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

CHAPTER 20

REVIEW OF THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN–FIFTH EDITION: CRITIQUE, COMMENTARY, AND INDEPENDENT ANALYSES

Gary L. Canivez and Marley W. Watkins

WISC–V REVIEW

Description

The Wechsler Intelligence Scale for Children–Fifth Edition (WISC–V; Wechsler, 2014) is the latest edition of Wechsler's test of child intelligence with its origin dating back to the first Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949). The WISC–V is a major revision of the Wechsler Intelligence Scale for Children–Fourth Edition (WISC–IV; Wechsler, 2003) with national standardization for youth ages 6 to 16 years. The WISC–V includes an *Administration and Scoring Manual*, an *Administration and Scoring Manual Supplement*; a *Technical and Interpretive Manual*; three stimulus books; a Record Form; two response booklets (Coding and Symbol Search [Response Booklet 1], Cancellation [Response Booklet 2]); a scoring key for Symbol Search, scoring templates for Coding and for Cancellation; and the standard Block Design set. While the *WISC–V Administration and Scoring Manual* includes norms and analyses tables for the Summary and Primary Analysis pages, norms and analysis tables for the Ancillary and Complementary Analysis and Process Analysis pages are included in the *WISC–V Administration and Scoring Manual Supplement*.

Pearson also makes available a *WISC–V Technical and Interpretive Manual Supplement: Special Group Validity Studies with Other Measures and Additional Tables* (Pearson, 2014), which is available as a free download at http://downloads.pearsonclinical.com/images/Assets/WISC-V/WISC-V-Tech-Manual-Supplement.pdf. Within this supplement are full correlation matrices and descriptive statistics by age. This is a welcome addition and a positive contrast to the WISC–IV^{UK}; there the publisher did not provide a technical manual disclosing psychometric characteristics of the UK standardization sample; the publisher also refused to make available standardization sample correlation matrices and descriptive statistics necessary for fully understanding the psychometric characteristics of WISC–IV^{UK} scores (Canivez, Watkins, James, James, & Good, 2014).

As with earlier editions, the WISC–V includes numerous subtests that provide estimates of general intelligence consistent with Wechsler's "global capacity" definition of intelligence (Wechsler, 1939, p. 229) but also are combined to measure various group factors. The WISC–V, like the WISC–IV, overlaps in age with the *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (WPPSI–IV; Wechsler, 2012) (age 6 years through 7 years 7 months) and the *Wechsler Adult Intelligence*

Scale–Fourth Edition (WAIS–IV; Wechsler, 2008) (age 16 years) to allow clinicians the opportunity to select the more appropriate instrument depending on referral question and child characteristics.

Development

The *WISC–V Technical and Interpretive Manual* notes that revision goals included updating theoretical foundations, increasing developmental appropriateness, increasing user-friendliness, improving psychometric properties, and enhancing clinical utility and that these goals were based on considerations of structural models of intelligence, neurodevelopmental and neurocognitive research, psychometric results, clinical utility, and clinicians' practical needs. Subsequently, around 15 pages of text were devoted to an explication of evidence to justify each goal. Although not explicitly mentioned, this revision's recent normative sample removes the threat of normative obsolescence (Wasserman & Bracken, 2013).

Evolution of the Wechsler scales based on references to intelligence structure suggested by J. B. Carroll (1993a, 2003, 2012), Cattell and Horn (1978), Horn (1991), and Horn and Blankson (2012) denote a hierarchical structure with general intelligence and group factors of verbal comprehension (VC), visual spatial (VS), fluid reasoning (FR), working memory (WM), and processing speed (PS) that is consistent with what has come to be known as Cattell-Horn-Carroll (CHC; McGrew, 1997, 2005) theory. Thus, measurement of intelligence by the WISC–V continues to include narrow ability subtests (16), group factors (5), and general intelligence (Spearman, 1927).

Modifications and simplification of instructions and item phrasing were reportedly studied in children ages 4:6 to 5:11 and incorporated in the WISC–V. The number of demonstration, sample, and teaching items were increased. The number of items with time bonuses was reduced. Discontinue rules within subtests were reduced, and for most primary and secondary subtests it is now three consecutive zero-point responses. Test stimuli included in the stimulus books are attractive, in full color, and visually engaging. Materials also appear to be of high quality and likely to withstand the demands of frequent use without significant deterioration. The *WISC–V Administration and Scoring Manual,* like other recent editions, includes the crack-back binding to allow the manual to stand during administration.

Word Reasoning and Picture Completion subtests from the WISC–IV were eliminated, and Visual Puzzles and Figure Weights (present in the WAIS–IV) and Picture Span (adapted from Picture Memory in the WPPSI–IV) were added. Five "complementary scale" subtests (Naming Speed Literacy, Naming Speed Quantity, Immediate Symbol Translation, Delayed Symbol Translation, and Recognition Symbol Translation) were added but are not measures of intelligence. Subtests retained from the WISC–IV had administration, item content, and scoring changes. It was reported that all retained subtests had both low-difficulty and high-difficulty items added to achieve adequate floor and ceiling levels.

Organization and subtest administration order of the WISC–V reflects a new four-level organization. At the Full scale level, the FSIQ is composed of seven primary subtests across the five domains: Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed; if one of the FSIQ subtests is invalid or missing, a secondary subtest from within the same domain may be substituted. Only one substitution is allowed. Administration of these seven subtests should take around 50 minutes. The primary index scale level is composed of 10 WISC–V subtests (primary subtests), which are used to estimate the five WISC–V factor index scores: Verbal Comprehension Index, Visual Spatial Index, Fluid Reasoning Index, Working Memory Index, and Processing Speed Index . No substitutions are allowed for the primary index scales. Administering the 10 primary subtests should take around 65 minutes. The ancillary index level is composed of five scales that are not factorially derived—Quantitative Reasoning, Auditory Working Memory, Nonverbal, General Ability, and Cognitive Proficiency—and reflect various combinations of primary and secondary subtests. The Complementary Index level is composed of three scales—Naming Speed, Symbol Translation, and Storage and Retrieval—derived from the newly created complementary subtests: Naming Speed Literacy, Naming Speed Quantity, Immediate Symbol Translation, Delayed Symbol Translation, and Recognition Symbol Translation. Complementary subtests are not intelligence subtests and may not be substituted for primary or secondary subtests.

In prior versions of the WISC, the FSIQ was based on 10 subtests; the WISC–V FSIQ is based on seven subtests. Additionally, the subtests that comprise the FSIQ differ between the WISC–IV and WISC–V: Only six of the WISC–V FSIQ subtests were used to compute the WISC–IV FSIQ. Similar changes in the underlying composition of the WISC–III and WISC–IV were noted and generated the caution that "research findings with previous WISCs are now less generalizable to the WISC–IV" (A. S. Kaufman, Flanagan, Alfonso, & Mascolo, 2006, p. 281). That caution can now be extended to the WISC–V. Although the general intelligence construct appears to be robust to changes in subtest composition (Johnson, te Nijenhuis, & Bouchard, 2008), the resulting measured FSIQ scores may differ (Floyd, Clark, & Shadish, 2008). This difference may be especially important when FSIQ scores are applied in high-stakes situations, such as Atkins cases (Taub, 2014).

Interpretation

The *WISC–V Administration and Scoring Manual* provides detailed and annotated descriptions of the sequential procedures (with examples) of transformation of raw scores to scaled scores and scaled scores to standard scores. It also explains the methods for calculating deviations (with examples) and use of tables for statistical significance and base rates (where available). Such detail should allow clinicians ample instruction for such critical derivations.

WISC–V interpretation considerations and methods presented in the manual begin with reporting and describing performance of the individual using the standard scores that indicate how the child performed relative to same-age peers. Percentile ranks, confidence intervals based on standard errors of measurement, and qualitative descriptors of performance further describe the child's performance. These are normative (nomothetic) interpretations. The qualitative descriptors in the WISC–V have changed from the traditional Wechsler qualitative descriptors and likely will be favorably received. The new descriptors are now symmetrical in terminology ranging from extremely high to extremely low. Terms of borderline, superior, and very superior have been abandoned.

The remaining analyses and interpretations are intra-individual comparisons (comparing the child's performance on different scales) and dependent on statistical significance of score differences (alpha levels now provided for .01, .05, .10, and .15), by age group or the overall sample. Primary index score strengths and weaknesses are ipsative comparisons, and the scores can be either compared to the mean primary index score or to the FSIQ. Users select the alpha level for the comparisons, which ranges from .01 to .15. Base rates for differences in the population can be based either on the overall sample or by ability level. Subtest score strengths and weaknesses are also ipsative comparisons, and subtest scores can be compared to the mean of all 10 primary subtest scores or the mean of the seven FSIQ subtests. Users also select the alpha level for comparisons, which ranges from .01 to .15. Pairwise difference scores can be calculated for all possible combinations of the five primary index scores (10 comparisons) with statistical significance of the difference based on the user-selected alpha level (.01–.15). Pairwise differences also have population base rates based on either the overall sample or ability level. There are also five specific subtest-level pairwise comparisons examining the difference between each of the two subtest indicators of the primary index scores; statistical significance is dependent on the user-selected alpha level (.01–.15).

Ancillary index scores may also be derived and reported as standard scores and include percentile rank, confidence intervals based on standard errors of measurement, and qualitative descriptors. Like the primary index scores, ancillary index scores are normative (nomothetic) interpretations. Four ancillary index score pairwise comparisons are provided with statistical significance based on user-selected alpha (.01–.15) and also include population base rates. Six ancillary index subtest pairwise comparisons may be calculated and also utilize user-selected alpha (.01–.15) and population base rates.

Complementary index scales may also be derived as subtest standard scores, and their combinations provide for three complementary index composite scores, which include percentile rank and confidence intervals. Last, there are a host of process scores and analyses including pairwise comparisons and base rates.

Analyses for specific learning disability identification include description of ability-achievement discrepancy (AAD) analysis with a preference for using regression-based discrepancy rather than the simple difference method. Learning disability identification using the pattern of strengths and weaknesses (PSW) is also described.

Technical Qualities

Standardization

The *WISC–V Technical and Interpretive Manual* includes detailed and extensive information regarding standardization procedures and the normative sample of 2,200 children between the ages of 6 and 16 years with 100 boys and 100 girls at each age level. Raw score to scaled score conversions are reported by 3-month blocks in the *WISC–V Administration and Scoring Manual* so that approximately 67 children are included in each 3-month block, well above the minimum number of 30 to 50 suggested by researchers (Kranzler & Floyd, 2013; Zhu & Chen, 2011). Normative data were collected between April 2013 and March 2014 and stratified according to the October 2012 U.S. census data to achieve proportional representation across key demographic variables of age, sex, race/ethnicity, parent education level (a proxy for socioeconomic status), and geographic region. Additionally, a representative proportion of children with special education diagnoses (developmental delay = 0.6%; intellectual disability = 1.6%; specific learning disability = 1.7%; speech/language impairment = 1.5%; attention-deficit/hyperactivity disorder = 1.1%; gifted/talented = 1.7%) were included and accounted for around 8% to 10% of the children in each age group. Table 3.1 of the *WISC–V Technical and Interpretive Manual* presents exclusionary criteria that prevented individuals from being included in the normative sample. Tables 3.2 through 3.5 illustrate close approximation to population percentages supporting generalizability to the United States as a whole.

Primary and secondary subtest scaled scores (mean [*M*] = 10, standard deviation [*SD*] = 3, Range = 1 to 19) for each of the age groups were derived from an inferential norming procedure using raw score means, *SDs*, and skewness estimates that were examined from linear through fourth-order polynomial regressions with comparison to theoretical distributions and growth curve patterns that produced percentiles for each raw score. Smoothing (method not disclosed) eliminated minor irregularities of scaled score progression. Item gradients (e.g., the change in scaled score created by a 1-point increase in raw scores) for the primary subtests were adequate per the standards provided by Wasserman and Bracken (2013).

Standard scores (*M* = 100, *SD* = 15) are used for all composite scores (FSIQ, primary index scores, ancillary index scores, complementary index scores) and complementary subtests.

Composite scores for the five primary index scales, ancillary index scales (except Nonverbal, General Ability, and Cognitive Proficiency), and complementary index scales range from 45 to 155, and the FSIQ, Nonverbal, General Ability, and Cognitive Proficiency composite scores range from 40 to 160. Thus, the floors and ceilings for composite scores are 3.7 to 4.0 *SDs*. Given these floors, index scores should be adequate for identification of children with mild to moderate intellectual disabilities but may be inadequate for children with severe to profound intellectual disabilities (Wasserman & Bracken, 2013). These ceilings should allow identification of most candidates for gifted programs but may not be adequate for identification of exceptionally gifted children (Wasserman & Bracken, 2013). Item gradients for the primary index scales were generally within acceptable limits except at the floors of the Fluid Reasoning and Working Memory index scores.

Age-equivalent scores are also provided despite the caution in the *WISC–V Technical and Interpretive Manual* of "common misinterpretation and psychometric limitations" (p. 53) and the long-standing admonitions against using them. Given the many weaknesses of age-equivalent scores and the potential for misuse, it might be advantageous to no longer provide them to examiners.

Reliability

Reliability estimates of WISC–V scores reported in the *WISC–V Technical and Interpretive Manual* were derived using three methods: internal consistency, test-retest (stability), and interscorer agreement. Internal consistency estimates were produced by Spearman-Brown corrected split-half correlations for all subtests except Coding, Symbol search, Cancellation, Naming Speed Literacy, Naming Speed Quantity, Immediate Symbol Translation, and Delayed Symbol Translation, as these are speeded tests. For these subtests, the short-term test-retest (stability) method was used to estimate reliability. Table 4.1 in the *WISC–V Technical and Interpretive Manual* presents internal consistency reliability estimates for the WISC–V primary and secondary subtests, process scores, and composite scores by age. Average coefficients across the 11 age groups for the composite scores ranged from .88 (Processing Speed Index) to .96 (FSIQ and General Ability Index) and were higher than those obtained for subtests and process scores; a typical and expected result.

