SCHOOL PSYCHOLOGY REVIEW
2022, VOL. 51, NO. 5, 619–633
https://doi.org/10.1080/2372966X.2020.1816804

RESEARCH INTO PRACTICE

Assessing the Psychometric Utility of IQ Scores: A Tutorial Using the Wechsler Intelligence Scale for Children–Fifth Edition

Marley W. Watkins[a] and Gary L. Canivez[b]

[a]Baylor University; [b]Eastern Illinois University

ABSTRACT

IQ tests provide numerous scores, but valid interpretation of those scores is dependent on how precisely each score reflects its intended construct and whether it provides unique information independent of other constructs. Thus, IQ scores must be evaluated for their reliability and dimensionality to determine their psychometric utility. As a tutorial, the Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V) scores were evaluated and it was demonstrated that the WISC-V is multidimensional, but only the Full-Scale IQ (FSIQ) was found to be sufficiently reliable for clinical use. WISC-V group *factors* were not well defined and WISC-V index (i.e., factor) *scores* were contaminated with variance from other constructs and insufficiently reliable for clinical decisions. Clinicians were encouraged to go beyond structural goodness of fit and evaluate IQ test scores in terms of their reliability and ability to provide information that is not available from the general ability score as well their predictive and treatment validity. Software was provided to assist in that evaluation.

IMPACT STATEMENT

IQ tests provide numerous scores, but valid interpretation of those scores is dependent on how precisely each score reflects its intended construct and whether it provides unique information independent of other constructs. Thus, IQ scores must be evaluated for their reliability and dimensionality to determine their psychometric utility. This article describes an evidence-based approach that clinicians can employ to assess the psychometric utility of IQ scores, supplies software tools to assist in that analysis, and provides a tutorial example using Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V) scores.

ARTICLE HISTORY
Received June 10, 2020
Accepted August 20, 2020

KEYWORDS
intelligence, WISC-V, children, assessment, evidence-based

ASSOCIATE EDITOR
Shane Jimerson

The Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V; Wechsler, 2014a) is one of the most frequently used tests by school psychologists (Benson et al., 2019; Groth-Marnat & Wright, 2016; L. T. Miller et al., 2020). However, the WISC-V can produce 35 separate scores (Carlson et al., 2016; Wechsler, 2014b). Even with its abbreviated primary battery, the WISC-V generates one global score, five primary index scores, and 10 subtest scores. Psychologists must decide which, if any, of these 16 scores to interpret.

Validity

Interpretation of WISC-V scores can only be justified with validity evidence (American Educational Research Association [AERA] et al., 2014). Several types of evidence must be integrated when evaluating the validity of test scores, most often, evidence about test content (content validity), internal structure (structural or factorial validity), and relationships to other variables (external validity). Following this template, Wechsler (2014b) provided content, structural, and external validity evidence in the WISC-V technical manual.

Evidence regarding the internal structure of a test is vital because that internal structure serves as the statistical rationale for the test's scoring structure (Braden, 2013; Braden & Niebling, 2012; Canivez & Youngstrom, 2019; Furr, 2011). Evidence of structural validity is often obtained through factor analysis, which is a multivariate statistical technique that utilizes the variability among a set of scores to identify the underlying latent constructs or factors that theoretically caused that observed variability (Montgomery et al., 2018).

Factor analysis was first developed by Spearman (1904) to analyze mental test scores in support of his theory of intelligence and has developed over the ensuing decades

CONTACT Marley Watkins Marley_Watkins@baylor.edu Department of Educational Psychology, Baylor University, One Bear Place #97301, Waco, TX 76798-7301, USA.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

 DOI: 10.1080/2372966X.2020.181680

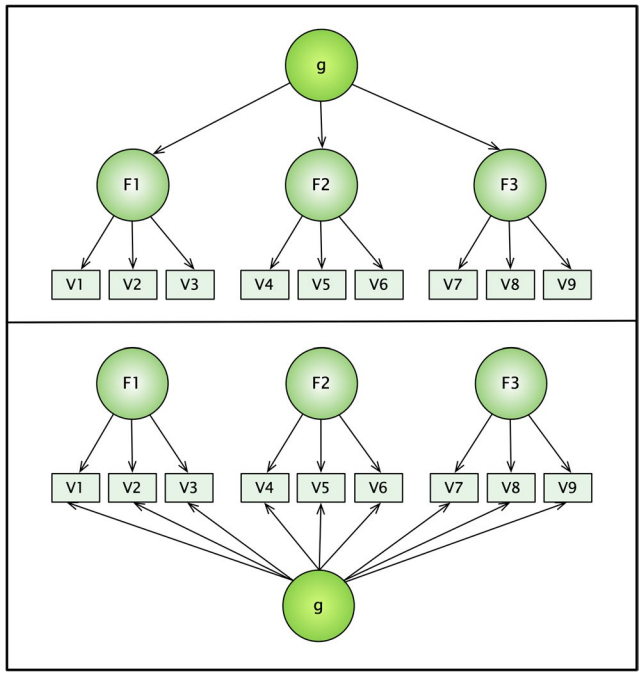
into a family of multivariate methods, roughly demarcated into exploratory and confirmatory approaches. Detailed information regarding the structure of latent constructs and factor analytic models is readily available (Bornovalova et al., 2020; Brown, 2015; Brunner et al., 2012; Chen et al., 2012; Chen & Zhang, 2018; Gorsuch, 1983; Gustafsson & Åberg-Bengtsson, 2010; Markon, 2019; Reise, 2012; Reise et al., 2013, 2018), Sellbom & Tellegen, 2019; Styck, 2019; Watkins, 2018).

In essence, factor analysis is used to verify that the internal structure of a scale (i.e., its dimensionality) "is consistent with expectations regarding the construct(s) that the scale is intended to measure" (Flora & Flake, 2017, p. 82). That is, the scale's actual structure should match its theoretical structure (Furr, 2011). Factor analysis has now become "central to the validation of measurement constructs" (Jewsbury & Bowden, 2017, p. 44) and "dictates the number of meaningful scores that a scale produces" (Furr, 2011, p. 7). For example, Wechsler (2014b) submitted WISC-V subtest scores to a confirmatory factor analysis and concluded that five factors were responsible for the variability of its subtest scores. Subsequently, the WISC-V scoring structure of five primary index scores was based on those five factors.

Questions About Validity and Multidimensionality

IQ tests have almost always been found to be hierarchically structured with multiple factors or dimensions (Beaujean & Benson, 2019). This multidimensional structure is reflected in modern theories of intelligence (Carroll, 1993) that are typically illustrated with path models as in Figure 1 (Brunner et al., 2012; Canivez & Youngstrom, 2019). In path diagrams, ovals represent factors and rectangles represent measured variables. Directional relationships between variables are indicated by single-headed arrows and nondirectional (correlational) relationships by double-headed arrows. A simplified higher-order factor structure is illustrated in the top panel of Figure 1. In this model, intelligence is assumed to consist of an overarching general factor (*g*) that influences the group factors (F1–F3) which, in turn, influence the measured variables or subtests (V1–V9), but there are no direct relations between the general factor and the subtests. This creates group factor scores that are impure measures of their purported factors because they are influenced by both general *and* group factors and therefore "represent a collection of different attributes" (Beaujean & Benson, 2019, p. 129) and psychologists "will not know which attribute to invoke to account for a particular score" (Gustafsson & Åberg-Bengtsson, 2010, p. 97). In sum, group factor IQ scores (i.e., WISC-V index scores) are conceptually complex and lack a univocal interpretation (Chen & Zhang, 2018; Ferrando & Lorenzo-Seva, 2019b).

Figure 1. Simplified Model of Intelligence Expressed in a Higher-Order Measurement Model (top panel) and Transformed With the Schmid-Leiman Procedure (bottom panel)



Note. Latent variables (factors) are represented by circles, measured variables by rectangles, and the direction of causal influence by directional arrows.

