtests and related index scores. Additionally, there are several countries

(i.e., Italy) that continue to rely on adapted versions of the WISC–IV as those measures have yet to be revised (Kush & Canivez, 2019).

Although measurement invariance for the WISC–V across countries has been purportedly established by van de Vijver et al. (2019), they failed to examine the hierarchical measurement model posited by the test publisher. Instead separate first- and second-order analyses were conducted and although support for the five-factor model was reported, it is important to note that oblique (correlated) factors models omit the influence of general intelligence and can inflate the importance of first-order variables (Gignac & Kretzschmar, 2017). Also absent in these invariance studies were examinations of rival bifactor structures which previous research suggests provide the best fit to WISC–V data (e.g., Canivez, Watkins & Dombrowski, 2017).

Contemporary International WISC–V Validation Concerns

In some cases, technical manuals for international versions of the WISC–V (e.g., Wechsler, 2015) report extensive psychometric information that appears to allow practitioners to independently evaluate the adequacy of the WISC–V relative to established test development and validation standards (i.e., American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; International Test Commission [ITC], 2013, 2016), while in other cases (e.g., Wechsler, 2016a), foundational reliability and validity information is missing and users are directed to consult the U.S. WISC–V Technical and Interpretive Manual.

Whereas it may be possible for users to indirectly glean some insight about what a test measures from the results of validity studies conducted on other versions of that test, this approach is problematic. According to van de Vijver (2016), every translated version of a test has "some mismatch between the source and target version" (p. 368). Many of the international

versions of the WISC–V require substantive modification of U.S. items and omission of whole subtests and index scores. Thus, it cannot be assumed that an adapted version of the test automatically measures the same constructs as the source test. As noted in the *ITC Guidelines for Translating and Adapting Tests*,

> The norms, validity evidence, and reliability evidence of a test in its source language version do not automatically apply to other possible adaptations of the test into different cultures and languages. Therefore, empirical validity and reliability evidence of any new versions developed *must* [emphasis added] also be presented (ITC, 2016, p. 22).

In particular, the guidelines emphasize the use of appropriate statistical techniques (i.e., EFA and/or CFA) as part of gold-standard test validation procedures. Additionally, the *Standards for Educational and Psychological Testing* (AERA et al., 2014), which represent the gold standard in guidance on testing in the United States and many other countries, outline a blueprint for strong test validation that includes multiple forms of reliability and validity evidence (e.g., evidence based on test content, response processes, internal structure, relations to other variables, and the consequences of testing). It should be noted that the joint testing standards (AERA et al., 2014) are often used as an organizing framework for the reporting of psychometric evidence in the technical manuals for major intelligence tests.

Based on these guidelines, we highlight several limitations concerning test validation, reliability, and validity that have been identified in the empirical literature regarding various translated and adapted versions of the WISC–V. These limitations are not intended to be exhaustive. Rather, we focus on the most salient validation concerns for practitioners administering and interpreting these measures in clinical practice.

Inadequate Normative Samples

The U.S. version of the WISC–V was normed on a nationally representative sample of 2,200 participants ages 6-16 years, one of the larger standardization samples among available commercial school-age ability measures. In contrast, the normative samples for some of the international versions of the WISC–V fail to meet, in some cases, even *de minimis* guidelines for normative sampling. For example, ITC (2016) guidelines suggest that an adapted version of a test should contain at least 300 normative participants. Alternatively, Alfonso and Flanagan (2008) describe an acceptable norming sample as containing a minimum of 1,000 participants with at least 100 participants in each 1-year age bracket throughout the age span of the test. While the normative samples for the Canadian (N = 880) and UK (N = 415) versions exceed recommended ITC guidelines, the Spanish (N = 220) version of the WISC–V does not. It should be noted that all three of the aforementioned versions fall well short of the more stringent guidelines proposed by Alfonso and Flanagan (2008).

In the WISC–V Spanish Manual (Wechsler, 2017b), it is reported that the equating sample was divided into five age groups (2-year intervals) and that no group consisted of more than 60 participants. In the case of the WISC–V UK (Wechsler, 2016a), the number of participants in each age bracket is not disclosed though the sample size alone (N = 415) indicates that there are likely far less than the required number of participants in each age bracket. The WISC–V Spain (N = 1,008), German WISC–V (N = 1,411), and French WISC–V (N = 1,049) versions were normed on much larger samples. Even so, with large subject to item ratios and overall sample size (i.e., > 1,000), error rates as high as 30% have been observed in factor analysis simulations (Osbourne & Costello, 2004). We are not presently aware of any

international version of the WISC–V that compares to the sample size employed by the source test.