WISC–V primary and secondary subtest internal consistency estimates ranged from .81 (Symbol Search) to .94 (Figure Weights) while process scores ranged from .80 (Digit Span Backward) to .88 (Block Design Partial). Internal consistency estimates across the 11 age groups ranged from .96 to .97 for the FSIQ, from .84 to 94 for primary index scores, from .91 to .96 for ancillary index scores, and from .75 to .93 for process scores. Reliability estimates for the complementary subtests, process, and composite scores are provided in Table 4.2 of the *WISC–V Technical and Interpretive Manual.* Average coefficients across the 11 age groups ranged from .90 to .94 for composite scores and from .82 to .89 for subtests and process scores. Internal consistency reliability coefficients ≥ .90 have been recommended for high-stakes decisions (Kranzler & Floyd, 2013), which arguably include decisions about diagnosis as well as decisions about remedial or tailored instructional interventions for individual children (Stone, Ye, Zhu, & Lane, 2010). The Figure Weights, Arithmetic, and Digit Span subtests met that standard. Among the primary index scores, only the Processing Speed Index failed to meet the .90 standard.

Standard errors of measurement based on reliability coefficients from Table 4.1 are presented in Table 4.4 of the *WISC–V Technical and Interpretive Manual* and are the basis for estimated true score confidence intervals reported in the *WISC–V Administration and Scoring Manual* Tables A.2 through A.7 and in the *WISC–V Administration and Scoring Manual Supplement* Tables C.1 to C.5 and C.7 to C.9. Formulae for the estimated true score confidence interval *and* the obtained score confidence interval are provided in the *WISC–V Technical and Interpretive Manual*. Those clinicians preferring to use the obtained score confidence interval (when interest is in estimating the true score *at the time of the evaluation* and not the long-term/future estimate [Glutting, McDermott, & Stanley, 1987]) should be able to produce them from the provided formula and the detailed example (p. 62). Due to the generally high reliability estimates in Table 4.1, estimated true score and obtained score confidence intervals will likely be quite close.

Reliability estimates in Table 4.1 and standard errors of measurement in Table 4.4 should be considered best-case estimates because they do not consider other major sources of error, such as transient error, administration error, or scoring error (Hanna, Bradley, & Holen, 1981), which influence test scores in clinical assessments. Another factor that must be considered is the extent to which subtest scores reflect portions of true score variance due to a hierarchical general intelligence factor *and* variance due to specific group factors because these sources of true score variance are conflated. Later in this chapter, model-based reliability estimates will be provided to illustrate the contrast with important consequences for interpretation.

Short-term test-retest stability estimates were provided for WISC–V scores where the WISC–V was twice administered to a sample of 218 children (demographic descriptive statistics are provided in the *WISC–V Technical and Interpretive Manual* Table 4.6) with retest intervals ranging 9 to 82 days (M = 26 days). Uncorrected stability coefficients were .91 for the FSIQ, .68 (Fluid Reasoning Index) to .91 (Verbal Comprehension Index) for primary index scores; .76 (Quantitative Reasoning Index) to .89 (General Ability Index) for ancillary index scores; and .63 (Picture Concepts) to .89 (Vocabulary) for primary and secondary subtests. Corrected (for variability) stability coefficients were slightly higher. Kranzler and Floyd (2013) also recommended that short-term test-retest stability coefficients should be ≥ .90 for high-stakes decisions. Only the Vocabulary subtest along with the Verbal Comprehension Index, FSIQ, and General Ability Index met that standard. Mean differences across the retest interval were mostly small but reflected some practice effects, particularly for Processing Speed subtests (and the Processing Speed Index). Long-term stability (retest interval exceeding 1 year) estimates of the WISC–V were not expected to be included in the *WISC–V Technical and Interpretive Manual* but should be examined in the coming years. Not included in stability examinations were ipsative-based strengths and weaknesses or pairwise difference scores that are significant components of WISC–V interpretation.

Interscorer agreement was estimated by double-scoring most WISC–V subtests for all standardization sample record forms by two independent scorers. Because most WISC–V subtests have simple and objective criteria, interscorer agreement ranged from .97 to .99, which is extremely high. What is unknown is the degree to which clinicians not trained or employed by the test publisher achieve such impressive agreement when they administer and score the WISC–V because "there are innumerable sources of error in giving and scoring mental tests" (Terman, 1918, p. 33). Changes in standardized administration of cognitive tests, even something as minor as voice inflection, have been shown to influence test scores (D. Lee, Reynolds, & Willson, 2003). Likewise, examiner familiarity and examinee characteristics may impact test scores (Fuchs & Fuchs, 1986; Szarko, Brown, & Watkins, 2013). In fact, considerable evidence suggests that such positive results are improbable among clinicians. For example, a recent study revealed large examiner effects among 448 examiners who tested 2,783 children with the WISC–IV (McDermott, Watkins, & Rhoad, 2014), and there is a long history of examiner inaccuracy, especially on the verbal portions of Wechsler scales (Babad, Mann, & Mar-Hayim, 1975; Moon, Blakey, Gorsuch, &

Fantuzzo, 1991; Oakland, Lee, & Axelrod, 1975; Slate, Jones, Murray, & Coulter, 1993).

Validity

The *WISC–V Technical and Interpretive Manual* chapter on validity references *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [AP], & National Council on Measurement in Education [NCME], 1999), although the new edition of the *Standards* (AERA, APA, & NCME, 2014) preceded the WISC–V in publication and could have been used. Presentation of evidence for WISC–V validity was structured around the *Standards*, which reflect Messick's (1995) unified validity theory that prescribes evidence based on test content, response processes, internal structure, relations with other variables, and consequences of testing.

Validity evidence based on test content is a nonempirical approach. In the WISC–V, test content was reportedly informed through review of literature and item and subtest review by experts and advisory panel members (specialists in child psychology, neuropsychology, and/or learning disabilities), a list of which is provided in the *WISC–V Technical and Interpretive Manual*. Discussion of evidence based on response processes in the manual highlighted both retention of subtests from previous versions for which such evidence was claimed as well as interviewing children regarding their rationale for selecting responses or problem-solving strategies used to complete various items. Modifications to item content and instructions were noted as a result of these procedures.

Evidence based on internal structure is one of the most important aspects for construct validity in order to understand relations between subtests and their correspondence to theoretical and latent constructs. Two approaches to examination of the internal structure are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is the method of extracting latent factors from the correlation matrix of the indicators based on their convergent and divergent relationships and allows "the data to speak for themselves" (J. B. Carroll, 1995, p. 436). CFA is a method of proposing various theoretical measurement models and empirically testing which model (or models) best fits the data. EFA and CFA are considered complementary procedures, each answering somewhat different questions, and greater confidence in the latent factor structure is achieved when EFA and CFA are in agreement (Gorsuch, 1983). Further, J. B. Carroll (1995) and Reise (2012) noted that EFA procedures are particularly useful in suggesting possible models to be tested in CFA.

The *WISC–V Technical and Interpretive Manual* describes data supporting a priori hypotheses regarding subtest correlations reflecting convergent and divergent (discriminant) validity as evidence of construct validity within the internal structure section. The average correlations (Fisher transformations) and descriptive statistics for the total normative sample are presented in Table 5.1 of the manual. Several pages in the manual are devoted to description of how various subtests within the five primary factor indexes are moderately to highly correlated with each other, suggesting construct validity (convergent validity). Descriptions of lower correlations between subtests from different primary factors also illustrate construct validity (discriminant validity). However, regardless of the a priori hypotheses regarding these relationships and their differential correlations, full understanding of the complex relationships between all the subtests at the same time requires multivariate methods, such as EFA and CFA.

CFA reported in the *WISC–V Technical and Interpretive Manual* includes specification of numerous models starting with a one-factor model. All other models were higher order with a general intelligence factor indirectly influencing subtests via full mediation through two through five first-order factors. All CFA models are illustrated with subtest assignments to latent factors in Table 5.3 of the manual. Contemporary fit statistics were described and their meaning explained.

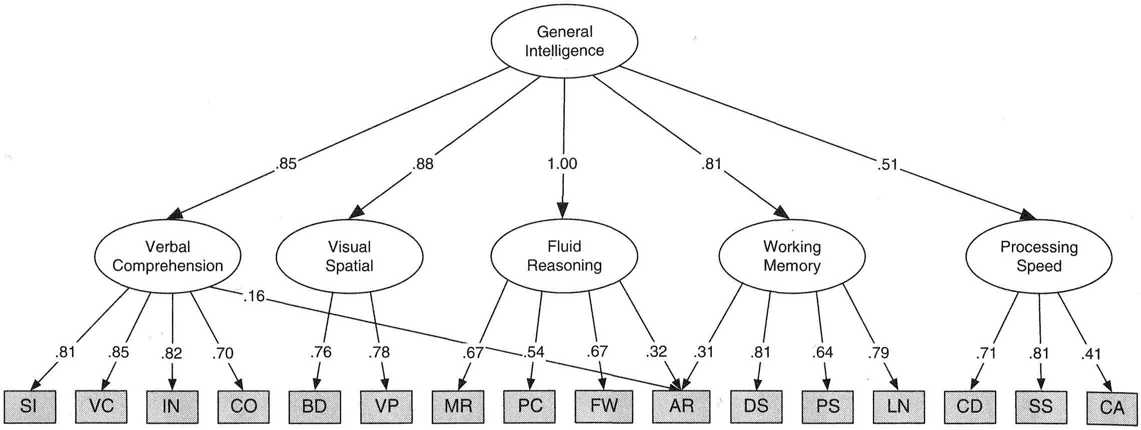

Figure 20.1 Higher-Order Measurement Model Adapted from Figure 5.1 (Wechsler, 2014), with standardized coefficients for WISC–V normative sample (N = 2,200) , ages 6-16, for 16 primary and secondary subtests.
Note: SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, PC = Picture Concepts, FW = Figure Weights, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, CD = Coding, SS = Symbol Search, CA = Cancellation.

The standardized measurement model for the preferred five-factor higher-order (hierarchical) model for WISC–V primary and secondary subtests for the total normative sample is presented in Figure 5.1 of the WISC–V *Technical and Interpretive Manual* and adapted here as Figure 20.1. This "best-fitting" model includes a higher-order general intelligence dimension with five first-order factors (Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, Processing Speed). The 16 subtest indicators are uniquely associated with one latent first-order factor except for Arithmetic, which was cross-loaded on Verbal Comprehension, Fluid Reasoning and Working Memory. This preferred measurement model includes a standardized path coefficient of 1.00 between the higher-order general intelligence factor and the Fluid Reasoning factor, which indicates that they are redundant. This final model was also reported to fit five different age groupings (6–7, 8–9, 10–11, 12–13, 14–16) equally well.

Finally, Figure 5.2 in the *WISC–V Technical and Interpretive Manual* illustrates the five-factor higher-order (hierarchical) model as applied to only the 10 primary subtests. In this model there are no cross-loadings, and each first-order factor has two subtest indicators. Like the 16-subtest CFA, the standardized path coefficient of .99 between the higher-order general intelligence factor and the fluid reasoning factor indicates redundant dimensions.

Regardless of factor structure suggested by either EFA or CFA, models must be evaluated by comparisons to external criteria. Evidence based on relations with other variables presents numerous comparisons of the WISC–V with other measures of intelligence (WISC–IV [n = 242], WPPSI–IV [n = 105], WAIS–IV [n = 112], Kaufman Assessment Battery for Children–Second Edition [KABC–II; A. S. Kaufman & Kaufman, 2004; n = 89]), measures of academic achievement (Kaufman Test of Educational Achievement–Third Edition [KTEA–3; A. S. Kaufman & Kaufman, 2004; n = 207], Wechsler Individual Achievement Test–Third Edition [WIAT–III; Pearson, 2009a; n = 211]), a measure of adaptive behavior (Vineland Adaptive Behavior Scales–Second Edition [Vineland–II; Sparrow, Cicchetti, & Balla, 2005; n = 61]), and a measure of child behavior (Behavior Assessment System for Children–Second Edition Parent Rating Scales [BASC–2 PRS; Reynolds & Kamphaus, 2004; *n* = 2,302]) using nonclinical samples. With respect to comparisons of the WISC–V to other measures of intelligence, there appears to be good correspondence with moderate to high correlations between similar composite scores. The highest uncorrected correlations were observed between the WISC–V FSIQ and the WISC–IV FSIQ (.81), WPPSI–IV FSIQ (.74), WAIS–IV FSIQ (.84), and KABC–II MPI (.77). Comparisons between the WISC–V FSIQ and academic achievement tests (KTEA–3 and WIAT–III) produced zero-order Pearson correlations with achievement composite scores that were typically in the .60s and .70s and consistent with those reported by Naglieri and Bornstein (2003).

Correlations between the WISC–V and the Vineland–II were largely low to near zero, indicating divergent validity because the WISC–V and Vineland measure different psychological constructs (intelligence versus adaptive behavior). Comparisons of the WISC–V with the BASC–2 PRS were somewhat limited given that Resiliency, Conduct Problems, Executive Functioning, and Attention Problems were the only BASC scales reported. Like the Vineland–II, correlations between the WISC–V and BASC–2 scores on these four scales were low to near zero and supportive of divergent validity. This finding also was expected, given the different psychological constructs the WISC–V and BASC–2 measure. Canivez, Neitzel, and Martin (2005) found similar results with the Wechsler Intelligence Scale for Children–Third Edition (WISC–III; Wechsler, 1991) in comparisons with the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993).

WISC–V performance among 13 special groups is summarized in the *WISC–V Technical and Interpretive Manual* (pp. 112–147). Groups included intellectually gifted, intellectual disability–mild severity, intellectual disability–moderate severity, borderline intellectual functioning, specific learning disorder–reading, specific learning disorder–reading and written expression, specific learning disorder–mathematics, attention-deficit/hyperactivity disorder, disruptive behavior, traumatic brain injury, English language learner, autism spectrum disorder with language impairment, and autism spectrum disorder without language impairment. Most are small groups of 20 to 30 individuals who were then compared to a randomly selected and demographically matched standardization subsample. Acknowledged limitations included small sample sizes, nonrandom selection of special group participants, data collected by independent examiners and researchers, and special group participants had predetermined classifications that might have been based on different selection criteria. For these reasons, these results must be considered preliminary and require replication with well-designed and controlled studies.

Generally, results indicated various WISC–V scores that were significantly different between the special group and the control participants and in expected directions. For example, individuals with intellectual giftedness scored higher than the control group, but individuals with specific disabilities scored lower than the control group. Such distinct group differences provide some preliminary evidence for construct validity.