Multidimensionality and uncertainty regarding score interpretation are not unique to IQ tests. For example, measures of personality and psychopathology often contain both general and group factors (Gomez et al., 2019; Reise et al., 2018; Rodriguez et al., 2016a, 2016b). Similar issues have been encountered with educational tests (Wainer & Feinberg, 2015) where a meaningful score was defined as "one that is reliable enough for its prospective use and one that has information that is not adequately contained in the total test score" (Wainer & Feinberg, 2015, p. 18).

As described by Furr (2011), "Each score obtained from a scale should reflect a single coherent psychological variable" (p. 26). Adhering to this admonition, test publishers often either implicitly or explicitly assume that each test score can be interpreted as a measure of a single construct that provides meaningful and reliable information independent of other constructs (Beaujean & Benson, 2019; Canivez & Youngstrom, 2019; Reise et al., 2013). For example, Wechsler (2014b) claimed that the WISC-V index scores are "reliable and valid measures of the primary cognitive constructs they intend to represent" (p. 149) so that the Verbal Comprehension Index (VCI), for example, "measures the child's ability to access and apply acquired

word knowledge," which involves "verbal concept formation, reasoning, and expression" (p. 157). Although encouraging clinicians to consider index scores within an ecological context, Wechsler (2014b) suggested an interpretation system based on a comparison of index scores with each other and to the Full-Scale IQ (FSIQ) to identify strengths and weaknesses across the Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed cognitive domains.

Other successive-level approaches for the interpretation of IQ test scores have been developed for clinicians (e.g., Flanagan & Alfonso, 2017; Groth-Marnat & Wright, 2016; Kaufman et al., 2016; Sattler et al., 2016). Typically, these approaches assume that IQ scores are homogeneous, but they also "go beyond the information contained in the FSIQ or the index scores" (Sattler et al., 2016, p. 175) by using intraindividual or within-person comparisons of scores (an approach described as ipsative by McDermott et al., 1992, and idiographic by Freeman & Chen, 2019) to identify score patterns or profiles assumed to reflect cognitive strengths and weaknesses that, in turn, underpin recommendations for remedial strategies, classroom modifications, instructional accommodations, curricular modifications, targeted interventions, and program placements (Groth-Marnat & Wright, 2016; Kaufman et al., 2016; J. L. Miller et al., 2016; Sattler et al., 2016; Wechsler, 2014b).

These approaches to IQ score interpretation have achieved widespread use by school psychologists (Benson et al., 2020; J. L. Miller et al., 2016; Sotelo-Dynega & Dixon, 2014) and trainers (Lockwood & Farmer, 2020; L. T. Miller et al., 2020). For example, a recent survey found that they were routinely employed by a majority of school psychologists (Kranzler et al., 2020). In contrast, researchers have consistently criticized these approaches for inadequate reliability, validity, and diagnostic utility (Beaujean & Benson, 2019; Freeman & Chen, 2019; Glutting et al., 1997; Kranzler et al., 2016, 2020; McDermott et al., 1992; McGill, 2016, 2018; McGill et al., 2018; Styck et al., 2019; Watkins, 2003, 2009). As a consequence, IQ score interpretation is currently "in a state of disarray" (Beaujean & Benson, 2019, p. 126).

Purpose

Given this disarray, psychologists must possess considerable expertise in psychometrics to competently interpret scores from IQ tests (AERA et al., 2014; Beaujean & Benson, 2019; Gould et al., 2013; Reynolds & Milam, 2012). Unfortunately, instruction in psychometrics has been neglected in graduate training (Aiken et al., 2008; Canivez, 2019; Charter, 2003; Perham, 2010). For example, Aiken et al. (2008) estimated that a majority of doctoral psychology students were unable to assess the reliability or validity of tests. Given that expertise in psychometrics is unlikely to develop without guidance and instruction (Canivez, 2019), this article describes an evidence-based approach that clinicians can employ to assess the psychometric utility of IQ scores, supplies software tools to assist in that analysis, and provides a tutorial example using WISC-V scores.

An Evidence-Based Approach to IQ Score Interpretation

Specialists in educational and psychological measurement have developed methods to determine how precisely test scores reflect their intended constructs and whether scores provide sufficient unique information independent of each other (Brunner et al., 2012; Canivez & Youngstrom, 2019; Chen et al., 2012; Chen & Zhang, 2018; Ferrando & Lorenzo-Seva, 2018; Ferrando & Navarro-González, 2018; Reise et al., 2013, 2018; Rodriguez et al., 2016a, 2016b; Styck, 2019). Those methods undergird this evidence-based approach to IQ score interpretation.

Preliminary Information

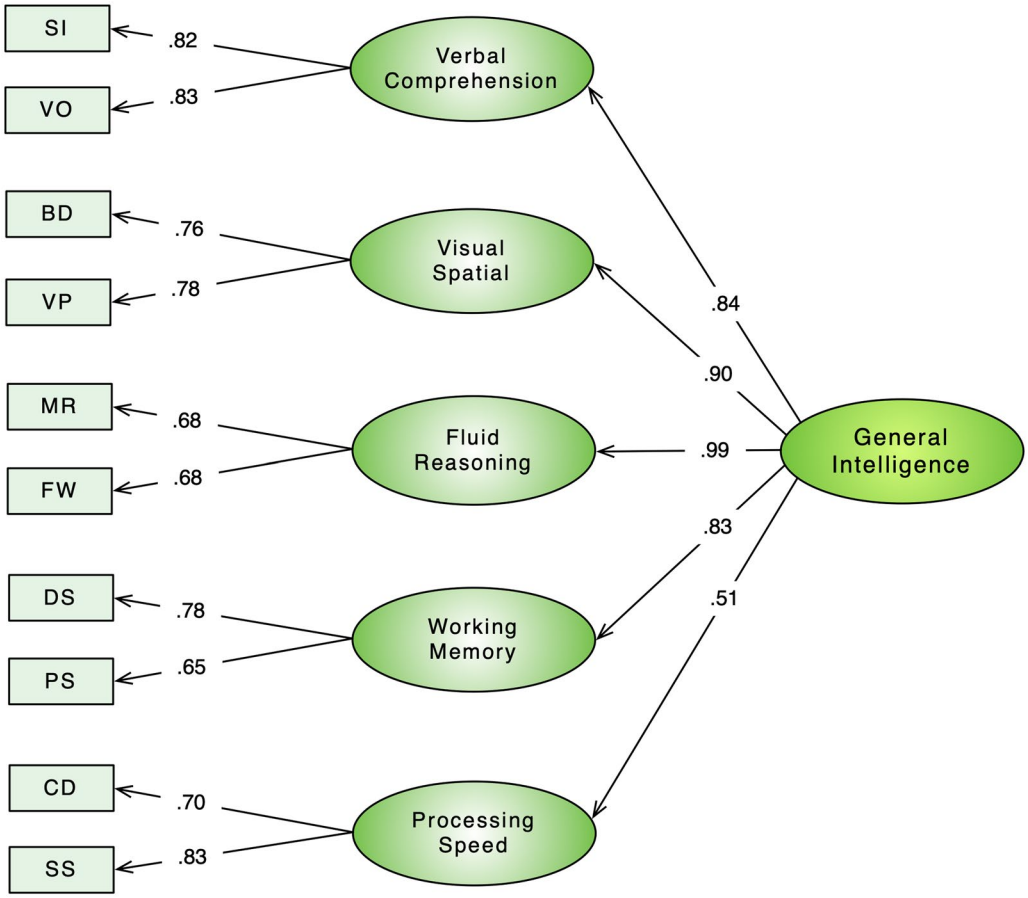
Preliminary information is needed for computation of the indices that will subsequently be used to judge the psychometric utility of IQ scores (Benson et al., 2018; Canivez & Youngstrom, 2019; Chen et al., 2012; Reise, 2012; Reise et al., 2013, 2018; Rodriguez et al., 2016a, 2016b). Specifically, results from a factor analysis with the same number of factors as specified in its scoring structure (e.g., five group factors and one general factor for the WISC-V). This information is frequently displayed in the test's technical manual as a higher-order factor model as illustrated in Figure 2 for the WISC-V.