For some instruments, concerns regarding dilution of sample size appear to be a relatively recent phenomenon. In a review of the WISC–V Canadian, Cormier, Kennedy, and Aquilina (2016) highlighted that the two previous versions of the test were normed on 1,100 participants successively and that the current normative sample is approximately 30% smaller than previous editions with no justification provided for the decrease in sampling power. Our review of test technical manuals suggests that a similar pattern has also been observed for the UK and Spanish versions of the test.

Another issue to consider is the representativeness of the normative sample for intended examinees. The WISC–V Spanish is principally designed to assess the intellectual abilities of bilingual Spanish-speaking students in the United States. It is reported in the Manual (Wechsler, 2017b) that "Effort was made to ensure that each [demographic] region was represented in the sample" (p. 87). However, inspection of the proportions of the Hispanic population in the U.S. represented by each of the four major regions (Table 4.2, p. 87) reveal that the South region was significantly oversampled (54.1%) and the West (15.5%) region was significantly under sampled. This is particularly concerning given the fact that two of the three states with the largest proportion of Spanish speaking students in the U.S. do not appear to be represented by the participating sites listed in the Appendices despite the likelihood that measure will be used by many bilingual school psychologists to examine dual language examinees within those jurisdictions.

Beyond the control of the test publisher, practitioners can also use the WISC-V in ways that raise questions about adequacy of the normative sample. For example, the WISC–V UK is

widely used in the Republic of Ireland even though the normative sample for that instrument does not contain any participants from that country. As noted by Furr and Bacharach (2014), scores on norm-referenced tests are of little value when there is doubt whether an examinee is a member of the reference population. Whereas we stipulate that such practices are likely borne out of limited access to instrumentation, there are no separate norms for Irish children nor equivalence studies examining performance between Irish and UK examinees in the WISC–V UK Technical Manual (Wechsler, 2016a). Until such information is furnished, WISC–V UK scores for Irish children should be interpreted with caution.

Limited Reliability Evidence

Concerns about the reliability of WISC–V scores have implications for validity as reliability is necessary but insufficient for establishing the validity of a test. There are different reliability estimates that assess distinctive aspects of measurement precision: internal consistency coefficients assess the consistency of responses across test items, test-retest (stability) coefficients examine the accuracy of measurement over time, and interscorer agreement coefficients evaluate the degree to which examiners score test items consistently. All are important in their own right.

Whereas multiple forms of reliability evidence are reported in the Technical Manual for the U.S. WISC–V (Wechsler, 2014b), less reliability evidence is reported for most international versions of the WISC–V. Although internal consistency estimates are commonly reported, short-term stability and interscorer agreement are rarely, if ever, assessed. Even the reporting of internal consistency estimates does not always comport with best practice guidelines. For example, in the WISC–V Spanish Manual (Wechsler, 2017b), split-half coefficients are reported for non-speeded subtests and no estimates are provided for any of the index or composite scores

or speeded subtests. The use of the split-half method to establish internal consistency reliability has been questioned by measurement scholars. According to Geisinger (2013), "there is simply no reason to use these procedures today" (p. 41). Unfortunately, those are the only reliability estimates reported for the WISC–V Spanish, and a common practice for international versions of the WISC–V.

In other circumstances, the procedures used to obtain reliability coefficients are not fully disclosed. In the WISC–V UK Technical Manual (Wechsler, 2016a), average internal consistency coefficients are reported but the type of coefficient and the method used to calculate those estimates are not reported. Instead, readers are directed to the U.S. Technical Manual for a more in-depth discussion of the procedures used to calculate reliability estimates. However, such an approach to scale validation is likely problematic as all reliability estimates are the property of the scores for a test for a specific group of examinees (Geisinger, 2013). Thus, it cannot be assumed that these estimates will generalize from one group to another. Put simply, important reliability information, in particular the stability of WISC–V scores in normative samples, is presently unknown for most of the international versions of the instrument. The WISC–V Spain is a notable exception as the disclosure in its Manual is comparable to the U. S version.