The *WISC–V Technical and Interpretive Manual* noted that results from these studies "demonstrate the differential sensitivity of the WISC–V to specific and general cognitive deficits exhibited by children commonly evaluated in clinical settings" (p. 112).

The manual further asserted that this information about group mean differences "provides evidence for the clinical utility and discriminant validity of the WISC–V subtests and composites" (p. 112). Unfortunately these conclusions are insufficiently supported by comparisons of distinct groups, which provide necessary *but not sufficient* evidence for clinical utility. Differences between groups (discriminative validity) do not automatically translate into accurate decisions about individuals (clinical utility).

Rather, methods and analyses examining conditional probabilities of diagnostic efficiency statistics are required for accurate clinical (individual) decisions (Kessell & Zimmerman, 1993; Swets, 1996; Treat & Vicken, 2012; Wasserman & Bracken, 2013). It has long been known that the base rates of clinical disorders, cut scores used for individual decisions, and the like are all vital for determining clinical utility (Meehl & Rosen, 1955). The distinction between classical validity and clinical utility has been repeatedly demonstrated with Wechsler scores (Devena & Watkins, 2012; Watkins, 2005; Watkins, Glutting, & Youngstrom, 2005); its absence in the WISC–V *Technical and Interpretive Manual* (other than a somewhat confusing presentation under the rubric of consequential validity) is disappointing.

COMMENTARY

As might have been expected, the foreword of the *WISC–V Technical and Interpretive Manual* was enthusiastically positive about the WISC–V. Such unbridled enthusiasm without regard to the psychometric limitations and past failures when numerous subtest and composite score comparisons were put to the empirical test is perhaps premature. Although there are a number of positive changes and elements in the WISC–V, there continue to remain glaring omissions previously pointed out in reviews of the WAIS–IV (Canivez, 2010), WPPSI–IV (Canivez, 2014b), and WISC–IV (Canivez & Kush, 2013) that must be examined.

Failure to Provide Results from EFA

The *WISC–V Technical and Interpretive Manual* explicitly preferred CFA over EFA methods rather than taking advantage of both methods. EFA and CFA are considered complementary procedures, each answering somewhat different questions, and greater confidence in the latent factor structure is achieved when EFA and CFA are in agreement (Gorsuch, 1983). Further, J. B. Carroll (1995) and Reise (2012) both noted that EFA procedures are particularly useful in suggesting possible models to be tested in CFA. In fact, J. B. Carroll (1998) suggested that "CFA should derive its initial hypotheses from EFA results, rather than starting from scratch or from a priori hypotheses ... [and] CFA analyses should be done to check my EFA analyses" (p. 8).

The fact that two WISC–IV subtests were deleted (Word Reasoning and Picture Completion), three new subtests were added (Visual Puzzles, Figure Weights, and Picture Span), and items in all WISC–V subtests were new or revised suggests that relationships among retained and new subtests might result in associations and latent structure unanticipated by a priori conceptualizations (Strauss, Spreen, & Hunter, 2000). The absence of EFA results is most disappointing, given prior criticism of their absence in other Wechsler manuals (Canivez, 2010, 2014b). Because of this lacuna in the WISC–V *Technical and Interpretive Manual*, EFA results for the total WISC–V normative sample are included later in this chapter.

CFA Methods

Figure 5.1 and 5.2 in the *WISC–V Technical and Administration* mislabel the latent construct of general intelligence. What is labeled "Full Scale" in these figures should be General Intelligence, which is the name of the latent construct. The Full Scale IQ is an observed variable and an estimate of the latent construct general intelligence. Also, there is no FSIQ utilizing all 16 subtests; in fact, only seven WISC–V subtests are used to produce the FSIQ.

Unfortunately, reports of the CFA analyses in the *WISC–V Technical and Administration Manual* were not adequately informative (Boomsma, 2000). For example, there was no indication of the method used to scale the models for identification. A brief footnote to Table 5.4 indicated that weighted least squares (WLS) estimation was applied. However, "use of an estimation method other than ML [maximum likelihood] requires explicit justification" (Kline, 2011, p. 154), and no explanation was provided for the choice of WLS. WLS typically is used for categorical or nonnormal data and may not produce chi-square values nor approximate fit indices equivalent to those produced by ML estimation (Yuan & Chan, 2005). Further, WLS requires very large sample sizes (Hu, Bentler, & Kano, 1992) and may be more sensitive to model misspecification than ML estimation (Olsson, Foss, Troye, & Howell, 2000). For these and other reasons, Brown (2006) concluded that WLS is "not a good estimator choice" (p. 388). We were unable to replicate these analyses because raw data are needed for WLS estimation.

Figure 5.1 in the *WISC–V Technical and Interpretive Manual* (modified and presented as Figure 20.1 in this chapter) presents the final and publisher-preferred standardized measurement model for the hierarchical five-factor model for the 16 primary and secondary subtests with the total normative sample (ages 6–16 years). This complex model (due to cross-loadings included for the Arithmetic subtest) is problematic for several reasons. First, the standardized path of 1.00 between the latent general intelligence (Full Scale) factor and Fluid Reasoning factor means that fluid reasoning is isomorphic with the hierarchical *g* factor. This is a major threat to discriminant validity and an indication that the WISC–V may be overfactored when five group factors are included.

The relationship between fluid reasoning and general intelligence is a long-standing puzzle, and there are practical and theoretical issues remaining to be resolved (M. R. Reynolds, Keith, Flanagan, & Alfonso, 2013). Both Vernon (1965) and J. B. Carroll (2003) questioned whether general and fluid factors could be distinguished. However, Golay, Reverte, Rossier, Favez, and Lecerf (2013) used Bayesian structural equation modeling (BSEM) rather than traditional CFA methods with the French WISC–IV and found that the fluid reasoning factor *did not* load at unity on the general intelligence factor when allowing small nonzero subtest cross-loadings rather than fixing them to zero as is typically done in frequentist CFA. The relationship was still very high but less likely to be identical. Whether this BSEM result is unique to the French WISC–IV or is also observed in other Wechsler tests such as the WISC–V should be examined.

The *WISC–V Technical and Interpretive Manual* remarked on the propensity of the chi-square test to identify trivial differences with large samples but subsequently used chi-square difference tests of nested models to identify the preferred five-factor model. However, the same sensitivity to large samples is true for chi-square difference tests (Millsap, 2007), suggesting that the model differences reported in the manual might be trivial. For example, Table 5.4 in the manual reveals that the difference between models 4a and 5a was statistically significant but those two models exhibited identical comparative fit index (CFI) and root mean squared error of approximation (RMSEA) values. Likewise, the preferred five-factor higher-order model was significantly different from other five-factor models, but all exhibited identical CFI and RMSEA values (e.g., .98 and .04, respectively). Cheung and Rensvold (2002) demonstrated, in the context of factorial invariance, that practical differences independent of sample size and model complexity could be identified by ΔCFI > .01.

Figure 5.2 in the *WISC–V Technical and Interpretive Manual* presents the final and publisher-preferred standardized measurement model for the hierarchical five-factor model for the 10 primary subtests with the total normative sample (ages 6–16 years). Although this model does not include cross-loadings and thus represents simple structure (a desired feature of test structure), it is still problematic because the path from the latent general intelligence factor and

Fluid Reasoning factor is .99, suggesting that fluid reasoning is isomorphic with general intelligence. This likely indicates overfactoring when including five first-order factors. Again, it is possible that this result is an artifact of CFA methods and fixing cross-loadings to zero when they are actually small nonzero values (Golay et al., 2013). BSEM methods will help to determine this phenomenon in the WISC–V; however, independent analyses using BSEM requires access to the standardization raw data and cannot be based on summary data from the manual.

Variance Decomposition

Another problem is that the publisher did not provide decomposed variance estimates to illustrate how much subtest variance is due to the hierarchical *g* factor and how much is due to the specific group factors. This is a glaring omission because clinicians and researchers are unable to judge the adequacy of the group factors (Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed) based on how much unique variance they capture when purged of the effects of general intelligence. Because cross-loadings are included in the preferred measurement model for all 16 WISC–V subtests (Figure 5.1 in the *WISC–V Technical and Interpretive Manual*), it is not easy to use the standardized path coefficients to decompose the variance estimates. These problems were pointed out in reviews of the WAIS–IV (Canivez, 2010) and WPPSI–IV (Canivez, 2014b) as well as in a commentary regarding the WISC–IV and WAIS–IV (Canivez & Kush, 2013). The publisher was admonished to include such estimates and information to no avail.

Fortunately, the measurement model presented in Figure 5.2 of the manual exhibits simple structure so it is relatively straightforward to decompose the variance estimates from the standardized loadings. Table 20.1 in this chapter presents the subtest and factor variance estimates based on Figure 5.2. Table 20.1 reveals that most subtest variance is associated with the general intelligence factor and substantially smaller portions of subtest variance are uniquely related to the first-order factors, except in the case of Processing Speed, which includes tasks little related to general intelligence.

Further inspection of Table 20.1 shows that the higher-order *g* factor accounted for 34.8% of the total variance and 67.6% of the common variance. Thus, WISC–V measurement is dominated by the higher-order general intelligence factor. At the first-order level, Verbal Comprehension accounted for an additional 4.0% of the total variance and 7.0% of the common variance; Visual Spatial accounted for an additional 2.3% of the total variance and 4.0% of the common variance; Fluid Reasoning accounted for an additional 0.2% of the total variance and 0.4% of the common variance; Working Memory accounted for an additional 3.2% of the total variance and 5.6% of the common variance; and Processing Speed accounted for an additional 8.7% of the total variance and 15.4% of the common variance. Given the extremely low variance attributable to Fluid Reasoning, there seems to be little justification for its inclusion.

Model-Based Reliability

It has long been known that classical estimates of reliability are biased (Raykov, 1997) and model-based estimates, such as omega-hierarchical (ω_h) and omega-subscale (ω_s), have been recommended as superior replacements (Gignac & Watkins, 2013). In a review of the WPPSI–IV, Canivez (2014b) noted that model-based reliability coefficients should have been included to allow clinicians and researchers to judge the merits and interpretability of the claimed latent factors. Based on the decomposed factor loadings in Table 20.1, ω_h and ω_s coefficients were computed to estimate latent factor reliabilities. Omega coefficients should exceed .50 at a minimum, but .75 would be preferred (Reise, 2012; Reise, Bonifay, &

Table 20.1 Decomposed Sources of Variance in the WISC–V 10 Primary Subtests for the Total Normative Sample (N = 2,200) According to the Higher-Order Model (Figure 5.2, *WISC–V Technical and Interpretive Manual*)

	General		Verbal Comprehension		Visual Spatial		Fluid Reasoning		Working Memory		Processing Speed			
Subtest	b	S^2	b	S^2	b	S^2	b	S^2	b	S^2	b	S^2	h^2	u^2
Similarities	.689	.474	0.442	0.196									.670	.330
Vocabulary	.697	.486	0.452	0.204									.690	.310
Block Design	.684	.468			.335	.112							.580	.420
Visual Puzzles	.702	.493			.342	.117							.610	.390
Matrix Reasoning	.673	.453					.082	.007					.460	.540
Figure Weights	.673	.453					.130	.017					.470	.530
Digit Span	.647	.419							.437	.191			.610	.390
Picture Span	.540	.291							.359	.129			.420	.580
Coding	.357	.127									.602	.363	.490	.510
Symbol Search	.423	.179									.715	.511	.690	.310
Total Variance		.348		.040		.023		.002		.032		.087		
Common Variance		.676		.070		.040		.004		.056		.154		
	$\omega_h =$	.823	$\omega_s =$	.238	$\omega_s =$	.144	$\omega_s =$	.015	$\omega_s =$	.210	$\omega_s =$	.548		

Note: b = standardized loading of subtest on factor, S^2 = variance explained in the subtest, h^2 = communality, u^2 = uniqueness, ω_h = omega hierarchical, ω_s = omega subscale

Haviland, 2013). The ω_h coefficient for general intelligence (.823) was high and sufficient for scale interpretation; however, the ω_s coefficients for the five WISC–V first-order factors (Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed) were considerably lower, ranging from .015 (Fluid Reasoning) to .548 (Processing Speed). Thus, the WISC–V first-order factors, with the possible exception of Processing Speed, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise et al., 2013).

Continuing Problems with the Arithmetic Subtest

Examination of Figure 5.1 in the *WISC–V Technical and Interpretive Manual* illustrates the continuing difficulties with the Arithmetic subtest with its cross-loading with Verbal Comprehension, Fluid Reasoning, and Working Memory. Canivez and Kush (2013) pointed out problems with Arithmetic in the WISC–IV and WAIS–IV due to cross-loadings modeled by Weiss, Keith, Zhu, and Chen (2013a) and by Weiss, Keith, Zhu, and Chen (2013b). Arithmetic had its origin in the Wechsler scales as a verbal subtest, but beginning with the Wechsler Intelligence Scale for Children–Revised (WISC–R; Wechsler, 1974a) factor-analytic studies found Arithmetic, Digit Span, and Coding formed a small third factor (so-called Freedom from Distractibility). The attempt to strengthen that small third factor by adding a new subtest (Symbol Search) in the WISC–III produced the opposite effect by pulling Coding away to form a new fourth factor, so-called

Processing Speed, a name both Keith (1997) and Kranzler (1997) questioned. This left Arithmetic and Digit Span to measure the small third factor renamed Working Memory. Subsequent analyses of national standardization samples of the WISC–IV found that Arithmetic loaded on a memory factor or a fluid reasoning factor or both (Cornoldi, Orsini, Cianci, Giofrè, & Pezzuti, 2013; Fina, Sánchez-Escobedo, & Hollingworth, 2012; Golay et al., 2013; Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Weiss et al., 2013b). In fact, Arithmetic may be more of a quantitative reasoning task, as suggested by the CHC conceptualization, but there are no other quantitative reasoning tasks with which it can associate. That supposition was corroborated by a study that found that Arithmetic migrated to the Quantitative Reasoning factor when marker tests of quantitative reasoning and memory were included with subtests from the WISC–III (Watkins & Ravert, 2013). It might be time to abandon Arithmetic or provide more tasks that measure quantitative reasoning to adequately measure that broad ability.