Factor Transformation

As previously described, higher-order factor models conflate general and group factor variance, making interpretation of factor scores ambiguous. However, conceptual clarity can be attained with a mathematical transformation of the higher-order model via the Schmid-Leiman (S-L; Schmid & Leiman, 1957) procedure as illustrated in the bottom panel of Figure 1 (Brunner et al., 2012; Chen & Zhang, 2018; Gustafsson & Åberg-Bengtsson, 2010). In this transformed model, the general and group ability factors are all directly related to the indicator variables (i.e., subtest scores) and are uncorrelated with each other (i.e., they are orthogonal). This enhances "the interpretability of higher order and lower order factors" (Brunner et al., 2012, p. 808) by estimating the direct and unique influence

 DOI: 10.1080/2372966X.2020.181680

Figure 2. Higher-Order Measurement Model With Standardized Coefficients for the Wechsler Intelligence Scale for Children–Fifth Edition



Note. Adapted from figure 5.2 of Wechsler (2014b). The higher-order general intelligence latent construct was mislabeled "FSIQ" by Wechsler (2014b).

of each factor on each subtest score. In fact, Carroll (1993) used the S-L transformation when developing his influential model of intelligence that was one theoretical foundation for the WISC-V (Wechsler, 2014b). Additionally, statistical simulations have found S-L results to be an accurate factor recovery method (Giordano & Waller, 2020). Statistically, the higher-order model is nested within the orthogonal model even though these models have different conceptual meanings that might otherwise be theoretically or statistically important (Bornovalova et al., 2020; Chen et al., 2012; Chen & Zhang, 2018; Markon, 2019; Reise et al., 2018; Sellbom & Tellegen, 2019).

Variance Decomposition

Next, decomposition of the test's variance is accomplished using the S-L transformed model as input. That is, separating the test's variance into that due to a general factor (variance common to all measured variables), group factors (variance uniquely shared by a group of measured variables), and uniqueness (i.e., reliable variance unique to a single observed variable plus measurement error).

Indices of Score Utility

The variance contributions of general and group factors and their interrelationships allow the computation of several indices that assess the reliability of test scores and whether the test is best viewed as unidimensional or multidimensional.[1] Together, these indices guide decisions about the utility of IQ scores.

Reliability

Strong reliability is one of the fundamental requirements of evidence-based assessment (Hunsley & Mash, 2008). The reliability of factors and factor scores can be estimated with the *H* index (Hancock & Mueller, 2001) and omega coefficients, respectively (McDonald, 1999; Rodriguez et al., 2016a, 2016b; Watkins, 2017). Factors are latent constructs that are used for theoretical or conceptual purposes, whereas factor scores are mathematically derived estimates of those constructs used to make clinical decisions. For example, verbal comprehension (VC) is a latent construct thought to be measured by the WISC-V, and the VCI is a score estimated to represent that construct.

Factors

A factor might be identified but not reliably specified. The reliability of factors can be estimated with the *H* index (Hancock & Mueller, 2001). *H* is the correlation between a factor and an optimally weighted factor score and is considered a measure of construct reliability or replicability that quantifies how well a latent variable is represented by a set of indicators (Hancock & Mueller, 2001). According to Mueller and Hancock (2019), *H* is "an estimate of the correlation that a factor is expected to have with itself over repeated administrations" (p. 455). *H* values lower than .80 suggest that the factor is not well defined and will not replicate across studies nor provide accurate path coefficients if included in statistical models (Ferrando & Lorenzo-Seva, 2018; Mueller & Hancock, 2019; Rodriguez et al., 2016a).

Factor Scores

The reliability of unit-weighted factor scores can be indexed with omega coefficients, which make fewer and more realistic assumptions than traditional alpha coefficients (Watkins, 2017). Omega (ω) estimates the proportion of variance in a unit-weighted factor score that is attributable to all modeled sources of common variance. Omega-hierarchical (ω_h) estimates the proportion of variance in a unit-weighted factor score that is attributable to a single target factor after removing the variance due to all other sources. Thus, ω indicates how precisely a score measures the blend of general and group factors, whereas ω_h specifies how precisely a score measures a single factor independent of all other factors.[2] A comparison of ω to ω_h reveals how the reliability of a factor score has been inflated by multidimensionality. Omega can also be seen as a validity measure because it addresses the proportion of variance contributed by latent constructs (Brunner et al., 2012; Gustafsson & Åberg-Bengtsson, 2010).

Like traditional estimates of internal consistency reliability, omega indexes the total systematic variance in each unit-weighted score, whatever its source, and its magnitude should probably be judged similarly. Unfortunately, there is no consensus on what constitutes adequate reliability: experts have suggested minimums as low as .70 and as high as .96 (Kelley, 1927; Kline, 1998). However, evidence-based assessment guidelines recommend minimal values of .80 to .90 for clinical applications (Hunsley & Mash, 2008), and a review of cognitive test score reliability in the professional literature found an average of .85 (Charter, 2003). Thorndike and Thorndike-Christ (2010) argued that reliability estimates for making decisions about individuals should reach .80 at a minimum. Consequently, .80 was recognized as the guideline for judging omega estimates in this study, although .90 might be a preferable minimum for confident interpretation of IQ scores (Kranzler & Floyd, 2013).

There is also no universally accepted guideline for acceptable or adequate levels of ω_h for clinical decisions, but values less than .50 indicate that less than 50% of the score variance is due to the target factor, making "meaningful interpretation of [those scores] arguably impossible" (Gignac & Watkins, 2013, p. 658). Consequently, .75 might be a preferable guideline for confident score interpretation (Canivez & Youngstrom, 2019; Reise, 2012; Reise et al., 2013; Watkins, 2017). Some researchers (Giofrè et al., 2019) have accepted ω_h values lower than .50 and cited Gignac and Kretzschmar (2017) for support. However, Gignac and Kretzschmar proposed those lower guidelines "within the context of pure research" (p. 140) and "did not mean for those guidelines to be applied to clinical interpretation" (G. Gignac, personal communication, October 21, 2019).

Dimensionality

Although an IQ score is likely to be multidimensional, it is possible that it is *essentially* unidimensional. That is, unidimensional enough that the score can be interpreted as a measure of its purported construct without excessive bias (Rodriguez et al., 2016a, 2016b). As described by Reise et al. (2013), the assumption of unidimensionality "is a convenient fiction, sometimes useful in applied contexts" (p. 136).

Two indices contribute to an evaluation of test dimensionality: (a) percentage of uncontaminated correlations (PUC) and (b) explained common variance (ECV). PUC is the proportion of subtest correlations that are uncontaminated by multidimensionality. PUC values ≥ .80 support essential unidimensionality (Rodriguez et al., 2016a, 2016b). ECV is an index of general factor strength computed as a ratio of the variance explained by the general factor to the total common variance. ECV values ≥ .70 suggest that minimal bias would result from estimating a unidimensional factor even for data that are multidimensional (Gu et al., 2017; Rodriguez et al., 2016a, 2016b; Sellbom & Tellegen, 2019). However, ECV decreases in importance as an indicator of bias as the PUC increases (Rodriguez et al., 2016b), so an IQ score might be considered essentially unidimensional if ECV and PUC are both ≥ .70 (Gu et al., 2017; Rodriguez et al., 2016a, 2016b; Sellbom & Tellegen, 2019).

WISC-V TUTORIAL EXAMPLE

The WISC-V (Wechsler, 2014a) primary battery contains 10 core subtests, each with a population mean of 10 and standard deviation of 3. Five unit-weighted

 DOI: 10.1080/2372966X.2020.181680

factor index scores are produced from those 10 subtests: the VCI from the Similarities (SI) and Vocabulary (VO) subtests; the Visual Spatial Index (VSI) from the Block Design (BD) and Visual Puzzles (VP) subtests; the Fluid Reasoning Index (FRI) from the Matrix Reasoning (MR) and Figure Weights (FW) subtests; the Working Memory Index (WMI) from the Digit Span (DS) and Picture Span (PS) subtests; and the Processing Speed Index (PSI) from the Coding (CD) and Symbol Search (SS) subtests. The factor index and FSIQ scores each have a population mean of 100 and standard deviation of 15. A higher-order measurement model was provided by Wechsler (2014b) and an adapted version is presented in Figure 2.