Inconsistent Procedures for Evaluating Internal Structure

Evaluation of a test's internal structure (i.e., structural validity) is a critical aspect of test validation as structural validity addresses a necessary, but not singularly sufficient condition for construct validity (Keith & Kranzler, 1999). Typically, this is accomplished through EFA/CFA analyses as these procedures provide the statistical rationale for the development of test scores (McGill & Dombrowski, 2017). Not surprisingly, test standards (e.g., ITC, 2016) require the

disclosure of the results of an investigation of a test's internal structure. Yet, this information was not fully reported or, in some cases, completely omitted from some international technical manuals. For example, in the WISC–V Spanish Manual (Wechsler, 2017b), no factor analytic results or subtest intercorrelations were provided without compelling justification for these omissions. The latter omission is particularly egregious because it prevents users from being able to independently examine important technical aspects of the test without access to proprietary information from the test publisher. Instead, users are again directed to the U.S. WISC–V Technical and Interpretive Manual, suggesting "This information also provides relevant evidence for evaluating the validity of the WISC–V Spanish" (Wechsler, 2017b, p. 91).

Curiously, these omissions represent a significant departure from the procedures employed to validate the previous version of the instrument where EFA/CFA results were fully disclosed in the WISC–IV Spanish Manual (Wechsler, 2005). Inexplicably, the opposite pattern has been observed for the WISC–V UK where structural validity evidence was absent from the WISC–IV UK (Wechsler, 2004); yet, included in the most recent revision (Wechsler, 2016a). However, only model fit statistics were disclosed and the structural models containing standardized path coefficients were not presented so it is not possible for users of the test to examine the local fit of the proposed interpretive model. As previously mentioned, when this information is disclosed, a problematic loading of 1.0 between *g* and Fluid Reasoning is frequently observed indicating that those dimensions are empirically redundant (i.e., there is little, if any, unique variance in the Fluid Reasoning construct that cannot be explained by psychometric *g*). It is important to note that while global fit indices may indicate that a preferred model provides adequate fit, inspection of local fit may reveal a problematic parameter indicating that a model may not be tenable.

Even when structural validity data are reported, an emerging body of independent research has raised serious questions about the statistical methods used to claim support for the posited WISC–V measurement model for various international versions. For example, Lecerf and Canivez (2018) examined the factor structure of the French WISC–V using best practices in EFA and CFA procedures and found that a bifactor expression of a four-factor model consistent with Wechsler Theory provided a better fit to the normative data when compared to the hierarchical five-factor model preferred by the test publisher. Furthermore, the variance explained by the group-specific factors that were identified ranged from 2% to 5% and was dwarfed by the variance explained by general intelligence (37%). Resulting indices of interpretive relevance (i.e., Rodriguez, Reise, & Haviland, 2016) indicated that only the FSIQ can be interpreted confidently among the index and composite scores. Consistent with previous U.S. WISC–V research, EFA and CFA results did not support separation of Visual-Spatial (VS) and Fluid Reasoning (FR). Similar results have been obtained for the Canadian (Watkins, Dombrowski, & Canivez, 2018), UK (Canivez, Watkins, & McGill, 2019), German (Bünger, Grieder, & Canivez, 2019), and Spain (Fenollar-Cortés & Watkins, 2019) editions of the WISC–V.

Watkins et al. (2018) found evidence for a five-factor bifactor model if a VS-FR covariance was added, similar to the alternative five-factor model suggested by Reynolds and Keith (2017) when reexamining the U.S. version, although it was not superior to the more parsimonious four group factor version. A recent German WISC–V measurement invariance analysis across gender (Pauls, Daseking, & Petermann, 2019) supported invariance of the hierarchical five-factor model posited by the test publisher, but no other models were considered. Even so, the variance accounted for, and the interpretive relevance of the group-specific factors

located in these studies, was again low with the exception of Processing Speed and conclusions indicated strong support for FSIQ interpretation.