Incremental Validity Considerations

Zero-order Pearson correlations between the WISC–V subtests, primary index scores, and ancillary index scores with the KTEA–3 and WIAT–III subtest and composite scores reported in the *WISC–V Technical and Interpretive Manual* do not account for the hierarchical nature of the WISC–V and resulting complex associations with academic achievement. As illustrated previously, WISC–V subtests measure both general intelligence variance *and* some group ability variance, but zero-order Pearson correlations between primary index scores or ancillary index scores with KTEA–3) or WIAT–III) scores conflate the general intelligence and specific group ability variance. Examination of incremental validity of primary index or ancillary index scores *beyond* that of the FSIQ (Haynes & Lench, 2003; Hunsley, 2003; Hunsley & Meyer, 2003) is necessary because the WISC–V is interpreted across multiple levels and scores and primary and ancillary index scores conflate general and group factor variance.

Canivez (2010, 2014b) argued in reviews of the WAIS–IV and WPPSI–IV that hierarchical multiple regression analyses should have been included in the respective technical manuals but such analyses are absent from the *WISC–V Technical and Interpretive Manual.* Studies applying hierarchical multiple regression analyses have supported the dominance of the FSIQ in accounting for academic achievement variance and substantially less (and often trivial amounts) of achievement variance attributable to the factor index scores (e.g., Canivez, 2013a; Canivez et al., 2014; Freberg, Vandiver, Watkins, & Canivez, 2008; Glutting, Watkins, Konold, & McDermott, 2006; Glutting, Youngstrom, Ward, Ward, & Hale, 1997; J. J. Ryan, Kreiner, & Burton, 2002; Watkins, Glutting, & Lei, 2007). Interestingly, similar results have been found for the prediction of job training and work performance of adults (Ree, Earles, & Teachout, 1994). It may be that these limited portions of achievement test score variance accounted for by first-order factor index scores is related to the generally smaller portions of subtest variance apportioned to the first-order factor scores identified through hierarchical EFA and CFA.

Measurement Bias

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) describe three ways that measurement bias might make test scores unfair for subgroups of the population. First, differential item functioning (DIF) could result when "equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership" (p. 51). Second, predictive bias could be exhibited if group membership influences the prediction of a criterion. Finally, structural bias could result if the construct being measured has different meanings dependent on group membership. Test Standard 3.0 calls for test developers and users to analyze test scores to ensure the absence of item, predictive, and structural bias for relevant subgroups. Although few details were provided, DIF within the WISC–V was analyzed and dismissed. Predictive bias and structural bias were not addressed. Methods of evaluating these types of measurement bias are well known and should have been applied (C. R. Reynolds & Ramsay, 2003). It would also have been desirable to have provided descriptive statistics for WISC–V scores disaggregated by race/ethnicity and parent education level so that users would have more information regarding score variations across these groups. Similarly, reliability estimates across race/ethnicity and sex groups would have been useful (C. R. Reynolds & Milam, 2012).

Selective Reporting and Review of Scientific Literature

There is a rather selective reporting of empirical literature including omission of contradictory evidence, reliance on studies tangential to the issue at hand, dependence on methodologically flawed studies, failure to specify the limitations of the cited research, and focusing on inessential aspects of the cited research. These practices are most egregious in Chapter 6 of the *WISC–V Technical and Interpretive Manual*, which is devoted to interpretation of WISC–V scores. Similar criticisms were made about the evidence presented in the WISC–IV manual. For example, Braden and Niebling (2012) found that "extensive discussion of how to identify intraindividual strengths and weaknesses by using Index, subtest, and within-subtest responses does not include discussion of contradictory findings available in the literature" (p. 744) and that "no evidence is cited or provided in direct support of these claims" (p. 745) for basing educational and clinical interventions on an analysis of cognitive strengths and weaknesses. These criticisms were also made in reviews of the WAIS–IV (Canivez, 2010) and WPPSI–IV (Canivez, 2014b).

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) demand that "test users should be provided with clear explanation of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations" (p. 102) and that test documentation should disclose the "validity of recommended [score] interpretations" (p. 126). Further, "when interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided" (p. 27). Given that focus, the content of Chapter 6 of the manual should meet the evidential requirements of the *Standards*.

In the interest of space, only four major examples are presented to illustrate our assertion that Chapter 6 fails to meet the *Standards*. First, the manual asserts that the differences that occur between WISC–V scores from a single WISC–V administration are important considerations in interpreting a child's performance. This assertion was followed by a presentation on intersubtest scatter as well as pairwise comparisons of subtest and index scores. No evidence was presented to support the validity of these score interpretations. However, previous research often has revealed critical flaws in such ipsative measurement methods (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Likewise, previous research has clearly shown that subtest scatter and other subtest comparisons exhibit little to no diagnostic utility (Kramer, Henning-Stout, Ullman, & Schellenberg, 1987; Watkins, 1999) and are not stable across time (Borsuk, Watkins, & Canivez, 2006; Watkins & Canivez, 2004). Given this evidence, Hunsley and Mash (2007) concluded that "an evidence-based approach to the assessment of intelligence would indicate that nothing is to be gained, and much is to be potentially lost, by considering subtest profiles" (p. 32). Similar opinions have been expressed by other assessment experts (e.g., Braden & Shaw, 2009;

Kamphaus, Reynolds, & Vogel, 2009; Kranzler & Floyd, 2013; Lilienfeld, Ammirati, & David, 2012; McDermott, Fantuzzo, & Glutting, 1990; C. R. Reynolds & Milam, 2012). Nevertheless, no limitations or cautions were provided in the *WISC–V Technical and Interpretive Manual* and no contradictory studies were reported.

Second, the *WISC–V Technical and Interpretive Manual* claims that "there is strong psychometric and clinical support for interpreting the WISC–V index scores as reliable and valid measures of the primary cognitive constructs they intend to represent" (p. 149) and concludes that analysis of primary index scores (e.g., Verbal Comprehension Index, Visual Spatial Index, etc.) "is recommended as the principal level of clinical interpretation" (p. 157). However, "no score yielded by intelligence tests (or any other measurement instrument) is a pure measure of the construct it targets" (Floyd, Reynolds, Farmer, & Kranzler, 2013, p. 399). Rather, the WISC–V index scores are "contaminated" by: (a) systematic variance of the general factor, (b) random error variance, and (c) systematic specific variance of each subtest that is not shared with any other subtest.

Sources of variance for the WISC–V have been itemized in Table I.10 in Appendix I in the downloadable resources: www.wiley.com/go/itwiscv and clearly show that general intelligence accounts for the bulk of the variance of the index scores. Using a primary index score as the principal level of interpretation ignores the contributions of general intelligence, error, and specific variance. Additionally, the factor index scores have demonstrated poor temporal stability (Watkins & Canivez, 2004; Watkins & Smith, 2013) and little incremental predictive validity of academic achievement (Canivez et al., 2014; Glutting et al., 2006; Parkin & Beaujean, 2012). At present, "there is very little evidence to suggest that the subtest or composite score differences on intelligence tests can be used to improve decisions about individuals" (Kranzler & Floyd, 2013, p. 86). This conclusion was affirmed by Schneider (2013b), who stated, "there is little evidence that clinicians are able to measure the non-*g* portions of group factors with precision, make valid inferences about them, and use this knowledge to help individuals" (p. 187). However, the *WISC–V Technical and Interpretive Manual* provides no cautions and does not present contrary opinions regarding the appropriate level of interpretation.

The third major example in Chapter 6 of the manual is that many of the interpretive suggestions are based on the "pattern of scores on the composites and subtests" (p. 156). WISC–V scores are assumed to be "reliable and valid measures of the primary cognitive constructs they intend to represent" (p. 149). Thus, the score profile identifies cognitive strengths and weaknesses that are assumed to underlie learning problems. Logically, interventions could then be individualized to match the specific cognitive strengths and weaknesses of each examinee. This approach exemplifies an aptitude-treatment interaction (ATI) model where it is assumed that learners will differentially respond to interventions that capitalize on their cognitive strengths (Cronbach & Snow, 1977).

The *WISC–V Technical and Interpretive Manual* asserts that profiles are only hypotheses that must be "corroborated or refuted by other evaluation results, background information, or direct behavioral observations" (p. 157). However, the two references cited to support this statement are books that do not include primary research results. In essence, they are sources with similar opinions that do not contribute experimental evidence. Further, what evidence, exactly, is needed to corroborate or refute any particular hypothesis? This uncertainty leaves considerable subjectivity in interpretation of cognitive profiles, which has been shown to increase error rates (Aspel, Willis, & Faust, 1998).

The admonition to confirm or refute WISC–V score interpretations with other information also ignores the likelihood of reducing overall validity if multiple measures include some with low reliability or validity. As noted by Faust (2007), "prediction is often maximized by identifying a relatively small set of the most valid and minimally redundant predictors, rather than trying to integrate many variables" (p. 35). Likewise, the admonition to simultaneously consider a large amount of complex information implies that clinicians are able to combine multiple sources of information holistically and arrive at accurate judgments, which is unlikely (Faust, 1989; Ruscio & Stern, 2006). Thus, the belief that an ill-defined analysis by clinicians of an unspecified set of data will accurately adjudicate hypotheses ignores what is known about clinical judgment (Lilienfeld et al., 2012; Watkins, 2009), especially the power of confirmation bias (Nickerson, 1998).

Early analyses of aptitude-treatment interactions in education were negative (Kavale & Mattson, 1983). By 1990, Glutting and McDermott had concluded that "traditional IQ tests have not met the challenge of providing effective aptitude-treatment interactions (ATIs) for evaluating how children best learn, or for determining how a particular child's style of learning is different from the styles manifested by other children" (p. 296). Although ATIs are clinically popular, the evidence against them has accumulated over the past several decades. For example, Good, Vollmer, Creek, Katz, and Chowdhri (1993) conducted a study with the Kaufman Assessment Battery for Children (A. S. Kaufman & Kaufman, 1983) and found no benefit from matching instructional approaches to cognitive strengths. Other reviews of the literature found no support for ATIs (Canivez, 2013b; Gresham & Witt, 1997; Kamphaus, Winsor, Rowe, & Kim, 2012; Macmann & Barnett, 1997; McDermott et al., 1990; Reschly, 1997; Watkins, 2003, 2009; Watkins et al., 2005). One review of ATIs concluded that "the evidence showing cognitive assessment is useful for matching interventions under an ATI model is lacking, and in some cases, it is demonstrably negative.... [N]or do cognitive test developers provide evidence of ATI outcomes to support their claims that tests are valuable for this purpose" (Braden & Shaw, 2009, p. 107). None of this contradictory evidence is mentioned in the *WISC–V Technical and Interpretive Manual.*

The final example from Chapter 6 of the manual is the presentation of PSW methods and the caution (that appears twice in the manual) that cognitive profiles are "not intended to diagnose specific disorders" (p. 149). However, there is an unmistakable implication that patterns of cognitive strengths and weaknesses can and should be used in the diagnosis of learning disabilities. For example, the final portion of the chapter is devoted to explication of a PSW model based on other PSW methods that are explicitly designed for the identification of learning disabilities (Flanagan, Ortiz, & Alfonso, 2007; Hale et al., 2008). The *WISC–V Technical and Interpretive Manual* claims that this PSW "model is a legally acceptable and clinically sound approach for helping practitioners identify SLDs [specific learning disabilities] and develop intervention plans based on a child's strengths and weaknesses. Use of this type of model is good clinical practice and adds weight to an eligibility or diagnostic decision" (p. 183). Additionally, the Pearson clinical website lists "identifying and diagnosing learning disabilities/disorders" as one application of the WISC–V (http://www.pearsonclinical.com/psychology/products/100000771/wechsler-intelligence-scale-for-childrensupsupfifth-edition--WISC-V.html#tab-details). How do such claims *not encourage* clinicians to use cognitive profiles for diagnostic purposes?

The assertion that PSW approaches have a sound legal foundation is also dubious. Zirkel (2014) noted that use of a PSW model relies on an inaccurate interpretation of Individuals with Disabilities Education Act regulations and that "the legally required evaluation does not necessarily include—per the aforementioned OSEP [Office of Special Education Programs] interpretation—an assessment of psychological or cognitive processing" (Zirkel, 2013, p. 95). Why the publisher includes legal advice, let alone *questionable* legal advice, in the *WISC–V Technical and Interpretive Manual* is a mystery (C. R. Reynolds & Milam, 2012).

The claim in the *WISC–V Technical and Interpretive Manual* that PSW models are "research-based" (p. 183) is debatable. There are three major models for using cognitive strengths and weaknesses to assist in the identification of children with learning disabilities (Flanagan et al., 2007; Hale et al., 2008; Naglieri & Das, 1997). The accuracy of those models was evaluated in a simulation study that found that all three failed to identify a large number of positive cases and falsely identified an even larger number of negative cases. Theoretically, these results suggest that an ATI paradigm would be iatrogenic because the misidentified children would not receive treatments matched to their true ability profiles (Stuebing, Fletcher, Branum-Martin, & Francis, 2012).

Subsequently, the accuracy of two of those PSW models (Flanagan et al., 2007; Hale et al., 2008) was evaluated for adolescents with a history of failure to respond to academic interventions (Miciak, Fletcher, Vaughn, Stuebing, & Tolar, 2014). Results revealed that there was poor agreement between the two models in identifying children with learning disabilities (kappa of –.05 to .31), and there was no pattern of academic skills that distinguished the children identified by these models. Based on these results, the authors concluded that "until empirical research provides more evidence for the validity, reliability, and utility of PSW methods, resources may be better allocated toward directly assessing important academic skills" (p. 35).

The only supportive evidence for PSW methods presented in the *WISC–V Technical and Interpretive Manual* was by authors of PSW models, although five specific references "that document empirically proven links between cognitive processes and achievement domains" (p. 183) were provided. However, all five sources were authored by developers of PSW models, and little experimental evidence was provided in any of them. To the contrary, a quantitative review of the evidence of the treatment validity of instruction based on putative cognitive strengths and weaknesses found that "a minority of reviewed studies supported the efficacy of cognitive interventions; fewer still when the cognitive component was not paired with an academic intervention" (Kearns & Fuchs, 2013, p. 285). Considered in conjunction with the ATI evidence, it appears that, when put to the test, PSW approaches fail in both identification and intervention. None of this antithetical evidence is provided in the *WISC–V Technical and Interpretive Manual*.