Factor Transformation

Figure 3 illustrates use of the MacOrtho program (Watkins, 2020) to input the first-order factor loadings reported by Wechsler (2014b), and Figure 4 displays the S-L transformation of that first-order structure. The resulting orthogonal model appears to be appropriate given that the general factor is loaded by all indicator variables (from .357 for CD to .697 for VO; Chen & Zhang, 2018; Eid et al., 2017; Sellbom & Tellegen, 2019). This information will subsequently be used to determine how precisely WISC-V scores reflect their intended constructs and whether WISC-V scores provide unique information independent of each other.

Variance Decomposition

The sources of variance in the WISC-V for this normative sample, based on results of the S-L transformation by the MacOrtho program and variance decomposition by the Omega program, are presented in Figure 5. Considerable research has indicated that subtests contain too little specific variance to be useful (McDermott et al., 1992). As long ago as 1959, for example, Cohen (1959) concluded that subtests were "quite inadequate to serve as a basis for a subtest-specific rationale" (p. 290), and similar judgments have been repeated over the ensuing decades (Styck et al., 2019; Watkins, 2003; Zaboski et al., 2018). Unique variance exceeded the communality (i.e., variance contributed by general and group factors) for the MR, FW, PS, CD, and SS subtests. Similarly, the variance contributed by the general factor exceeded the variance due to the corresponding group factor for all subtests except CD and SS.

Figure 3. Input for MacOrtho Software to Perform a Schmid-Leiman Transformation of the Wechsler Intelligence Scale for Children–Fifth Edition Higher-Order Model

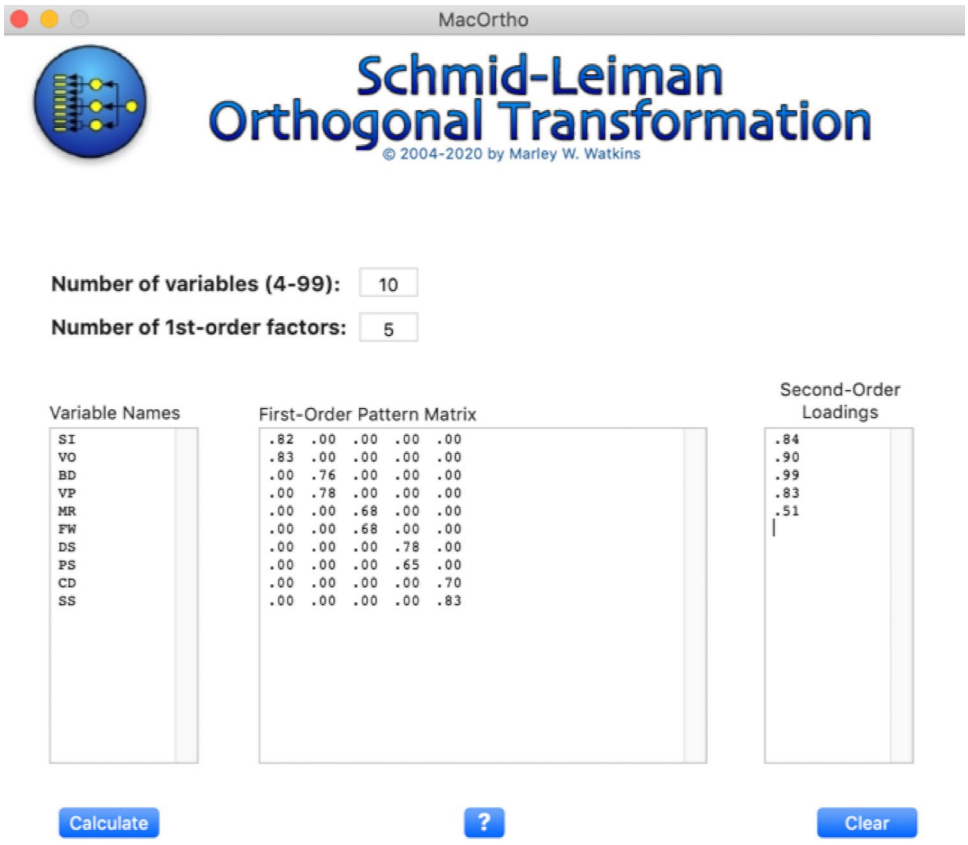
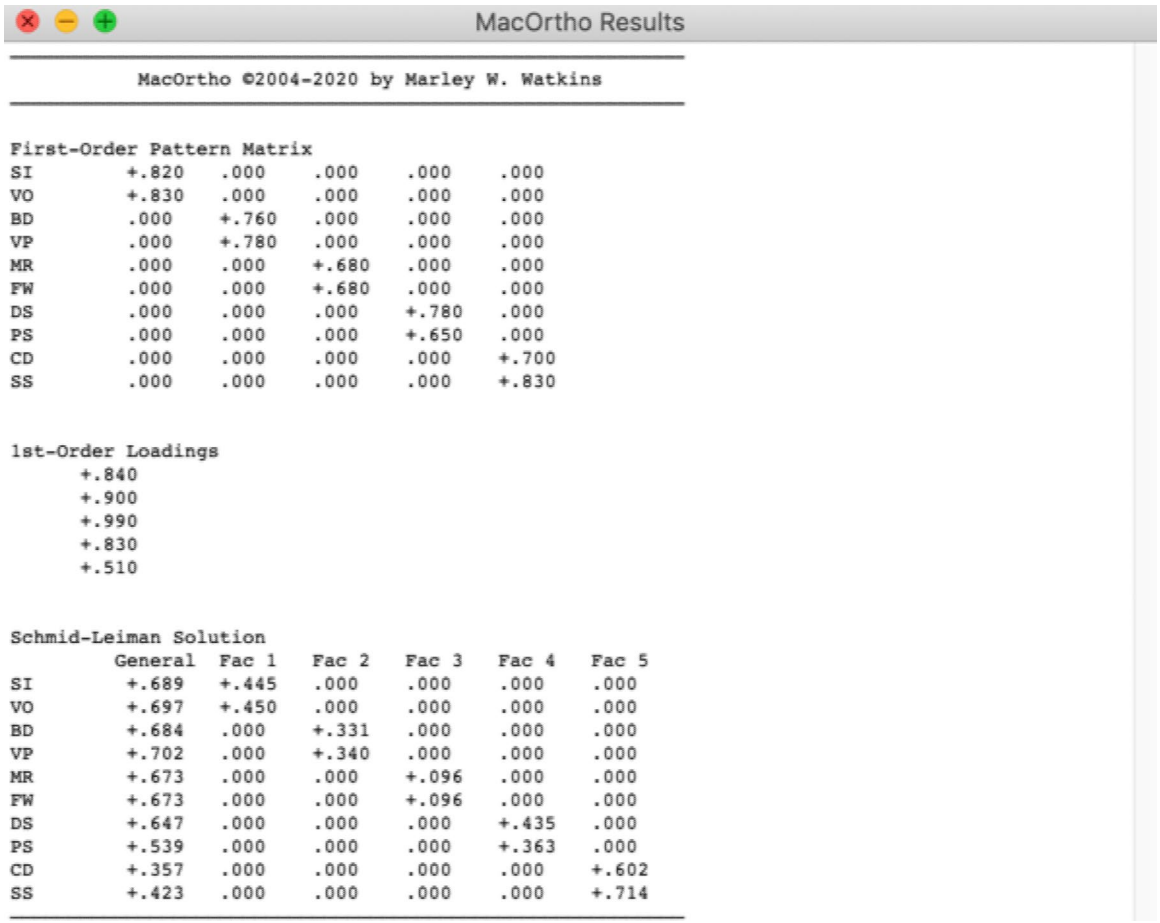


Figure 4. Schmid-Leiman Transformation of the Wechsler Intelligence Scale for Children–Fifth Edition Higher-Order Model Computed by MacOrtho Software



Note. SI = Similarities, VO = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, and SS = Symbol Search.