Use of Score Equating to Establish Construct Validity

A final concern is the use of score equating to seemingly establish the construct validity of a translated test. Essentially, equating ensures that scores are exchangeable across all available forms of a test. As previously mentioned, no structural validity information is reported in the WISC–V Spanish Manual (Wechsler, 2017b). Instead, the publisher reports the results of a study *equating* scores from the WISC–V and the WISC–V Spanish as a primary means for establishing WISC–V Spanish validity. It is asserted that "Because the WISC–V Spanish has been subjected to equating procedures, the same evidence [WISC–V Technical Manual] supports WISC–V Spanish validity" (p. 91). However, this approach to test validation is conceptually problematic (van der Linden, 2013). Equating, in this instance, assumes that a bilingual Spanish-speaking student would obtain the same score on both the WISC–V and the WISC–V Spanish regardless of their language proficiency or the test administered. Or put another way, a bilingual assessor accepting the logic of score equating could potentially elect to administer one of those measures and automatically bypass assessing with the other as it would be assumed the examinee would receive the same scores on both tests, a scenario at odds with best practice bilingual assessment guidelines (e.g., Rhodes, Ochoa, & Ortiz, 2005). Given the reality that the vast majority of referred bilingual examinees are rarely equally proficient in L1 and L2 skills (Cummins, 1984), the equating assumptions for the WISC–V Spanish are likely specious.

Additionally, there are well-known requirements for score equating: (a) the two tests should measure the same constructs, (b) have equal reliability, and (c) it should not matter which version of the test the examinee takes (von Davier, 2013). Some of these assumptions appear to

have been violated in the equating of the WISC–V Spanish. To illustrate, the average internal consistency reliability estimates for eight of the 10 primary subtests from the respective WISC–V/WISC–V Spanish manuals were extracted and converted using Fisher's *r* to *z* transformation to determine whether the pairwise differences were significantly different. Results reported in Table 1 indicate that statistically significant differences between internal consistency estimates were observed for over half (62%) of the subtests. These results and the omission of important validity information for the WISC–V Spanish suggests that several of the fundamental assumptions for equating may not hold. Furthermore, the degraded means and standard deviations reported for the WISC–V Spanish equating sample suggest that the WISC–V Spanish is likely not assessing posited constructs as well as the source test. In conclusion, score equating should not be used to shortcut long-established construct validation procedures (Ryan & Brockmann, 2018).

Discussion

The present article highlights a number of problems related to the reporting of reliability and validity evidence in the technical manuals for translated and adapted versions of the WISC–V. In some circumstances, fundamental aspects of reliability and validity are not assessed in a meaningful way. In others, such as the WISC–V UK and Canadian WISC–V, structural validity evidence is presented, but users are not able to fully judge whether the measurement model for the instrument is tenable nor whether the scores that are potentially interpreted by users are sufficiently stable over time. The WISC–V Spanish is a notable exception as its Manual reports comparable psychometric information to the U. S. version. These observations buttress concerns raised by prominent assessment scholars regarding the methodological quality in cross-national testing and research practices (e.g., Byrne et al., 2016).

In spite of this evidentiary lacuna, a wealth of independent factor analytic studies of numerous international editions of the WISC–V (Bünger, Grieder, & Canivez, 2019; Canivez, Watkins, & McGill, 2019; Fenollar-Cortés & Watkins, 2019; Lecerf & Canivez, 2018; Watkins, Dombrowski, & Canivez, 2018) have emerged suggesting that many WISC–V versions may not locate posited constructs nor measure the ones that are located well enough to permit confidant clinical interpretation of those indices (with the exception of general intelligence and Processing Speed). These results are consistent with independent U.S. WISC–V research and the long-standing debate about its structural validity. Much discussion in the literature has focused on whether the WISC–V measures four or five factors (e.g., Canivez, Watkins, & Dombrowski, 2016, 2017; Reynolds & Keith, 2017). While this may seem like a trivial matter to some, it is critically important for practitioners who use the instrument in clinical practice as they are likely to interpret scores that are presented as sacrosanct (Lilienfeld, et al., 2013; Weidman, Steckler, & Tracy, 2017). Of additional concern, similar to the U.S. WISC–V, Bünger and colleagues (2019) found that the models tested for the WISC–V German contained fewer degrees of freedom than information about model parameters in the Technical Manual would suggest. The cause of these discrepancies remains unclear.