WISC–V Review Summary

There are many positive aspects of the WISC–V. Specifically, the WISC–V includes a large, demographically representative normative sample that allows generalizability of individual WISC–V performance to the U.S. population at large. Changes to instructions, subtest discontinue (ceiling) rules, and attractive, well-constructed materials are also major advantages. In particular improvements in instructions were noted for Block Design, Picture Concepts, the Working Memory subtests, and the Processing Speed subtests. Inclusion of subtests, such as Visual Puzzles and Figure Weights, two subtests that appear better indicators of reasoning abilities, is also quite positive. Additionally, the publisher is commended for providing correlation matrices and descriptive statistics so that independent researchers can study some aspects of the WISC–V.

As described, there are also problems with the WISC–V. Many of these problems would be ameliorated if there was greater adherence to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Additionally, a more objective, scientific approach to the presentation of evidence regarding the WISC–V should be cultivated. Clinicians are ultimately responsible for use of the WISC–V, and they must be given complete, accurate, and objective information on which to base their judgments. It must be noted that many of the problems discussed in this chapter have been reported in prior test reviews or articles. Failure to acknowledge or deal with them is a serious mistake that will not be in the long-term interest of the publisher or clinicians.

INDEPENDENT ANALYSES

Given our criticism of the structural validity analyses reported in the *WISC–V Technical and Interpretive Manual*, the remainder of this chapter is devoted to an independent examination of the WISC–V structure using both EFA *and* CFA methods, including presentation of variance estimates for subtests and factors as well as model-based latent factor reliability estimates. These analyses and results, which should have been included in the WISC–V manual, are presented in Appendix I in the downloadable resources: www.wiley.com/go/itwiscv to allow clinicians and researchers to assess additional psychometric features of the WISC–V in order to determine if various WISC–V scores possess sufficient evidence of validity for clinical interpretations.

Whereas the complete independent analyses are only available in the downloadable resources: www.wiley.com/go/itwiscv, Canivez and Watkins's conclusions from their analyses are presented here, followed by their general summary.

CONCLUSIONS

We were unable to replicate the structural validity results reported in the *WISC–V Technical and Interpretive Manual.* A comparison of our results in Table I.9 (in downloadable resources: www.wiley.com/go/itwiscv) to results in Table 5.4 of the manual reveals discrepant chi-square values as well as divergences in the reported degrees of freedom for most models. The *WISC–V Technical and Interpretive Manual* reported approximate fit statistics with two-digit precision so it is not possible to make accurate comparisons of our approximate fit statistics with three-digit precision. Further, it is not possible to tell if inadmissible solutions were obtained for the five-factor models (model specification errors such as negative variance) but approximate fit statistics were reported in the *WISC–V Technical and Interpretive Manual* or whether those models converged with proper statistical estimates.

Results from both EFA and CFA conducted in Appendix I (in downloadable resources: www.wiley.com/go/itwiscv) provide important considerations for clinical interpretation of basic scores from the WISC–V. Although the intention was to separate the former Perceptual Rasoning factor into separate Visual Spatial and Fluid Reasoning factors, it appears that this was not very successful, despite endorsement in the final measurement model selected by the publisher and development of standard scores for Visual Spatial and Fluid Reasoning. Had the publisher examined results from EFA or seriously considered the practical and theoretical issues created by the 1.00 loading of Fluid Reasoning on general intelligence in CFA, it would have been apparent that there were significant problems for separate Visual Spatial and Fluid Reasoning factors, given the available 16 WISC–V subtests. Results from EFA and CFA converged and suggest that the best representation of WISC–V measurement is a bifactor model with four specific group factors, but the limited portions of variance uniquely captured by the four specific group factors is low and the ω_S coefficients indicated too little true score variance associated with the four specific group factors, with the possible exception of Processing Speed, to warrant confident interpretation in clinical practice.

GENERAL SUMMARY

Professional standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) demand full and honest disclosure of psychometric features of all scores and comparisons, but sadly many are missing or obfuscated for the WISC–V. Given that numerous critiques and recommendations were known to the publisher (e.g., inclusion of EFA, bifactor CFA models, decomposed variance estimates for scores, provision of validity evidence for interpretations, disclosure of contradictory evidence, and model-based reliability coefficients for composite scores), there is an appearance of intentionality to the absence of these analyses from the *WISC–V Technical and Interpretive Manual*. Clinicians are unable to make evidence-based judgments regarding the psychometric fitness of WISC–V scores or the scientific soundness of interpretation schemes without complete and accurate information. Likewise, researchers cannot adequately understand how to integrate WISC–V scores into theoretical and practical models without complete and accurate information. "Bad usage of tests" (Buros, 1965, p. xxiv) is exacerbated by the great number of score comparisons and analyses promoted for the WISC–V. Users should remember that "just because the test or its scoring software produces a score, you need not interpret it" (Kranzler & Floyd, 2013, p. 95). Furthermore, users must be mindful of the advice of Weiner (1989) that the ethical psychologist will "(a) know what their tests can do and (b) act accordingly" (p. 829). It is our hope that the information in this review and our independent analyses will provide the information necessary for clinicians and researchers to follow this sage advice.

APPENDIX I

FACTOR ANALYSES (CHAPTER 20)

Gary L. Canivez and Marley W. Watkins

Given our criticism of the structural validity analyses reported in the *WISC–V Technical and Interpretive Manual* (Wechsler, 2014), the remainder of Chapter 20—included as Appendix I in the downloadable resources: www.wiley .com/go/itwiscv—is devoted to an independent examination of the WISC–V structure using both exploratory factor analysis (EFA) *and* confirmatory factor analysis (CFA) methods, including presentation of variance estimates for subtests and factors as well as model-based latent factor reliability estimates. These analyses and results should have been included in the *WISC–V Technical and Interpretive Manual*, and are presented here to allow clinicians and researchers to assess additional psychometric features of the WISC–V in order to determine if various WISC–V scores possess sufficient evidence of validity for clinical interpretations.

WISC–V EXPLORATORY FACTOR ANALYSES

In the first section of Appendix I, we present a series of EFA on the WISC–V.

Participants and Procedure

Participants were members of the WISC–V normative sample (N = 2,200) who ranged in age from 6 to 16 years. Demographic characteristics are detailed in the *WISC–V Technical and Interpretive Manual*. The WISC–V 16 subtest correlation matrix for the full standardization sample was obtained from Table 5.1 of that manual; that table was produced by averaging correlations from the 11 WISC–V age groups through Fisher transformations.

Analyses

Principal axis exploratory factor analyses (Fabrigar, Wegener, MacCallum, & Strahan, 1999) were used to analyze the WISC–V standardization sample correlation matrix using SPSS 21 for Macintosh OSX. Multiple criteria as recommended by Gorsuch (1983) were examined to determine the number of factors to retain and included eigenvalues > 1 (Kaiser, 1960), the scree test (Cattell, 1966), standard error of scree (SE_{Scree}; Zoski & Jurs, 1996), Horn's parallel analysis (HPA; Horn, 1965), and minimum average partials (MAP; Velicer, 1976). Because the scree test is a subjective criterion, the SE_{Scree} as programmed by Watkins (2007) was used because it was reported to be the most accurate objective scree method (Nasser, Benson, & Wisenbaker, 2002).

HPA and MAP were included as they are typically more accurate and are helpful so as not to overfactor (Frazier & Youngstrom, 2007; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). HPA indicated meaningful factors when eigenvalues from the WISC–V standardization sample data were larger than eigenvalues produced by random data containing the same number of participants and factors. Random data and resulting eigenvalues for HPA were produced using the Monte Carlo PCA for Parallel Analysis computer program (Watkins, 2000) with 100 replications to provide stable eigenvalue estimates. Retained factors were subjected to promax (oblique) rotation (k = 4; Gorsuch, 1983). Setting k to 4 produced greater hyperplane count compared to k = 2 with the present data. Salient factor pattern coefficients were defined as those ≥.40, but where factor pattern coefficients were between .30 and .39, subtests were designated as "aligned" with the latent factor.

J. B. Carroll (1995) argued that EFA results "should be shown on the basis of orthogonal factors, rather than oblique, correlated factors. I insist, however, that the orthogonal factors should be those produced by the Schmid-Leiman, 1957, orthogonalization procedure" (p. 437). Accordingly, the first-order factor correlation matrix was factor analyzed (principal axis) and first-order factors were orthogonalized by removing all variance associated with the second-order dimension using the Schmid and Leiman (1957) procedure as programmed in the MacOrtho computer program (Watkins, 2004). This transforms "an oblique factor analysis solution containing a hierarchy of higher-order factors into an orthogonal solution which not only preserves the desired interpretation characteristics of the oblique solution, but also discloses the hierarchical structuring of the variables" (Schmid & Leiman, 1957, p. 53).

The Schmid-Leiman (SL) orthogonalization procedure produces an approximate exploratory bifactor (Holzinger & Swineford, 1937) solution (Canivez, in press), has a proportionality constraint (Yung, Thissen, & McLeod, 1999), and may be problematic with nonzero cross-loadings (Reise, 2012). Reise (2012) also noted two additional and more recent alternative exploratory bifactor methods that do not include proportionality constraints: analytic bifactor (Jennrich & Bentler, 2011) and target bifactor (Reise, Moore, & Maydeu-Olivares, 2011). The present application of the SL orthogonalization procedure was selected because there are numerous studies of its application with Wechsler scales (Canivez & Watkins, 2010a, 2010b; Golay & Lecerf, 2011; Watkins, 2006) and with other intelligence tests (Canivez, 2008, 2011; Canivez, Konold, Collins, & Wilson, 2009; Dombrowski & Watkins, 2013; Dombrowski, Watkins, & Brogan, 2009; Nelson & Canivez, 2012; Nelson, Canivez, Lindstrom, & Hatt, 2007), which facilitates direct comparison of WISC–V results to these other studies. For convenience, this method is labeled SL bifactor (Reise, 2012).

Omega-hierarchical and omega-subscale (Reise, 2012) were estimated as model-based reliability estimates of the latent factors (Gignac & Watkins, 2013). F. F. Chen, Hayes, Carver, Laurenceau, and Zhang (2012) noted that "for multidimensional constructs, the alpha coefficient is complexly determined, and McDonald's (1999) omega-hierarchical (ω_h) provides a better estimate for the composite score and thus should be used" (p. 228). These same problems are inherent with other internal consistency estimates such as split-half or KR-20. ω_h is the model-based reliability estimate for the general intelligence factor independent of the variance of group factors. Omega-subscale (ω_s) is the model-based reliability estimate of a group factor with all other group *and* general factors removed (Reise, 2012). Omega estimates (ω_h and ω_s) may be obtained from EFA SL bifactor solutions and were produced here using the *Omega* program (Watkins, 2013), which is based on the tutorial by Brunner, Nagy, and Wilhelm (2012) and the work of Zinbarg, Revelle, Yovel, and Li (2005) and Zinbarg, Yovel, Revelle, and McDonald (2006).

Results

Exploratory Factor Analysis of the 16 WISC–V Primary and Secondary Subtests

Principal axis (principal factors) EFA (SPSS v. 21) produced a Kaiser-Meyer-Olkin Measure of Sampling Adequacy coefficient of .938 (more than adequate according to Kaiser, 1974) and Bartlett's Test of Sphericity was 15,619.3, p < .0001, indicating that the correlation matrix was not random. Communality estimates ranged from .183 (Cancelation) to .735 (Vocabulary) and the *Mdn* = .560.

Factor Extraction Criteria Comparisons Of the six methods to determine how many factors to retain, only the publisher recommended theoretical structure suggested five factors. Minimum average partials indicated one factor; eigenvalues > 1, scree, and parallel analysis each recommended two factors; and the standard error of scree indicated three factors. Figure I.1 presents scree plots from parallel analysis for the 16 WISC–V primary and secondary subtests. Because it has been suggested that it is better to overextract than underextract (Fava & Velicer, 1992; Gorsuch, 1997; Wood, Tataryn, & Gorsuch, 1996), as overextracting allows examination of the performance of smaller factors, EFA began with extracting five factors to examine subtest associations based on the publisher's suggested structure.

First-Order EFA: Five WISC–V Factor Extraction Table I.1 presents results of the extraction of five WISC–V factors with promax rotation. Subtest g-loadings ranged from .219 (Cancellation) to .773 (Vocabulary). What was immediately apparent was that the five-factor model is overfactored. Factor 5 had only one salient subtest pattern coefficient (Figure Weights), and no other subtests were aligned with the fifth factor. Factors cannot be defined by one indicator. This overextraction further resulted in Matrix Reasoning, Picture Concepts, and Arithmetic failing to have salient loadings on any individual factors.

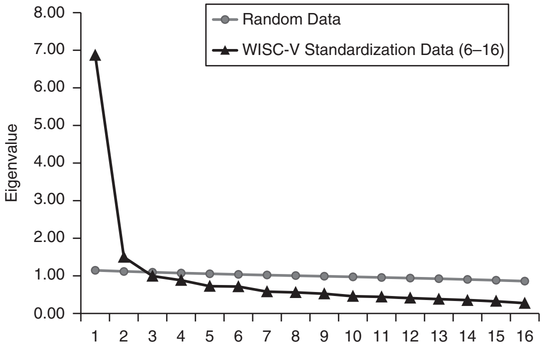

Figure I.1 Scree Plots for Horn's Parallel Analysis for WISC–V Standardization Sample (N = 2,200) 10 Primary Subtests Adapted from Figure 5.1(Wechsler, 2014d), with standardized coefficients, for WISC–V standardization sample (N = 2,200) 16 Subtests. SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, PC = Picture Concepts, FW = Figure Weights, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter–Number Sequencing, CD = Coding, SS = Symbol Search, CA = Cancellation.

Table I.1 illustrates robust Verbal Comprehension (Similarities, Vocabulary, Information, Comprehension), Working Memory (Digit Span, Picture Span, Letter–Number Sequencing), and Processing Speed (Coding, Symbol Search, Cancellation) factors with theoretically consistent and salient subtest associations. The hypothesized Visual Spatial factor (Block Design, Visual Puzzles) also emerged intact. Arithmetic, while failing to exhibit a salient loading on any factor, was moderately aligned with the Working Memory factor. Fluid Reasoning did not emerge as a viable latent factor. The moderate to high factor correlations presented in Table I.1 (.401–.726) imply a higher-order or hierarchical structure that requires explication (Gorsuch, 1983). Thus, ending analyses at this point would be premature for full understanding of the WISC–V structure.