The general factor accounted for 38.4% of the total variance and the group factors contributed another 18.4%, leaving 43.2% unexplained. The general factor accounted for 67.7% of the common variance, more than twice the amount contributed by the combined group factors. The FR factor was particularly weak, accounting for only 0.2% to 0.3% of the total and common variance, respectively. The relative variance contributions of all WISC-V subtest and composite scores are detailed in the Omega program output (see Figure 5) and visually illustrated in Figure 6.

Indices of Score Utility

Figure 5 also displays indices of score utility produced by the Omega program for the WISC-V normative data; that is, *H* and omega values to judge reliability as well as ECV and PUC values to evaluate dimensionality. Those indices are also reported in Table 1 in a format that could be used as a checklist for other instruments.

Reliability

In terms of *factor* reliability, as judged by the *H* index, the five WISC-V group factors were not well defined and will be unlikely to replicate across studies (i.e., $H < .80$). In contrast, the general factor was well defined and should replicate ($H = .873$). That is, an optimal composite of WISC-V subtests can explain 87% of the variability in the general factor, 33% of the variability in the VC factor, 20% of the variability in the VS factor, 2% of the variability in the FR factor, 28% of the variability in the WM factor, and 62% of the variance in the PS factor.

How precisely a score measures the blend of general and group constructs is indexed by the ω coefficient. In that regard, only the FSIQ and VCI were reliable enough for high-stakes decisions about individuals (ω = .904 and .810, respectively). How precisely a score measures a single construct independent of all other constructs is indexed by ω_h, with values less than .50 making "meaningful interpretation of [those scores] arguably impossible" (Gignac & Watkins, 2013, p. 658). By this standard, the VCI, VSI,

 DOI: 10.1080/2372966X.2020.181680

Figure 5. Variance Decomposition and Psychometric Utility Indicators From Omega Software Based on the Schmid-Leiman Transformation of the Wechsler Intelligence Scale for Children–Fifth Edition Higher-Order Model

Omega Reliability for Bifactor Models

Variable	Factor	g Loading	g Variance	Group Loading	Group Var.	Communality (Total Var)	Unique Var	ECV
SI	VC	.689	.475	.445	.198	.673	.327	.706
VO	VC	.697	.486	.450	.203	.688	.312	.706
BD	VS	.684	.468	.331	.110	.577	.423	.810
VP	VS	.702	.493	.340	.116	.608	.392	.810
MR	FR	.673	.453	.096	.009	.462	.538	.980
FW	FR	.673	.453	.096	.009	.462	.538	.980
DS	WM	.647	.419	.435	.189	.608	.392	.689
PS	WM	.539	.291	.363	.132	.422	.578	.688
CD	PSP	.357	.127	.602	.362	.490	.510	.260
SS	PSP	.423	.179	.714	.510	.689	.311	.260

	g	VC	VS	FR	WM	PSP	Unique
Total variance	.384	.040	.023	.002	.032	.087	.432
Common Var. (ECV)	.677	.071	.040	.003	.057	.154	
Omega	.904	.810	.744	.632	.678	.740	
Omega(h)	.823	.238	.141	.013	.211	.548	
Relative Omega	.910	.294	.190	.020	.312	.740	
Factor correlation	.907	.488	.376	.112	.460	.740	
H	.873	.334	.202	.018	.278	.617	
PUC	.889						
FDI	.934	.578	.450	.135	.527	.785	
Source of Explained Variance in Factor Scores							
Unique	.096	.190	.256	.368	.322	.260	
This Factor	.823	.238	.141	.013	.211	.548	
Other Factors	.081	.572	.603	.620	.467	.192	

Note. SI = Similarities, VO = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search, VC = Verbal Comprehension factor, VS = Visual Spatial factor, FR = Fluid Reasoning factor, WM = Working Memory factor, PSP = Processing Speed factor, g = general factor, *H* is from Mueller and Hancock (2019), ECV = explained common variance, and PUC = percent of uncontaminated correlations.

FRI, and WMI scores were uninterpretable. The ω_h index of .548 for the PSI met this minimal criterion but failed to reach the level preferred for confident score interpretation (i.e., ≥ .75; Canivez & Youngstrom, 2019; Reise, 2012; Reise et al., 2013; Watkins, 2017).

Dimensionality

According to Wechsler (2014b), the WISC-V is multidimensional, offering five group factor index scores (VCI, VSI, FRI, WMI, and PSI) and one general factor score (FSIQ). If each score is essentially unidimensional it might be interpreted without excessive bias. PUC values ≥ .80 and ECV values ≥ .70 would signal essential unidimensionality (Gu et al., 2017; Rodriguez et al., 2016a, 2016b; Sellbom & Tellegen, 2019), which was achieved by the FSIQ (PUC = .89, ECV = .68) but none of the five group factor index scores.

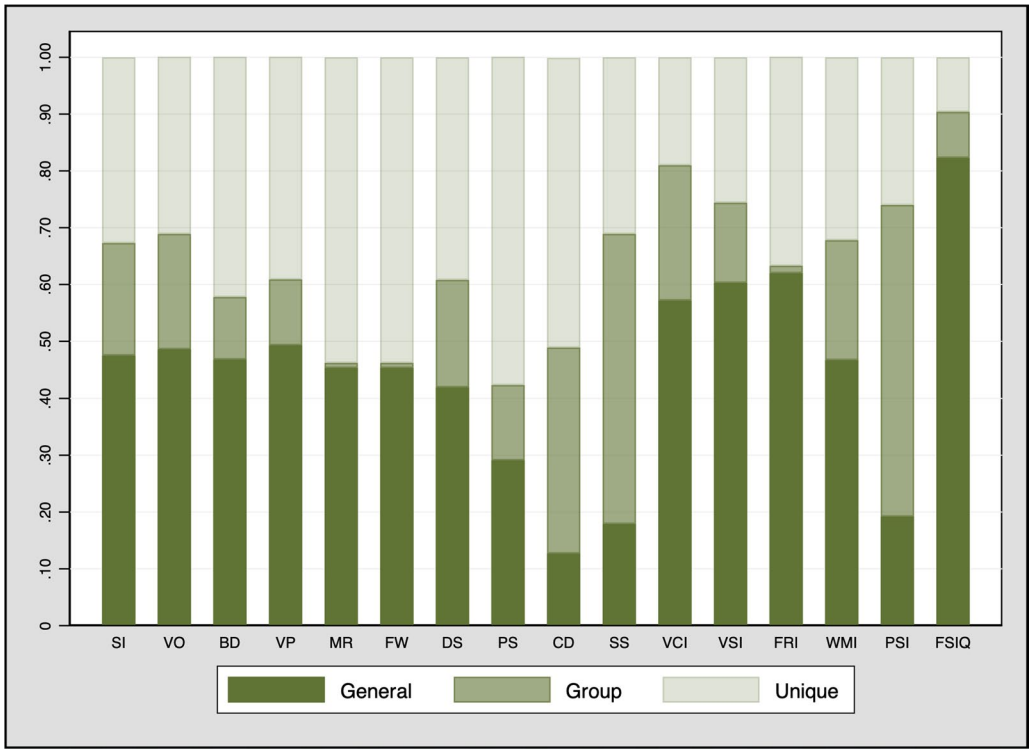
Replication

Reliability estimates from the field may differ from estimates derived from a test's standardization sample (Thorndike & Thorndike-Christ, 2010), necessitating a replication of normative results in clinical samples. Fortunately, Canivez et al. (2020) recently provided reliability and dimensionality indicators from a large clinical sample. Results from both normative and clinical samples were in agreement as demonstrated in Table 1.

GENERAL DISCUSSION

Modern IQ tests provide numerous subtest and composite scores but valid interpretation of those scores is dependent on how reliably each score reflects its intended construct and whether it provides unique information independent of other constructs (Brunner et al., 2012; Canivez & Youngstrom, 2019; Chen et al., 2012; Ferrando &

Figure 6. General, Group, and Unique Sources of Variance in Wechsler Intelligence Scale for Children–Fifth Edition Subtest and Composite Scores



Note. The *Y*-axis displays total variance for subtests and explained common variance for composite scores.