Best Practices in the Translation and Adaptation of Ability Measures

As previously mentioned, the *ITC Guidelines for Translating and Adapting Tests*, provide clear and specific guidance about best practices for translating and adapting ability measures for use in clinical practice and we encourage practitioners to consult them when appraising the quality of the documentation provided in test technical manuals (https://www.intestcom.org/page/16). These guidelines require users to be provided with: (a) a detailed explanation of the procedures used to translate test items, (b) evidence of multiple forms

of reliability, (c) evidence of validity (in particular, structural validity), and (d) a description of the normative group that permits users to ascertain whether an examinee is a member of that reference population. Put simply, test publishers should always abide by these and other relevant guidelines. When considering whether to adopt a new test, we encourage practitioners to specifically evaluate whether structural validity evidence is present given the relationship between those analyses and test scores. When such information is absent, practitioners are encouraged to consider partnering with independent researchers to establish the validity of their measures (i.e., International School Psychology Association, 2011). Moving forward, we encourage test publishers to adopt a consistent framework for reporting technical information in their manuals. Consistent with most U. S. test publishers, the elements previously outlined from the joint testing standards (AERA et al., 2014) provide a useful organizing framework for consideration.

Limitations

Whereas the present article raises substantive concerns about use of the WISC–V in some international settings, there have been substantial contributions of the test publisher over the course of the last 30 years in adapting the WISC for use in other countries and cross-cultural contexts and these contributions should not be overlooked. Adapting, norming, and translating a test requires a significant amount of time and resources. Without these investments, many practitioners around the world would likely lack access to quality instrumentation.

It is also important to acknowledge that all tests have flaws. In many respects the WISC–V is an exemplary instrument. Our goal is not to dissuade users from adopting translated and adapted versions of the WISC–V. We simply wish to highlight significant limitations that have been identified in the empirical literature so that practitioners are better informed users of these

tests. As Weiner (1988) cogently noted, ethical psychologists must "(a) know what their tests can do and (b) act accordingly" (p. 829).

Conclusion

The limitations discussed in the present article have important implications for school and educational psychologists who engage in the clinical assessment of intelligence using translated and adapted WISC–V versions. The *ITC Guidelines on Test Use* (ITC, 2013) state that competent test users will determine that a test's documentation provides sufficient information to enable accurate evaluation of the reliability and validity of the test for relevant populations. When such information is missing from a technical manual it may be difficult for practitioners to practice confidently within the scope of these guidelines. We contend that users of translated and adapted versions of the WISC–V should not be routed to U.S. materials to locate required information about their tests; especially when that information is not applicable for other versions of the measurement instrument. Further, international school psychologists should not automatically assume that an adapted/translated version of a test measures intended constructs or captures expected relationships between posited constructs across different cultural groups absent the necessary psychometric information to establish these hypotheses as fact (Byrne, 2016).

References

Alfonso, V. C., & Flanagan, D. P. (2008). Assessment of preschool children: A framework for evaluating the adequacy of the technical characteristics of norm-referenced measurements. In B. Mowder, F. Rubinson, & A. Yasik (Eds.), *Evidence based practice in infant and early childhood psychology* (pp. 129-166). New York: John Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement on Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence, 3,* 121-136. Doi: 10.3390/jintelligence3040121

Beaujean, A. A. (2016). Reproducing the Wechsler Intelligence Scale for Children-Fifth Edition: Factor model results. *Journal of Psychoeducational Assessment, 34,* 404-408. doi: 10.1177/0734282916642679

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan. G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology, 72,* 29-48. Doi: 10.1016/j.jsp.2018.12.004

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24,* 383-405. doi: 10.1076/jcen.24.3.383.981

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Bünger, A., Grieder, S., & Canivez, G. L. (2019, July). Confirmatory factor analysis of the German WISC-V primary and secondary subtests. In G. L. Canivez (Chair), *Validity investigations for international versions of the WISC-V: Informing evidence based assessment*. Symposium conducted at the meeting of the International School Psychology Association, Basel, CH.

Byrne, B. M. (2016). Adaptation of assessment scales in cross-national research: Issues, guidelines, and caveats. *International Perspectives in Psychology: Research, Practice, Consultation, 5,* 51-65. Doi: 10.1037/ipp0000042

Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC-V for four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools, 55,* 741-769. Doi: 10.1002/pits.22138

Canivez, G. L., & Kush, J. C. (2013). WAIS-IV and WISC-IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). *Journal of Psychoeducational Assessment, 31,* 157-169. Doi: 0.1177/0734282913478036

Canivez, G. L., McGill, R. J., Dombrowski, S. C., Watkins, M. W., Pritchard, A. E., & Jacobson, L. A. (2020). Construct validity of the WISC-V in clinical cases: Exploratory and confirmatory factor analyses of the 10 primary subtests. *Assessment, 27,* 274-296. doi: 10.1177/1073191118811609

Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children-Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson, *Intelligent testing with the WISC-V* (pp. 683-702). Hoboken, NJ: Wiley.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children–Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 28,* 975-986. doi.org/10.1037/pas0000238

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children-Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29,* 458-472. doi: 10.1037/pas0000358

Canivez, G. L., Watkins, M. W., & McGill, R. J. (2019). Construct validity of the Wechsler Intelligence Scale for Children-Fifth UK Edition: Exploratory and confirmatory factor analyses of the 16 primary and secondary subtests. *British Journal of Educational Psychology, 89,* 195-224. doi: 10.1111/bjep.12230

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30,* 429-452. Doi: 10.1207/s15327906mbr3003_6

Cormier, D. C., Kennedy, K. E., & Aquilina, A. M. (2016). Test review: Wechsler Intelligence Scale for Children, Fifth Edition: Canadian (WISC–VCDN) by D. Wechsler. *Canadian Journal of School Psychology, 31,* 322-334. Doi: 10.1177/0829573516648941

Cummins, J. (1984) *Bilingual education and special education: Issues in assessment and pedagogy* San Diego, CA: College Hill

Dombrowski, S. C., Beaujean, A. A., McGill, R. J., Benson, N. F., & Schneider, W. J. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An example featuring the WISC-V. *Structural Equation*

Modeling, 26, 847-869. doi: 10.1080/10705511.2019.1622421

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22,* 90-104. Doi: 10.1007/s40688-017-0125-2

Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children-Fifth Edition with the 16 primary and secondary subtests. *Intelligence, 53,* 194-201. Doi: 10.1016/j.intell.2015.10.009

Fenollar-Cortés, J., & Watkins, M. W. (2019). Construct validity of the Spanish version of the Wechsler Intelligence Scale for Children Fifth Edition (WISC-V^{Spain}). *International Journal of School and Educational Psychology, 7,* 150-164. doi: 10.1080/21683603.2017.1414006

Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-V assessment*. Hoboken, NJ: John Wiley.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Geisinger, K. F. (2013). Reliability. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology* (Vol. 1, pp. 21-42). Washington, DC: American Psychological Association.

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence, 62,* 138-147. Doi: 10.1016/j.intell.2017.04.001

International School Psychology Association (2011). *Code of ethics*. Amsterdam, The Netherlands: Author.

International Test Commission. (2013). *International guidelines for test use* (Version 1.2). Retrieved from https://www.intestcom.org.

International Test Commission. (2016). *The ITC guidelines for translating and adapting tests* (2nd ed.). Retrieved from https://www.intestcom.org.

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the cognitive assessment system does and does not measure. *School Psychology Review, 28,* 303-321.

Kranzler, J. H. (2016). Current practices and future directions for the assessment of child and adolescent intelligence in schools around the world. *International Journal of School and Educational Psychology, 4,* 213-214. doi: 10.1080/21683603.2016.1166762

Kush, J. C., & Canivez, G. L. (2019). Construct validity of the WISC-IV Italian Edition: A bifactor examination of the standardization sample: Chi niente sa, di niente dubita. *International Journal of School and Educational Psychology*. Advance online publication. doi: 10.1080/21683603.2018.1485601

Lecerf, T., & Canivez, G. L. (2018). Complementary exploratory and confirmatory factor analyses of the French WISC-V: Analyses based on the standardization sample. *Psychological Assessment, 30,* 793-808. doi: 10.1037/pas0000526

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review, 33,* 883-900. doi:

10.1016/j.cpr.2012.09.008

McGill, R. J., & Dombrowski, S. C. (2017). School psychologists as consumers of research: What school psychologists need to know about factor analysis. *Communiqué, 46 (1),* 16-18.

Meyer, E. M., & Reynolds, M. R. (2018). Scores in space: Multidimensional scaling of the WISC-V. *Journal of Psychoeducational Assessment, 36,* 562-575. doi: 10.1177/0734282917696935

Osbourne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research, and Evaluation, 9 (11),* 1-9.

Pauls, F., Daseking, M., & Petermann, F. (2019). Measurement invariance across gender on the second-order five-factor model of the German Wechsler Intelligence Scale for Children-Fifth Edition. *Assessment*. Advance online publication. doi: 10.1177/1073191119847762

Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fifth edition: What does it measure? *Intelligence, 62,* 31-47. doi: 10.1016/j.intell.2017.02.005

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practical guide*. New York: Guilford.