SL Bifactor Analyses: Five WISC–V First-Order Factors Results for the Schmid and Leiman orthogonalization of the higher-order factor analysis of the 16 WISC–V primary and secondary subtests are presented in Table I.2. All subtests (except for Matrix Reasoning and Picture Concepts, which had higher association with the Visual Spatial factor after removing their *g*-variance) were properly associated with their theoretically proposed factor. The hierarchical *g* factor accounted for 35.9% of the total variance and 66.3% of the common variance.

The general factor also accounted for between 4% and 50% (*Mdn* = 42%) of individual subtest variability. At the first-order level, VC accounted

Table I.1 WISC–V Exploratory Factor Analysis: Five Oblique Factor Solution for the Total Standardization Sample ($N = 2{,}200$)

WISC–V Subtest	General	F1: Verbal Comprehension		F2: Working Memory		F3: Visual Spatial		F4: Processing Speed		F5:Impermissible		
	S	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	h^2
SI	.751	**.776**	.805	.049	.595	.022	.575	–.005	.343	–.026	.584	.650
VC	.773	**.887**	.855	–.032	.584	.042	.602	–.055	.309	–.021	.606	.735
IN	.754	**.790**	.816	–.055	.573	–.007	.581	.002	.338	.095	.624	.669
CO	.660	**.721**	.708	.081	.534	–.026	.480	.052	.341	–.101	.479	.510
BD	.667	.033	.555	–.009	.515	**.598**	.732	.127	.445	.079	.582	.554
VP	.686	.046	.586	.022	.516	**.857**	.824	–.071	.331	–.063	.579	.684
MR	.635	.068	.546	.168	.557	.267	.595	.035	.361	.216	.592	.430
FW	.648	.027	.570	–.014	.584	.173	.614	–.038	.295	**.619**	.739	.560
PC	.518	.266	.489	.101	.431	.208	.465	.028	.284	–.008	.418	.275
AR	.725	.219	.466	*.311*	.629	.008	.446	.071	.337	.258	.452	.551
DS	.703	–.039	.565	**.852**	.814	.028	.515	–.042	.392	–.009	.569	.664
PS	.572	.004	.652	**.593**	.682	.095	.563	.000	.423	–.039	.660	.399
LN	.690	.094	.584	**.821**	.792	–.064	.469	–.043	.373	–.047	.538	.634
CD	.419	–.031	.276	.075	.388	–.063	.296	**.758**	.745	–.025	.264	.560
SS	.477	.023	.337	–.004	.405	.038	.377	**.769**	.777	–.042	.308	.605
CA	.219	.014	.151	–.133	.152	.035	.189	**.455**	.418	.024	.149	.183
Eigenvalue		6.87		1.50		1.00		0.88		0.73		
% Variance		40.31		6.39		3.55		3.02		0.89		
Promax Factor Correlations		F1: VC		F2: WM		F3: VS		F4: PS		F5		
Verbal Comprehension (VC)		–										
Working Memory (WM)		.713		–								
Visual Spatial (VS)		.700		.634		–						
Processing Speed (PS)		.417		.518		.463		–				
F5: Impermissible		.724		.707		.726		.401		–		

Note: WISC–V Subtests: SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter–Number Sequencing, CD = Coding, SS = Symbol Search, CA = Cancellation. *S* = Structure Coefficient, *P* = Pattern Coefficient, h^2 = Communality. General structure coefficients are based on the first unrotated factor coefficients (*g*-loadings). Salient pattern coefficients presented in bold (pattern coefficient ≥ .40) and aligned (.30–.39) in italic. Picture Concepts, Arithmetic, and Matrix Reasoning had no salient factor pattern coefficients.

Table I.2 Sources of Variance in the WISC–V for the Total Standardization Sample (N = 2,200) According to an Exploratory Bifactor Model (Orthogonalized Higher-Order Factor Model) with Five First-Order Factors

WISC–V	General		F1: VC		F2: WM		F3: VS		F4: PS		F5			
Subtest	b	S^2	B	S^2	b	S^2	b	S^2	b	S^2	b	S^2	h^2	u^2
SI	.689	.475	**.416**	**.173**	.027	.001	.013	.000	–.004	.000	–.014	.000	.649	.351
VC	.709	.503	**.476**	**.227**	–.018	.000	.024	.001	–.046	.002	–.011	.000	.732	.268
IN	.697	.486	**.424**	**.180**	–.030	.001	–.004	.000	.002	.000	.050	.003	.669	.331
CO	.597	.356	**.387**	**.150**	.045	.002	–.015	.000	.044	.002	–.053	.003	.513	.487
BD	.647	.419	.018	.000	–.005	.000	**.341**	**.116**	.107	.011	.042	.002	.548	.452
VP	.669	.448	.025	.001	.012	.000	**.489**	**.239**	–.060	.004	–.033	.001	.692	.308
MR	.619	.383	.036	.001	.092	.008	*.152*	*.023*	.030	.001	**.114**	**.013**	.430	.570
FW	.658	.433	.014	.000	–.008	.000	.099	.010	–.032	.001	**.328**	**.108**	.552	.448
PC	.488	.238	*.143*	*.020*	.056	.003	*.119*	*.014*	.024	.001	**–.004**	**.000**	.276	.724
AR	.695	.483	.118	.014	*.171*	*.029*	–.005	.000	.060	.004	**.137**	**.019**	.549	.451
DS	.671	.450	–.021	.000	**.468**	**.219**	.016	.000	–.035	.001	–.005	.000	.671	.329
PS	.543	.295	.002	.000	**.326**	**.106**	.054	.003	.000	.000	–.021	.000	.404	.596
LN	.649	.421	.050	.003	**.451**	**.203**	–.037	.001	–.036	.001	–.025	.001	.630	.370
CD	.371	.138	–.017	.000	.041	.002	–.036	.001	**.640**	**.410**	–.013	.000	.551	.449
SS	.425	.181	.012	.000	–.002	.000	.022	.000	**.649**	**.421**	–.022	.000	.603	.397
CA	.194	.038	.008	.000	–.073	.005	.020	.000	**.384**	**.147**	.013	.000	.191	.809
Total S^2		.359		.048		.036		.026		.063		.009	.541	.459
Common S^2		.663		.089		.067		.047		.116		.017		

Note: WISC–V Subtests: SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter–Number Sequencing, CD = Coding, SS = Symbol Search, CA = Cancellation. WISC–V Factors: VC = Verbal Comprehension, WM = Working Memory, VS = Visual Spatial, PS = Processing Speed, FR = Fluid Reasoning. b = loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness. Bold type indicates coefficients and variance estimates consistent with the theoretically proposed factor. Italic type indicates coefficients and variance estimates associated with an alternate factor (where cross-loading b was larger than for the theoretically assigned factor). Given the impermissibility of a five-factor solution, omega coefficients were not estimated for the five-factor model.

for an additional 4.8% of the total variance and 8.9% of the common variance, WM accounted for an additional 3.6% of the total variance and 6.7% of the common variance, VS accounted for an additional 2.6% of the total variance and 4.7% of the common variance, and PS accounted for an additional 6.3% of the total variance and 11.6% of the common variance. The underidentified Factor 5 accounted for an additional 0.9% of the total variance and 1.7% of the common variance. The general and specific group factors combined to measure 54.1% of the variance in WISC–V scores, resulting in 45.9% unique variance (combination of specific and error variance). Subtest specificity (variance unique to the subtest) estimates ranged from .14 to .55. Because of the underidentified fifth factor, omega coefficients were not estimated for the five-group factor solution.

First-Order EFA: Four WISC–V Factor Extraction Table I.3 presents results of the extraction of four WISC–V factors with promax rotation. What was immediately apparent was that the four-factor model appeared to be a better solution than the five-factor model and was very similar to the WISC–IV. Picture Concepts and Arithmetic again failed to exhibit salient loadings on any group factor, but Arithmetic was aligned with its theoretically appropriate Working Memory factor. Picture Concepts displayed evenly divided factor pattern coefficients on Verbal Comprehension and Perceptual Reasoning factors. Table I.4 illustrates robust Verbal Comprehension (Similarities, Vocabulary, Information, Comprehension), Working Memory (Digit Span, Picture Span, Letter–Number Sequencing), and Processing Speed (Coding, Symbol Search, Cancellation) factors with theoretically consistent and salient subtest associations. Block Design, Visual Puzzles, Matrix Reasoning, and Figure Weights converged and had salient factor pattern coefficients on a fourth factor, presumably Perceptual Reasoning. The moderate to high factor correlations presented in Table I.4 (.387–.747) imply a higher-order or hierarchical structure that required explication (Gorsuch, 1983). Thus, ending analyses at this point would again be premature for full understanding of the WISC–V structure.

SL Bifactor Analyses: Four WISC–V First-Order Factors Results for the Schmid and Leiman orthogonalization of the higher-order factor analysis are presented in Table I.4. All subtests (except for Picture Concepts, which had higher association with the Verbal Comprehension factor after removing *g*-variance) were properly associated with their theoretically proposed factor. The hierarchical *g* factor accounted for 35.5% of the total variance and 67.1% of the common variance.

The general factor also accounted for between 4% and 50% (*Mdn* = 39%) of individual subtest variability. At the first-order level, VC accounted for an additional 4.8% of the total variance and 9.2% of the common variance; WM accounted for an additional 3.4% of the total variance and 6.5% of the common variance; PR accounted for an additional 3.0% of the total variance and 5.6% of the common variance; and PS accounted for an additional 6.2% of the total variance and 11.6% of the common variance. The general and group factors combined to measure 53.0% of the variance in WISC–V scores, resulting in 47.0% unique variance (combination of specific and error variance). Subtest specificity (variance unique to the subtest) estimates ranged from .14 to .63.

Omega-hierarchical and omega-subscale coefficients were estimated based on the SL results in Table I.4. Because Picture Concepts had roughly equivalent secondary loadings on Verbal Comprehension *and* Perceptual Reasoning, omega coefficients were separately estimated with Picture Concepts assigned to Verbal Comprehension and then assigned to Perceptual Reasoning. Omega-subscale (ω_s) coefficients for Verbal Comprehension and Perceptual Reasoning were both lower when Picture Concepts was assigned to the respective group factor. The ω_h coefficients for general intelligence

Table I.3 WISC–V Exploratory Factor Analysis: Four Oblique Factor Solution for the Total Standardization Sample (*N* = 2,200)

	General	F1: Verbal Comprehension		F2: Working Memory		F3: Perceptual Reasoning		F4: Processing Speed		
WISC–V Subtest	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*h*²
Similarities (SI)	.752	**.766**	.805	.037	.597	.017	.615	.001	.323	.649
Vocabulary (VC)	.774	**.878**	.856	–.046	.586	.039	.641	–.047	.288	.735
Information (IN)	.754	**.794**	.815	–.033	.577	.062	.630	–.004	.316	.666
Comprehension (CO)	.660	**.703**	.707	.057	.534	–.082	.511	.064	.326	.506
Block Design (BD)	.669	–.011	.550	–.051	.511	**.738**	.750	.119	.427	.573
Visual Puzzles (VP)	.679	.024	.582	–.045	.514	**.815**	.781	–.045	.314	.612
Matrix Reasoning (MR)	.636	.071	.544	.198	.559	**.436**	.634	.016	.340	.431
Figure Weights (FW)	.638	.126	.569	.134	.537	**.503**	.657	–.072	.272	.454
Picture Concepts (PC)	.519	.249	.488	.080	.431	.227	.483	.031	.270	.274
Arithmetic (AR)	.724	.248	.464	*.373*	.627	.157	.476	.050	.322	.534
Digit Span (DS)	.704	–.049	.564	**.845**	.813	.029	.562	–.035	.373	.663
Picture Span (PS)	.573	–.012	.652	**.572**	.684	.085	.623	.008	.400	.396
Letter–Number Sequencing (LN)	.690	.085	.584	**.814**	.792	–.096	.517	–.033	.355	.634
Coding (CD)	.420	–.025	.273	.085	.384	–.070	.311	**.747**	.747	.562
Symbol Search (SS)	.477	.023	.333	–.007	.401	.030	.387	**.756**	.776	.603
Cancellation (CA)	.220	.018	.149	–.124	.150	.064	.194	**.443**	.419	.182
Eigenvalue		6.87		1.50		1.00		0.88		
% Variance		40.23		6.38		3.51		2.85		
Promax Based Factor Correlations		F1: VC		F2: WM		F3: PR		F4: PS		
F1: Verbal Comprehension (VC)		–								
F2: Working Memory (WM)		.716		–						
F3: Perceptual Reasoning (PR)		.747		.693		–				
F4: Processing Speed (PS)		.387		.490		.456		–		

Note: S = Structure Coefficient, *P* = Pattern Coefficient, h^2 = Communality. General structure coefficients are based on the first unrotated factor coefficients (*g*-loadings). Salient pattern coefficients (≥ .40) presented in bold and aligned pattern coefficients (.30-.39) in italic. Picture Concepts and Arithmetic had no salient factor pattern coefficients. Picture Concepts had no aligned factor pattern coefficients.