Table 1. Psychometric Utility of Wechsler Intelligence Scale for Children–Fifth Edition Factor Scores From Normative ($N = 2,200$) and Clinical ($N = 2,512$) Samples

	General	VC	VS	FR	WM	PS
Dimensionality						
PUC						
Norm	***.889***	—	—	—	—	—
Clinical	***.889***	—	—	—	—	—
Essential unidimensional (≥.80)	Yes	—	—	—	—	—
ECV						
Norm	.677	.071	.040	.003	.057	.154
Clinical	***.733***	.085	.041	.008	.049	.083
Essential unidimensional (≥.70)	Yes	No	No	No	No	No
Reliability						
Constructs (*H*)						
Norm	**.873**	.334	.202	.018	.278	.617
Clinical	***.902***	.413	.225	.050	.265	.411
Minimum reliability (≥.80)	Yes	No	No	No	No	No
Preferable reliability (≥.90)	Yes	No	No	No	No	No
Scores (ω)						
Norm	***.904***	**.810**	.744	.632	.678	.740
Clinical	***.924***	**.846**	**.823**	.746	.722	.641
Minimum reliability (≥.80)	Yes	Yes	No	No	No	No
Preferable reliability (≥.90)	Yes	No	No	No	No	No
Scores (ω_h)						
Norm	***.823***	.238	.141	.013	.211	**.548**
Clinical	***.860***	.300	.149	.032	.193	.345
Minimum reliability (≥.50)	Yes	No	No	No	No	Yes
Preferable reliability (≥.75)	Yes	No	No	No	No	No

Note. Metrics meeting minimum standards in bold and metrics meeting preferable standards in bold italic.
VC = Verbal Comprehension, VS = Visual Spatial, FR = Fluid Reasoning, WM = Working Memory, PS = Processing Speed, ω = omega, ω_h = omega hierarchical, *H* = construct replicability index, and PUC = percent of uncontaminated correlations. Metrics meeting minimum standards in bold and metrics meeting preferable standards in bold italic. Yes-No decision based on normative sample metric.

 DOI: 10.1080/2372966X.2020.181680

Lorenzo-Seva, 2018; Ferrando & Navarro-González, 2018; Reise et al., 2013, 2018; Rodriguez et al., 2016a, 2016b; Wainer & Feinberg, 2015). Accordingly, IQ scores must be evaluated for their reliability (measured by omega coefficients) and their dimensionality (measured by ECV and PUC indices) to determine their psychometric utility. These measures of utility are unlikely to be provided by test publishers, but can easily be generated with the tools provided in this tutorial.

In this example using WISC-V normative sample data, WISC-V group *factors* were not well defined by their indicators (i.e., subtests), and WISC-V index (i.e., factor) *scores* were unreliable and contaminated with variance from other constructs. Similar results have been found with other IQ tests, making this an almost universal conclusion (Benson et al., 2018; Canivez et al., 2019, 2020; Dombrowski et al., 2018, 2019; Fenollar-Cortés et al., 2019; Gignac & Watkins, 2013; Gomez et al., 2019; Styck, 2019; Watkins, 2006, 2018). In contrast, the WISC-V FSIQ score was essentially unidimensional and sufficiently reliable for clinical use. Similar results have been found for educational tests where subscores have generally added little value beyond total scores (Wainer & Feinberg, 2015).

These conclusions have been based on estimates of structural validity and assume that the factor structure is an accurate representation of the underlying structure of the test. If severely misspecified, factor analytic results may be biased estimates of population values. Potential symptoms of a structural mismatch include nonconvergence of factor models, poorly fitting factor models, and mathematically inadmissible parameter values (Brown, 2015; Chen & Zhang, 2018). Likewise, caution should be exercised when all subtests load at near-zero levels on a group factor, multiple subtests load at near-zero levels on the general factor, or the general factor accounts for a miniscule proportion of the test's variance (Eid et al., 2017; Gorsuch, 1983). Although validity evidence can be obtained from test publishers and independent researchers, "the test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is used" (AERA et al., 2014, p. 13). Thus, it is vital that reports of validity evidence, including those contributed by test publishers, be based on a transparent account of all measurement decisions and avoid questionable measurement practices (Flake & Fried, 2020).

Given that "validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible" (AERA et al., 2014, p. 22), the content validity of WISC-V scores, their stability over time, and their relationship with other variables must also be considered (Ferrando & Lorenzo-Seva, 2019a). The FSIQ has usually been more reliable than factor index scores across time (e.g., Watkins & Smith, 2013), and factor index scores have often exhibited little ability to predict academic achievement beyond the FSIQ (with the possible exception of the VCI) even in the presence of index score scatter (Canivez et al., 2014; Daniel, 2007; Freberg et al., 2008; Glutting et al., 1997; Oh et al., 2004; Styck, 2019; Watkins et al., 2007). Further, little evidence has been found to support the diagnostic utility and treatment validity of group factor scores (Benson et al., 2018; Braden, 2013; Braden & Niebling, 2012; Burns et al., 2016; Canivez & Youngstrom, 2019; Freeman & Chen, 2019; Kranzler et al., 2016; McGill, 2018; McGill et al., 2018; Styck & Watkins, 2013; Zaboski et al., 2018).

Nevertheless, expert recommendations for the interpretation of cognitive ability test scores (Flanagan & Alfonso, 2017; Groth-Marnat & Wright, 2016; Kaufman et al., 2016; J. L. Miller et al., 2016; Sattler et al., 2016; Wechsler, 2014b) are popular among school psychology practitioners and trainers (Benson et al., 2020; Lockwood & Farmer, 2020; L. T. Miller et al., 2020; Sotelo-Dynega & Dixon, 2014). These interpretational strategies often rest on evidence regarding the test's structure obtained via a factor analysis that is used to justify univocal interpretation of factor scores (Braden, 2013; Braden & Niebling, 2012; Canivez & Youngstrom, 2019). However, the fit of data to a model, as in a confirmatory factor analysis, does not guarantee that the resulting factors are precisely specified or that factor scores are reliable measures of the target construct that can provide accurate individual measurement (Beaujean & Benson, 2019; Benson et al., 2018; Ferrando & Lorenzo-Seva, 2019a, 2019b; Ferrando & Navarro-González, 2018; Rodriguez et al., 2016a, 2016b).

In sum, "to make valid and useful clinical judgments, clinicians must understand the dimensionality of the constructs they assess and of the measures used to asses them" (Haynes et al., 2019, p. 151). Therefore, clinicians are encouraged to eschew "*eminence-based* practices" in favor of "*evidence-based* practices" (Kranzler et al., 2020, p. 10) by responding in the affirmative to two questions *before* interpreting any IQ score: (a) Is this score a sufficiently reliable measure of the multidimensional constructs it purports to measure (i.e., ω at least $\geq .80$ and preferably $\geq .90$)? and (b) Is this score a sufficiently reliable measure of the single construct it purports to measure independent of other constructs (i.e., ω_h at least $\geq .50$ and preferably $\geq .75$)? A third question must be answered positively if the scores are to be quantitatively compared to identify intraindividual cognitive strengths and weaknesses (i.e., use of difference scores to create ipsative profiles). Namely, do these scores retain sufficient reliability for clinical decisions given that difference scores are less reliable than their constituent scores

(Thorndike & Thorndike-Christ, 2010)? For example, Farmer and Kim (2020) reported that the median WISC-V subtest and composite difference score reliability was .70 and .81, respectively. Given the psychometric limitations of difference and ipsative scores (Beaujean & Benson, 2019; McDermott et al., 1992), it is unlikely that cognitive profiles will exhibit sufficient reliability for clinical decisions (i.e., at least ≥ .80 and preferably ≥ .90). A final question must be answered in the affirmative if cognitive ability scores are to be used for diagnostic determinations or treatment recommendations. Specifically, is there evidence to support the diagnostic utility, predictive validity, or treatment validity of this score (AERA et al., 2014; Reynolds & Milam, 2012)? By following these evidence-based practices, practitioners can demonstrate that they know "what tests can and cannot do" (Weiner, 1989, p. 830).