Rodriguez, A., Reise, S. P., & Haviland, M . G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21,* 137-150. doi: 10.1037/met0000045

Ryan, J., & Brockmann, F. (2018). *A practitioner's introduction to equating: With primers*

on classical test theory and item response theory (Rev. ed.). Washington, DC: Council of Chief State School Officers.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73-163). New York: Guilford.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366. doi: 10.1177/0956797611417632

van de Vijver, F. J. R. (2016). Test adaptations. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 365-376). New York: Oxford University Press.

van de Vijver, F. J. R., Weiss, L. G., Saklofske, D. H., Batty, A., & Prifitera, A. (2019). A cross-cultural analysis of the WISC-V. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V. Clinical use and interpretation* (pp. 223-244). San Diego, CA: Academic Press.

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement, 50,* 249-285. doi: 10.1111/jedm.12014

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78,* 605-623. doi: 10.1007/S11336-013-9319-3

Watkins, M. W., Dombrowski, S. C., & Canivez, G. L. (2018). Reliability and factorial validity of the Canadian Wechsler Intelligence Scale for Children-Fifth Edition. *International Journal of School and Educational Psychology, 6,* 252-265. doi: 10.1080/21683603.2017.1342580

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children: Administration and scoring manual* (4th ed., UK). London, UK: Harcourt Assessment.

Wechsler, D. (2005). *Wechsler Intelligence Scale for Children: Manual*. (4th ed., Spanish). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children* (5th ed.). Bloomington, MN: NCS Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children: Technical and interpretive manual* (5th ed.). Bloomington, MN: NCS Pearson.

Wechsler, D. (2015). *Escala de inteligencia de Wechsler para niños-V: Manual técnico y de interpretación*. Madrid, Spain: Pearson Educacion.

Wechsler, D. (2016a). *Wechsler Intelligence Scale for Children: Administration and scoring manual*. (5th ed., UK). London, UK: Pearson Assessment.

Wechsler, D. (2016b). *WISC-V. Echelle d'intelligence de Wechsler pour enfants-5e édition*. Paris, France: Pearson France-ECPA.

Wechsler, D. (2017). *Wechsler Intelligence Scale for Children: Manual* (5th ed., Spanish). Bloomington, MN: NCS Pearson.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17,* 267-295. doi: 10.1037/emo0000226

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53,* 827-831. doi: 10.1207/s15327752jpa5304_18

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013a). Technical and practical issues in the structure and clinical invariance of the Wechsler Scales: A rejoinder to commentaries.

Journal of Psychoeducational Assessment, 31, 235-243. doi: 10.1177/0734282913478050

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013b). WISC-IV and clinical validation of the four- and five-factor interpretive approaches. *Journal of Psychoeducational Assessment, 31,* 114-131. doi: 10.1177/0734282913478032

Table 1

Internal Consistency Estimates for WISC–V (N = 2,200) and WISC–V Spanish (N = 220) Primary Subtests

	Internal Consistency Estimate	
Test	WISC–V	WISC–V Spanish
Similarities	.87	.92*
Vocabulary	.87	.94*
Block Design	.84	.86
Visual Puzzles	.89	.92*
Matrix Reasoning	.87	.89
Figure Weights	.94	.96*
Digit Span	.91	.90
Picture Span	.85	.91*
Coding	.81	[a]
Symbol Search	.82	[a]

Note: p values rounded to the nearest hundredth. [a] Estimates unavailable. * Difference between estimates is statistically significant ($p < .05$).

Author Bios

Ryan J. McGill, Ph.D., BCBA-D, NCSP is assistant professor and director of the school psychology program at the William & Mary School of Education. His research focuses on applied psychological measurement and SLD identification in school psychology.

Thomas J. Ward, Ph.D. is professor of education in the department of educational policy, planning, and leadership at the William & Mary School of Education. His research focuses on research methodology and applied measurement in education.

Gary L. Canivez, Ph.D. is professor of psychology at Eastern Illinois University. He is an elected member of the Society for the Study of School Psychology and a fellow of Division 5 of the American Psychological Association. His research focuses on applied psychological assessment and measurement. In particular, the clinical assessment of intelligence.