Table I.4 Sources of Variance in the WISC–V for the Total Standardization Sample *(N* = 2,200) According to an Exploratory Bifactor Model (Orthogonalized Higher-Order Factor Model) with Four First-Order Factors

	General		F1: Verbal Comprehension		F2: Working Memory		F3: Perceptual Reasoning		F4: Processing Speed			
WISC–V Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
Similarities (SI)	.690	.476	**.415**	**.172**	.020	.000	.009	.000	.001	.000	.649	.351
Vocabulary (VC)	.707	.500	**.476**	**.227**	–.024	.001	.020	.000	–.040	.002	.729	.271
Information (IN)	.690	.476	**.431**	**.186**	–.018	.000	.032	.001	–.003	.000	.663	.337
Comprehension (CO)	.602	.362	**.381**	**.145**	.030	.001	–.042	.002	.054	.003	.513	.487
Block Design (BD)	.642	.412	–.006	.000	–.027	.001	**.382**	**.146**	.101	.010	.569	.431
Visual Puzzles (VP)	.656	.430	.013	.000	–.024	.001	**.422**	**.178**	–.038	.001	.611	.389
Matrix Reasoning (MR)	.609	.371	.038	.001	.105	.011	**.226**	**.051**	.014	.000	.435	.565
Figure Weights (FW)	.612	.375	.068	.005	.071	.005	**.260**	**.068**	–.061	.004	.456	.544
Picture Concepts (PC)	.487	.237	*.135*	*.018*	.043	.002	**.118**	**.014**	.026	.001	.272	.728
Arithmetic (AR)	.685	.469	.134	.018	**.198**	**.039**	.081	.007	.043	.002	.535	.465
Digit Span (DS)	.681	.464	–.027	.001	**.450**	**.203**	.015	.000	–.030	.001	.668	.332
Picture Span (PS)	.551	.304	–.007	.000	**.304**	**.092**	.044	.002	.007	.000	.398	.602
Letter–Number Sequencing (LN)	.661	.437	.046	.002	**.433**	**.187**	–.050	.003	–.028	.001	.630	.370
Coding (CD)	.383	.147	–.014	.000	.045	.002	–.036	.001	**.636**	**.404**	.555	.445
Symbol Search (SS)	.435	.189	.012	.000	–.004	.000	.016	.000	**.644**	**.415**	.604	.396
Cancellation (CA)	.197	.039	.010	.000	–.066	.004	.033	.001	**.377**	**.142**	.186	.814
Total Variance		.355		.048		.034		.030		.062	.530	.470
Common Variance		.671		.092		.065		.056		.116		
Picture Concepts in VC	$\omega_h = .833$		$\omega_s = .216$		$\omega_s = .185$		$\omega_s = .167$		$\omega_s = .505$			
Picture Concepts in PR	$\omega_h = .834$		$\omega_s = .250$		$\omega_s = .185$		$\omega_s = .144$		$\omega_s = .505$			

Note: b = loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness, ω_h = omega-hierarchical, ω_s = omega-subscale. Bold type indicates coefficients and variance estimates consistent with the theoretically proposed factor. Italic type indicates coefficients and variance estimates associated with an alternate factor (where cross-loading *b* was larger than for the theoretically assigned factor). Omega-hierarchical and omega-subscale coefficients were estimated with Picture Concepts associated with Verbal Comprehension (VC) and also with Perceptual Reasoning (PR) to examine effects of differential assignment.

(.833 and .834) were high and sufficient for scale interpretation; however, the ω_S coefficients for the four WISC–V specific group factors (VC, WM, PR, PS) were considerably lower. Thus, the four specific WISC-IV group factors, with the possible exception of PS, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise, Bonifay, & Haviland, 2013).

Exploratory Factor Analyses of the 10 WISC–V Primary Subtests

Principal axis (principal factors) EFA (SPSS v. 21) produced a Kaiser-Meyer-Olkin Measure of Sampling Adequacy coefficient of .884 (more than adequate according to Kaiser, 1974) and the chi-square value from Bartlett's Test of Sphericity was 8,324.68, $p < .0001$, indicating that the correlation matrix was not random. Communality estimates ranged from .473 (Matrix Reasoning) to .742 (Visual Puzzles) with a median of .546.

Factor Extraction Criteria Comparisons Of the six methods to determine how many factors to extract, only the publisher recommended structure suggested five factors. Minimum average partials indicated one factor; eigenvalues > 1, scree the standard error of scree, and parallel analysis each recommended two factors. Figure I.2 presents scree plots from HPA for the WISC–V 10 primary subtests. Because it has been suggested that it is better to overextract than underextract (Fava & Velicer, 1992; Gorsuch, 1997; Wood et al., 1996), which allows examination of the performance of smaller factors, EFA again began with extracting five factors to examine subtest associations based on the publisher's suggested structure for the 10 primary subtests.

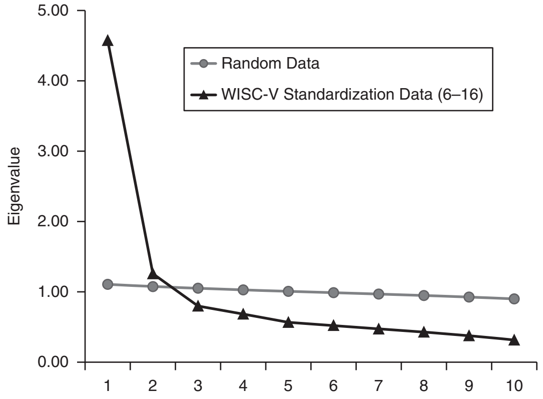

Figure I.2 Scree Plots for Horn's Parallel Analysis for WISC–V Standardization Sample (N = 2,200) 10 Primary Subtests

First-Order EFA: Five WISC–V Factor Extraction Table I.5 presents results of the extraction of five WISC–V factors from the 10 primary subtests with promax rotation. Subtest *g*-loadings ranged from .451 (Coding) to .741 (Vocabulary). When only the 10 primary subtests are included, salient factor pattern coefficients were produced for subtests on the theoretically consistent factors, and no salient cross-loadings were observed. The moderate to high factor correlations presented in Table I.6 (.374–.740) imply a higher-order or hierarchical structure that requires explication (Gorsuch, 1983).

SL Bifactor Analyses: Five WISC–V First-Order Factors Results for the Schmid and Leiman orthogonalization of the higher-order factor analysis are presented in Table I.6. All subtests were properly associated with their theoretically proposed factor after removing *g* variance and all subtests except Coding and Symbol Search had larger portions of subtest variance associated with the hierarchical general factor. The hierarchical *g* factor accounted for 36.9% of the total variance and 63.5% of the common variance.

The general factor also accounted for between 18.1% and 47.3% (*Mdn* = 42.1%) of individual subtest variability. At the first-order level, VS accounted for an additional 4.2% of the total variance and 7.2% of the common variance, VC accounted for an additional 4.4% of the total variance and 7.5% of the common variance, PS accounted for an additional 8.8% of the total variance and 15.1% of the common variance, WM accounted for an additional 3.0% of the total variance and 5.1% of the common variance,

Table I.5 WISC–V Exploratory Factor Analysis of the 10 Primary Subtests: Five Oblique Factor Solution for the Total Standardization Sample (*N* = 2,200)

WISC–V Subtest	General	F1: Visual Spatial		F2: Verbal Comprehension		F3: Processing Speed		F4: Working Memory		F5: Fluid Reasoning		
	S	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	h^2
SI	.724	.006	.568	**.747**	.800	.025	.336	.054	.576	.005	.607	.643
VC	.741	.036	.597	**.856**	.851	–.016	.303	–.025	.559	–.010	.617	.725
BD	.696	**.521**	.712	.024	.551	.127	.428	–.032	.514	.195	.626	.544
VP	.728	**.907**	.859	.018	.586	–.049	.313	.030	.521	–.081	.590	.742
MR	.650	.148	.570	.016	.536	.013	.350	.087	.555	**.485**	.674	.473
FW	.647	.208	.595	.102	.569	–.071	.279	.018	.522	**.462**	.668	.480
DS	.670	.007	.504	.065	.554	.022	.393	**.605**	.733	.093	.604	.548
PS	.580	.014	.426	–.015	.456	–.004	.341	**.738**	.700	–.046	.497	.491
CD	.451	–.048	.279	.001	.258	**.841**	.804	–.017	.358	–.019	.328	.651
SS	.495	.062	.358	.001	.311	**.693**	.726	.028	.400	–.015	.375	.531
Eigenvalue		4.57		1.26		0.80		0.69		0.57		
% Variance		41.64		8.67		3.75		3.16		1.07		
Factor Correlations		F1: VS		F2: VC		F3: PS		F4: WM		FR		
Visual Spatial (VS)		–										
Verbal Comprehension (VC)		.689		–								
Processing Speed (PS)		.417		.374		–						
Working Memory (WM)		.620		.673		.497		–				
Fluid Reasoning (FR)		.726		.731		.468		.740		–		

Note: WISC–V Subtests: SI = Similarities, VC = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search. *S* = Structure Coefficient, *P* = Pattern Coefficient, h^2 = Communality. General structure coefficients are based on the first unrotated factor coefficients (*g*-loadings). Salient pattern coefficients presented in bold (pattern coefficient ≥ .40).

Table I.6 Sources of Variance in the WISC–V 10 Primary Subtests for the Total Standardization Sample (N = 2,200) According to an Exploratory Bifactor Model (Orthogonalized Higher-Order Factor Model) with Five First-Order Factors

WISC–V	General		F1: VS		F2: VC		F3: PS		F4: WM		F5: FR			
Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
SI	.676	.457	.004	.000	**.433**	**.187**	.021	.000	.030	.001	.002	.000	.646	.354
VC	.688	.473	.022	.000	**.496**	**.246**	−.014	.000	−.014	.000	−.004	.000	.720	.280
BD	.652	.425	**.312**	**.097**	.014	.000	.108	.012	−.018	.000	.086	.007	.542	.458
VP	.667	.445	**.543**	**.295**	.010	.000	−.042	.002	.017	.000	−.036	.001	.743	.257
MR	.646	.417	.089	.008	.009	.000	.011	.000	.049	.002	**.213**	**.045**	.473	.527
FW	.642	.412	.125	.016	.059	.003	−.060	.004	.010	.000	**.203**	**.041**	.476	.524
DS	.653	.426	.004	.000	.038	.001	.019	.000	**.342**	**.117**	.041	.002	.547	.453
PS	.564	.318	.008	.000	−.009	.000	−.003	.000	**.417**	**.174**	−.020	.000	.493	.507
CD	.374	.140	−.029	.001	.001	.000	**.715**	**.511**	−.010	.000	−.008	.000	.652	.348
SS	.425	.181	.037	.001	.001	.000	**.589**	**.347**	.016	.000	−.007	.000	.529	.471
Total S^2		.369		.042		.044		.088		.030		.010	.582	.418
Common S^2		.635		.072		.075		.151		.051		.017		
	$\omega_h = .812$		$\omega_s = .228$		$\omega_s = .257$		$\omega_s = .538$		$\omega_s = .191$		$\omega_s = .059$			

Note: WISC–V Subtests: SI = Similarities, VC = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search. WISC–V Factors: VS = Visual Spatial, VC = Verbal Comprehension, PS = Processing Speed, WM = Working Memory, FR = Fluid Reasoning. *b* = loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness. Bold type indicates coefficients and variance estimates consistent with the theoretically proposed factor.

and FR accounted for an additional 1.0% of the total variance and 1.7% of the common variance. The general and specific group factors combined to measure 58.2% of the variance in WISC–V scores, resulting in 41.8% unique variance (combination of specific and error variance).

Omega-hierarchical and omega-subscale coefficients were estimated based on the SL results in Table I.6 to estimate the latent factor reliabilities. The ω_h coefficient for general intelligence (.812) was high and sufficient for scale interpretation; however, the ω_s coefficients for the five WISC–V specific group factors (VS, VC, PS, WM, FR) were considerably lower, ranging from .059 (FR) to .538 (PS). Thus, the five WISC–V first-order factors, with the possible exception of PS, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise et al., 2013).

First-Order EFA: Four WISC–V Factor Extraction Table I.7 presents results of the extraction of four WISC–V factors from the 10 primary subtests with promax rotation. Subtest *g*-loadings ranged from .450 (Coding) to .744 (Vocabulary). When only the 10 primary subtests are included, salient factor pattern coefficients were produced for subtests on the theoretically consistent factors. and no salient cross-loadings were observed. The moderate to high factor correlations presented in Table I.7 (.346–.742) imply a higher-order or hierarchical structure that required explication (Gorsuch, 1983).

SL Bifactor Analyses: Four WISC–V First-Order Factors Results for the Schmid and Leiman orthogonalization of the higher-order factor analysis are presented in Table I.8. All subtests were properly associated with their theoretically proposed factor after removing *g* variance, and all subtests except Coding and Symbol Search had larger portions of subtest variance associated with the hierarchical general factor. The hierarchical *g* factor accounted for 36.7% of the total variance and 64.9% of the common variance.

The general factor also accounted for between 13.7% and 47.7% (*Mdn* = 40.9%) of individual subtest variability. At the first-order level, PR accounted for an additional 3.9% of the total variance and 6.8% of the common variance; VC accounted for an additional 4.4% of the total variance and 7.8% of the common variance; PS accounted for an additional 8.7% of the total variance and 15.4% of the common variance; and WM accounted for an additional 2.8% of the total variance and 5.0% of the common variance. The general and specific group factors combined to measure 56.5% of the variance in WISC–V scores, resulting in 43.5% unique variance (combination of specific and error variance).

Omega-hierarchical and omega-subscale coefficients were estimated based on the SL results in Table I.8 to estimate the latent factor reliabilities. The ω_h coefficient for general intelligence (.800) was high and sufficient for scale interpretation; however, the ω_s coefficients for the four WISC–V factors (PR, VC, PS, WM) were considerably lower, ranging from .142 (PR) to .538 (PS). Thus, the four WISC–V first-order factors, with the possible exception of PS, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise et al., 2013).

WISC–V CONFIRMATORY FACTOR ANALYSES

Preference for the higher-order model (general intelligence as a superordinate dimension) without examination and direct comparison with a rival bifactor model (general intelligence as a breadth dimension) is unwarranted (Canivez & Kush, 2013; Gignac, 2008). Bifactor models allow all subtests to load directly on both the general factor and a group factor whereas

Table I.7 WISC–V Exploratory Factor Analysis of the 10 Primary Subtests: Four Oblique Factor Solution for the Total Standardization Sample (*N* = 2,200)

	General	F1: Perceptual Reasoning		F2: Verbal Comprehension		F3: Processing Speed		F4: Working Memory		
WISC–V Subtest	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*P*	*S*	h^2
Similarities (SI)	.725	.026	.619	**.726**	.796	.030	.320	.057	.598	.637
Vocabulary (VC)	.744	.045	.645	**.855**	.855	–.009	.286	–.044	.583	.732
Block Design (BD)	.705	**.777**	.760	–.029	.543	.105	.415	–.059	.530	.587
Visual Puzzles (VP)	.707	**.820**	.781	.020	.581	–.054	.301	–.041	.536	.613
Matrix Reasoning (MR)	.645	**.412**	.631	.061	.535	.014	.332	.233	.578	.432
Figure Weights (FW)	.643	**.468**	.648	.133	.566	–.070	.262	.157	.553	.447
Digit Span (DS)	.679	.011	.563	.022	.551	.010	.379	**.739**	.767	.589
Picture Span (PS)	.574	.002	.473	–.017	.455	.014	.330	**.664**	.660	.437
Coding (CD)	.450	–.050	.320	.009	.251	**.817**	.795	–.005	.360	.634
Symbol Search (SS)	.498	.064	.392	.001	.304	**.692**	.733	.023	.405	.542
Eigenvalue		4.57		1.26		0.80		0.69		
% Variance		41.45		8.56		3.39		3.09		
Promax Based Factor Correlations		F1: PR		F2: VC		F3: PS		F4: WM		
F1: Perceptual Reasoning (PR)		–								
F2: Verbal Comprehension (VC)		.742		–						
F3: Processing Speed (PS)		.449		.346		–				
F4: Working Memory (WM)		.719		.700		.483		–		

Note: S = Structure Coefficient, *P* = Pattern Coefficient, h^2 = Communality. General structure coefficients are based on the first unrotated factor coefficients (*g*-loadings). Salient pattern coefficients (≥ .40) presented in bold.