NOTES

1. Factor transformation via the S-L procedure can be directly implemented with the free *psych* package in the R software system (R Development Core Team, 2020), the free FACTOR program (Lorenzo-Seva & Ferrando, 2006), and the SPSS system with syntax provided by Wolff and Preising (2005). Alternatively, the S-L procedure can be indirectly computed from the validity information extracted from the test's technical manual with the MacOrtho program (Watkins, 2020). Variance decomposition and indices of score quality can be completed with a versatile spreadsheet contributed by Dueber (2017) that is available at https://uknowledge.uky.edu/edp_tools/1/. Similar metrics can be extracted from two R packages: BifactorIndicesCalculator at https://cran.r-project.org/web/packages/BifactorIndicesCalculator/index.html and *psych* at https://cran.r-project.org/web/packages/psych/index.html. The free Omega (Watkins, 2013) program can also accomplish these tasks.
2. The labels applied to omega coefficients have been inconsistent. Some authors use specific labels for omega coefficients applied to general and group factors. For example, ω for the amalgam of general and group factor variance in the general factor score (i.e., FSIQ), ω_s for the amalgam of general and group factor variance in the group factor scores (i.e., VCI, VSI, etc.), ω_h for the general factor variance in the general factor score, and ω_{hs} for the group factor variance in the group factor scores.

DISCLOSURE

The authors declare that they have no conflict of interest.

ORCID

Marley W. Watkins http://orcid.org/0000-0001-6352-7174
Gary L. Canivez http://orcid.org/0000-0002-5347-6534

REFERENCES

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *The American Psychologist*, *63*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Authors.

Beaujean, A. A., & Benson, N. F. (2019). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*, *23*(2), 126–137. https://doi.org/10.1007/s40688-018-0182-1

Benson, N. F., Beaujean, A. A., McGill, R. J., & Dombrowski, S. C. (2018). Revisiting Carroll's survey of factor-analytic studies: Implications for the clinical assessment of intelligence. *Psychological Assessment*, *30*(8), 1028–1038. https://doi.org/10.1037/pas0000556

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. https://doi.org/10.1016/j.jsp.2018.12.004

Benson, N. F., Maki, K. E., Floyd, R. G., Eckert, T. L., Kranzler, J. H., & Fefer, S. A. (2020). A national survey of school psychologists' practices in identifying specific learning disabilities. *School Psychology (Washington, D.C.)*, *35*(2), 146–157. https://doi.org/10.1037/spq0000344

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Braden, J. P. (2013). Psychological assessment in school settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology.* (Vol. 10, 2nd ed., pp. 291–314). Wiley.

Braden, J. P., & Niebling, B. C. (2012). Using the joint test standards to evaluate the validity evidence for intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues.* (3rd ed., pp. 739–757). Guilford.

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

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846. https://doi.org/10.1111/j.1467-6494.2011.00749.x

Burns, M. K., Petersen-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., Clayton, K., Hutcheson, S., Conner, C., Hosp, J., & VanDerHeyden, A. M. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association*, *31*(1), 28–42. https://doi.org/10.1037/spq0000117

Canivez, G. L. (2019). Evidence-based assessment for school psychology: Research, training, and clinical practice.

Contemporary School Psychology, *23*(2), 194–200. https://doi.org/10.1007/s40688-019-00238-z

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*(2), 274–296. https://doi.org/10.1177/1073191118811609

Canivez, G. L., Watkins, M. W., James, T., Good, R., & James, K. (2014). Incremental validity of WISC-IV(UK) factor index scores with a referred Irish sample: predicting performance on the WIAT-II(UK.). *The British Journal of Educational Psychology*, *84*(Pt 4), 667–684. https://doi.org/10.1111/bjep.12056

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. *The British Journal of Educational Psychology*, *89*(2), 195–224. https://doi.org/10.1111/bjep.12230

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical and policy implications. *Applied Measurement in Education*, *32*(3), 232–248. https://doi.org/10.1080/08957347.2019.1619562

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2016). *The twentieth mental measurements yearbook*. Buros Institute of Mental Measurements.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, *130*(3), 290–304. https://doi.org/10.1080/00221300309601160

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Chen, F. F., & Zhang, Z. (2018). Bifactor models in psychometric test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *Wiley handbook of psychometric testing.* (pp. 325–345). Wiley.

Cohen, J. (1959). The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology*, *23*(4), 285–299. https://doi.org/10.1037/h0043898

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: comment on Fiorello et al. (2007). *Applied Neuropsychology*, *14*(4), 291–295. https://doi.org/10.1080/09084280701719401

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*(1), 90–104. https://doi.org/10.1007/s40688-017-0125-2

Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2019). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation, and theory. *Assessment*, 107319111986982. Advance online publication. https://doi.org/10.1177/1073191119869828

Dueber, D. M. (2017). Bifactor indices calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. https://uknowledge.uky.edu/edp_tools/1/

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in ***G***-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Farmer, R. L., & Kim, S. Y. (2020). Difference score reliabilities with the RIAS-2 and WISC-V. *Psychology in the Schools*, *57*(8), 1273–1288. https://doi.org/10.1002/pits.22369

Fenollar-Cortés, J., López-Pinar, C., & Watkins, M. W. (2019). Structural validity of the Spanish Wechsler Intelligence Scale for Children–Fourth Edition in a large sample of Spanish children with attention-deficit hyperactivity disorder. *International Journal of School & Educational Psychology*, *7*(sup1), 2–14. https://doi.org/10.1080/21683603.2018.1474820

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*(5), 762–780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., & Lorenzo-Seva, U. (2019a). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement*, *79*(3), 437–461. https://doi.org/10.1177/0013164418824755

Ferrando, P. J., & Lorenzo-Seva, U. (2019b). On the added value of multiple factor score estimates in essentially unidimensional models. *Educational and Psychological Measurement*, *79*(2), 249–271. https://doi.org/10.1177/0013164418773851

Ferrando, P. J., & Navarro-González, D. (2018). Assessing the quality and usefulness of factor-analytic applications to personality measures: A study with the statistical anxiety scale. *Personality and Individual Differences*, *123*, 81–86. https://doi.org/10.1016/j.paid.2017.11.014

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

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

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences du Comportement*, *49*(2), 78–88. https://doi.org/10.1037/cbs0000069

Freberg, M. E., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III full scale IQ in predicting later academic achievement. *Applied Neuropsychology*, *15*(2), 131–139. https://doi.org/10.1080/09084280802084010

Freeman, A. J., & Chen, Y.-L. (2019). Interpreting pediatric intelligence tests: A framework from evidence-based medicine. In G. Goldstein, D. N. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment.* (4th ed., pp. 65–101). Academic Press.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. Sage.

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

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV.

Multivariate Behavioral Research, *48*(5), 639–662. https://doi.org/10.1080/00273171.2013.804398

Giofrè, D., Pastore, M., Cornoldi, C., & Toffalini, E. (2019). Lumpers vs. splitters: Intelligence in children with specific learning disorders. *Intelligence*, *76*, 101380–101310. https://doi.org/10.1016/j.intell.2019.101380

Giordano, C., & Waller, N. G. (2020). Recovering bifactor models: A comparison of seven methods. *Psychological Methods*, *25*(2), 143–156. https://doi.org/10.1037/met0000227

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*(3), 295–301. https://doi.org/10.1037/1040-3590.9.3.295

Gomez, R., Stavropoulos, V., Vance, A., & Griffiths, M. D. (2019). Re-evaluation of the latent structure of common childhood disorders: Is there a general psychopathology factor (P-factor)? *International Journal of Mental Health and Addiction*, *17*(2), 258–278. https://doi.org/10.1007/s11469-018-0017-3

Gorsuch, R. L. (1983). *Factor analysis*. (2nd ed.). Erlbaum.

Gould, J. W., Martindale, D. A., & Flens, J. R. (2013). Responsible use of psychological tests: Ethical and professional practice concerns. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *Oxford handbook of child psychological assessment.* (pp. 222–235). Oxford University Press.

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment*. (6th ed.)Wiley.

Gu, H., Wen, Z., & Fan, X. (2017). Examining and controlling for wording effect in a self-report measure: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 545–555. https://doi.org/10.1080/10705511.2017.1286228

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches.* (pp. 97–121). American Psychological Association.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudek, S. H. C. duToit, & D. F. Sorbom (Eds.). *Structural equation modeling: Present and future.* (pp. 195–216). Scientific Software.

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2019). *Scientific foundations of clinical assessment*. (2nd ed.). Taylor & Francis.

Hunsley, J., & Mash, E. J. (2008). *A guide to assessments that work*. Oxford University Press.

Jewsbury, P. A., & Bowden, S. C. (2017). Construct validity has a critical role in evidence-based neuropsychological assessment. In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice: Diagnostic and treatment evaluations.* (pp. 33–63). Oxford University Press.

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Wiley.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. Routledge.

Kranzler, J. H., Benson, N., & Floyd, R. G. (2016). Intellectual assessment of children and youth in the United States of America: Past, present, and future. *International Journal of School & Educational Psychology*, *4*(4), 276–282. https://doi.org/10.1080/21683603.2016.1166759

Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide*. Guilford.

Kranzler, J. H., Maki, K. E., Benson, N. F., Eckert, T. L., Floyd, R. G., & Fefer, S. A. (2020). How do school psychologists interpret intelligence tests for the identification of specific learning disabilities? *Contemporary School Psychology*, Advance online publication. https://doi.org/10.1007/s40688-020-00274-0

Lockwood, A. B., & Farmer, R. L. (2020). The cognitive assessment course: Two decades later. *Psychology in the Schools*, *57*(2), 265–283. https://doi.org/10.1002/pits.22298

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91. https://doi.org/10.3758/bf03192753

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*(1), 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education*, *25*(4), 504–526. https://doi.org/10.1177/002246699202500407

McDonald, R. P. (1999). *Test theory: A unified approach*. Erlbaum.

McGill, R. J. (2016). Invalidating the full scale IQ score in the presence of significant factor score variability: Clinical acumen or clinical illusion? *Archives of Assessment Psychology*, *6*(1), 33–63.

McGill, R. J. (2018). Confronting the base rate problem: More ups and downs for cognitive scatter analysis. *Contemporary School Psychology*, *22*(3), 384–393. https://doi.org/10.1007/s40688-017-0168-4

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

Miller, J. L., Saklofske, D. H., Weiss, L. G., Drozdick, L., Llorente, A. M., Holdnack, J. A., & Prifitera, A. (2016). Issues related to the WISC-V assessment of cognitive functioning in clinical and special groups. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives.* (pp. 287–343). Academic Press.

Miller, L. T., Bumpus, E. C., & Graves, S. L. (2020). The state of cognitive assessment training in school psychology: An analysis of syllabi. *Contemporary School Psychology*, Advance online publication. https://doi.org/10.1007/s40688-020-00305-w

Montgomery, A., Torres, E., & Eiseman, J. (2018). Using the joint test standards to evaluate the validity evidence for intelligence tests. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues.* (4th ed., pp. 841–852). Guilford.

Mueller, R. O., & Hancock, G. R. (2019). Structural equation modeling. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences.* (2nd ed., pp. 445–456). Routledge.

Oh, H.-J., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking

sample. *The Journal of Special Education*, *38*(3), 159–173. https://doi.org/10.1177/00224669040380030301

Perham, H. (2010). *Quantitative training of doctoral school psychologists: Statistics, measurement, and methodology curriculum* [Unpublished master's thesis]. Arizona State University.

R Development Core Team. (2020). *R: A language and environment for statistical computing [Computer program]*. R Foundation for Statistical Computing.

Reise, S. P. (2012). Invited Paper: The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *Wiley handbook of psychometric testing.* (pp. 677–707). Wiley.

Reynolds, C. R., & Milam, D. A. (2012). Challenging intellectual testing results. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony.* (6th ed., pp. 311–334). Oxford University Press.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Sattler, J. M., Dumont, R., & Coalson, D. L. (2016). *Assessment of children: WISC-V and WPPSI-IV*. Jerome M. Sattler Publisher.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive assessment practices: A survey of school psychologists. *Psychology in the Schools*, *51*(10), n/a–1045. https://doi.org/10.1002/pits.21802

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293. https://doi.org/10.2307/1412107

Styck, K. M. (2019). Psychometric issues pertaining to the measurement of specific broad and narrow intellectual abilities. In D. J. McFarland (Ed.), *General and specific mental abilities.* (pp. 80–107). Cambridge Scholars Publishing.

Styck, K. M., Beaujean, A. A., & Watkins, M. W. (2019). Profile reliability of cognitive ability subscores in a referred sample. *Archives of Scientific Psychology*, *7*(1), 119–128. https://doi.org/10.1037/arc0000064

Styck, K. M., & Watkins, M. W. (2013). Diagnostic utility of the culture-language interpretive matrix for the Wechsler Intelligence Scale for Children–Fourth Edition among referred students. *School Psychology Review*, *42*(4), 367–382.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education*. (8th ed.). Pearson.

Wainer, H., & Feinberg, R. (2015). For want of a nail: Why unnecessarily long tests may be impeding the progress of Western civilization. *Significance*, *12*(1), 16–21. https://doi.org/10.1111/j.1740-9713.2015.00797.x

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *Scientific Review of Mental Health Practice*, *2*(2), 118–141.

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment. 18*(1), 123–125. https://doi.org/10.1037/1040-3590.18.1.123

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology.* (4th ed., pp. 210–229). Wiley.

Watkins, M. W. (2013). *Omega* [Computer software]. Ed & Psych Associates. https://edpsychassociates.com/Watkins3.html

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, *31*(6-7), 1113–1126. https://doi.org/10.1080/13854046.2017.1317364

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Watkins, M. W. (2020). *MacOrtho* [Computer software]. Ed & Psych Associates. https://edpsychassociates.com/Watkins3.html

Watkins, M. W., Glutting, J. J., & Lei, P.-W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology*, *14*(1), 13–20. https://doi.org/10.1080/09084280701280353

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment*, *25*(2), 477–483. https://doi.org/10.1037/a0031653

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children–Fifth Edition*. Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children–Fifth Edition: Technical and interpretive manual*. Pearson.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment*, *53*(4), 827–831. https://doi.org/10.1207/s15327752jpa5304_18

Wolff, H.-G., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods*, *37*(1), 48–58. https://doi.org/10.3758/bf03206397

Zaboski, B. A., Kranzler, J. H., & Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Cattell-Horn-Carroll theory. *Journal of School Psychology*, *71*, 42–56. https://doi.org/10.1016/j.jsp.2018.10.001

AUTHOR BIOGRAPHICAL STATEMENTS

Marley W. Watkins, PhD, received his PhD in school psychology from the University of Nebraska–Lincoln and has held positions with the Deer Valley Unified School District,

Pennsylvania State University, Arizona State University, and Baylor University. He is currently a research professor in the Department of Educational Psychology at Baylor University. His research interests include professional issues, the psychometrics of assessment and diagnosis, individual differences, and computer applications. Dr. Watkins has published more than 170 peer-reviewed journal articles and made more than 125 presentations at professional conferences.

Gary L. Canivez, PhD, is professor of psychology at Eastern Illinois University and principally involved in the Specialist in School Psychology program. Before entering academia Dr. Canivez was a school psychologist for 8 years, was on the adjunct faculty of Arizona State University and Northern Arizona University, and was president of the Arizona Association of School Psychologists. Dr. Canivez currently serves as associate editor of *Archives of Scientific Psychology* and served as associate editor for *Psychological Assessment*. He is a consulting editor and frequent reviewer for numerous other school psychology, assessment, and clinically oriented journals. Dr. Canivez is a member of the Society for the Study of School Psychology, a Fellow of the Division (5) of Quantitative and Qualitative Methods and Division (16) of School Psychology of the American Psychological Association, and a Charter Fellow of the Midwestern Psychological Association. The author of over 100 research and professional publications and over 200 professional presentations and continuing professional development workshops, Dr. Canivez has research interests in psychological assessment and measurement pertaining to intelligence, achievement, personality, and psychopathology. His research has been supported by the National Institutes of Health/National Institute of Mental Health.