Table I.8 Sources of Variance in the WISC–V 10 Primary Subtests for the Total Standardization Sample (N = 2,200) According to an Exploratory Bifactor Model (Orthogonalized Higher-Order Factor Model) with Four First–Order Factors

	General		F1: Perceptual Reasoning		F2: Verbal Comprehension		F3: Processing Speed		F4: Working Memory			
WISC–V Subtest	b	S^2	b	S^2	b	S^2	b	S^2	b	S^2	h^2	u^2
Similarities (SI)	.676	.457	.012	.000	**.425**	**.181**	.026	.001	.029	.001	.639	.361
Vocabulary (VC)	.691	.477	.022	.000	**.500**	**.250**	−.008	.000	−.023	.001	.729	.271
Block Design (BD)	.660	.436	**.373**	**.139**	−.017	.000	.091	.008	−.030	.001	.584	.416
Visual Puzzles (VP)	.673	.453	**.394**	**.155**	.012	.000	−.047	.002	−.021	.000	.611	.389
Matrix Reasoning (MR)	.618	.382	**.198**	**.039**	.036	.001	.012	.000	.120	.014	.437	.563
Figure Weights (FW)	.618	.382	**.225**	**.051**	.078	.006	−.061	.004	.081	.007	.449	.551
Digit Span (DS)	.667	.445	.005	.000	.013	.000	.009	.000	**.380**	**.144**	.590	.410
Picture Span (PS)	.565	.319	.001	.000	−.010	.000	.012	.000	**.341**	**.116**	.436	.564
Coding (CD)	.370	.137	−.024	.001	.005	.000	**.706**	**.498**	−.003	.000	.636	.364
Symbol Search (SS)	.425	.181	.031	.001	.001	.000	**.598**	**.358**	.012	.000	.539	.461
Total Variance		.367		.039		.044		.087		.028	.565	.435
Common Variance		.649		.068		.078		.154		.050		
		ω_h = .800		ω_s = .142		ω_s = .255		ω_s = .538		ω_s = .173		

Note: b = loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness, ω_h = omega-hierarchical, ω_s = omega-subscale. Bold type indicates coefficients and variance estimates consistent with the theoretically proposed factor.

higher-order factors restrict subtests to indirect loadings on the general factor, mediated by the group factors. Bifactor CFA models have several technical benefits over EFA orthogonal solutions (Reise, 2012), have been found to fit data from other Wechsler scales (viz., Canivez, 2014a; Gignac & Watkins, 2013; Nelson, Canivez, & Watkins, 2013; Watkins, 2010; Watkins & Beaujean, 2014; Watkins, Canivez, James, James, & Good, 2013), and have been recommended for cognitive tests (Brunner et al., 2012; Canivez, in press; Gignac, 2005, 2006). Figures 5.1 and 5.2 in the *WISC–V Technical and Interpretive Manual* (and Figure 20.1 in Chapter 20 of *Intelligent Testing with the WISC–V*) illustrate higher-order models and Figures I.5 and I.6 in this appendix illustrate bifactor models.

Participants and Analyses

Participants were identical to those previously employed in EFA analyses; namely, the 2,200 participants in the WISC–V standardization sample. CFA was implemented with Mplus 7.3 (Muthén & Muthén, 2014). Covariance matrices were computed by Mplus from the correlation matrix, means, and standard deviations of the total normative sample reported in Table 5.1 of the *WISC–V Technical and Interpretive Manual*. Given the size of the normative sample and the multivariate normality of these data, maximum likelihood (ML) estimation of model parameters was employed. The ML estimator is "asymptotically consistent, unbiased, efficient, and normally distributed" (Lei & Wu, 2012, p. 167) and is the default method in Mplus. Scaling for identification of models was accomplished with marker variables (T. D. Little, Slegers, & Card, 2006), the default in Mplus.

The structural models specified in Table 5.3 of the *WISC–V Technical and Interpretive Manual* were tested. In addition, bifactor models were included in our analyses. See Figures 21.4 and 21.5 for a complete enumeration of the models examined in this study. Although there are no universally accepted cutoff values for approximate fit indices (McDonald, 2010), overall model fit was evaluated with the comparative fit index (CFI), root mean square error of approximation (RMSEA), Tucker-Lewis index (TLI), and the standardized root mean squared residual (SRMR). Higher values indicate better fit for the CFI and TLI whereas lower values indicate better fit for the SRMR and RMSEA. Applying the combinatorial heuristics of Hu and Bentler (1999), criteria for adequate model fit were CFI and TLI ≥ .90 along with SRMR ≤ .09 and RMSEA ≤ .08. Good model fit required CFI ≥ 0.95 with SRMR and RMSEA ≤ 0.06 (Hu & Bentler, 1999). For a model to be considered superior, it had to exhibit adequate to good overall fit and display meaningfully better fit (ΔCFI > .01 and ΔRMSEA > .015) than alternative models (Cheung & Rensvold, 2002). Additionally, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were consulted. Neither AIC nor BIC has a meaningful scale but the model with the smallest AIC and BIC values are most likely to replicate (Kline, 2011).

Finally, to assess the latent factor reliabilities, omega-hierarchical and omega-subscale (Reise, 2012) were estimated as more appropriate reliability estimates of the factors (Gignac & Watkins, 2013). ω_h is the model-based reliability estimate for the general intelligence factor with variability of group factors removed. Omega-subscale (ω_s) is the model-based reliability estimate of a specific group factor with all other group *and* general factors removed (Brunner et al., 2012; Reise, 2012). Omega estimates (ω_h and ω_s) may be obtained from CFA bifactor solutions and were produced here using the *Omega* program (Watkins, 2013).

Results

Results from CFAs are presented in Table I.9. For the 16 WISC–V primary and secondary

	Model 2		Model 3			Model 4a				Model 4b				Model 4c				Model 4d				Model 4a Bi-factor				
Subtest	F1	F2	F1	F2	F3	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4	g	F1	F2	F3	F4
SI	◆		◆			◆				▲				▲				▲				◆	◆			
VC	◆		◆			◆				▲				▲				▲				◆	◆			
IN	▲		▲			▲				▲				▲				▲				▲	▲			
CO	▲		▲			▲				▲				▲				▲				▲	▲			
BD		◆		◆			◆				▲				▲				▲			◆		◆		
VP		◆		◆			◆				▲				▲				▲			◆		◆		
MR		◆		◆			◆					▲			▲				▲			◆		◆		
FW		◆		◆			◆					▲			▲				▲			◆		◆		
PC		▲		▲			▲					▲			▲				▲			▲		▲		
AR	▲		▲					▲				▲			▲	▲		▲	▲	▲		▲			▲	
DS	◆		▲		▼			◆				▲				▲				▲		◆			◆	
PS		◆		◆				◆				▲				▲				▲		◆			◆	
LN	▲		▲					▲				▲				▲				▲		▲			▲	
CD		◆			◆				◆				▲				▲				▲	◆				◆
SS		◆			◆				◆				▲				▲				▲	◆				◆
CA		▲			▲				▲				▲				▲				▲	▲				▲

Figure I.3 Confirmatory Factor Models of WISC–V Subtests with Two to Four First-Order Factors σ = 16 subtest models, τ = 10 subtest models, and υ = both 10 and 16 subtest models. Models 2 to 4d include a higher-order general factor. SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, CD = Coding, SS = Symbol Search, and CA = Cancellation.

	Model 5a					Model 5b					Model 5c					Model 5d					Model 5e				
Subtest	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
SI	◆					▲					▲					▲					▲				
VC	◆					▲					▲					▲					▲				
IN	▲					▲					▲					▲					▲				
CO	▲					▲					▲					▲					▲				
BD		◆					▲					▲					▲					▲			
VP		◆					▲					▲					▲					▲			
MR			◆					▲					▲					▲					▲		
FW			◆					▲					▲					▲					▲		
PC			▲					▲					▲					▲					▲		
AR				▲				▲					▲	▲		▲			▲		▲		▲	▲	
DS				◆					▲					▲					▲					▲	
PS				◆					▲					▲					▲					▲	
LN				▲					▲					▲					▲					▲	
CD					◆					▲					▲					▲					▲
SS					◆					▲					▲					▲					▲
CA					▲					▲					▲					▲					▲

Figure I.4 Confirmatory Factor Models of WISC–V Subtests with Five First-Order Factors σ = 16 subtest models, τ = 10 subtest models, and υ = both 10 and 16 subtest models. Models 2 to 4d include a higher-order general factor. SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, CD = Coding, SS = Symbol Search, and CA = Cancellation.

Table I.9 CFA Fit Statistics for the WISC–V 16 Primary and Secondary Subtests and 10 Primary Subtests for the Total Standardization Sample (N = 2,200)

Model	χ^2	*df*	CFI	TLI	SRMR	RMSEA	RMSEA 90%	AIC	BIC
16 Subtests									
1	2539.7	104	.843	.819	.062	.103	.100, .107	4184.9	4458.3
2	2179.0	103	.866	.844	.059	.096	.092, .099	3826.3	4105.4
3	1351.1	101	.920	.904	.042	.075	.071, .079	3002.3	3292.9
4a	577.9	100	.969	.963	.030	.047	.043, .050	2231.1	2527.3
4b	756.6	100	.958	.949	.032	.055	.051, .058	2409.8	2706.0
4c	467.3	99	.976	.971	.027	.041	.037, .045	2122.5	2424.4
4d	433.6	98	.978	.974	.026	.039	.036, .043	2090.8	2398.4
4a Bifactor	**312.9**	**88**	**.986**	**.980**	**.021**	**.034**	**.030, .038**	**1990.2**	**2354.7**
5a*	Model specification error (negative variance estimate for Fluid Reasoning), improper model fit statistics not reported								
5b*	Model specification error (negative variance estimate for Fluid Reasoning), improper model fit statistics not reported								
5c*	Model specification error (negative variance estimate for Fluid Reasoning), improper model fit statistics not reported								
5d*	Model specification error (negative variance estimate for Fluid Reasoning), improper model fit statistics not reported								
5e*	Model specification error (negative variance estimate for Fluid Reasoning), improper model fit statistics not reported								
10 Subtests									
1	1297.1	35	.848	.804	.065	.128	.122, .134	3775.1	3945.9
2	1126.4	34	.868	.826	.066	.121	.115, .127	3606.4	3782.9
3	871.5	32	.899	.858	.057	.109	.103, .116	3355.4	3543.4
4a	185.0	31	.981	.973	.025	.048	.041, .054	2670.9	2864.6
4 Bifactor	**126.1**	**28**	**.988**	**.981**	**.022**	**.040**	**.033, .047**	**2618.0**	2828.8
5a	134.1	30	.987	.981	.023	.040	.033, .047	2622.0	2821.4

Note: CFI = comparative fit index, TLI = Tucker-Lewis index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, AIC = Akaike information criterion, and BIC = Bayesian information criterion. For bifactor models, identification achieved by constraining indicator loadings to equality if only two indicators for each factor. Bold text illustrates best-fitting model. *Model contains specification error.

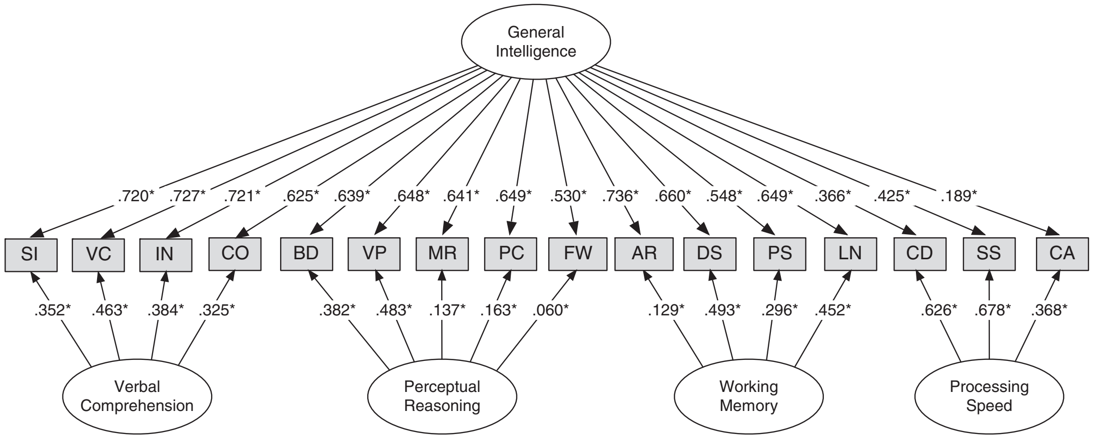

Figure I.5 Bifactor Measurement Model (4a Bifactor), with Standardized Coefficients, for WISC–V Standardization Sample *N* = 2,200, 16 Subtests. SI = Similarities, VC = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter–Number Sequencing, CD = Coding, SS = Symbol Search, CA = Cancellation. *p < .05.

subtests, all five of the models that included five first-order factors resulted in inadmissible solutions (i.e., negative variance estimates for the FR factor) potentially caused by misspecification of the models (Kline, 2011). In contrast, all five models that included four first-order factors demonstrated good fit to these data. No single four-factor model was superior in terms of ΔCFI > .01 and ΔRMSEA > .015, but AIC and BIC values were lowest for the bifactor version that collapsed the FR and VS dimensions (r = .91) into a single (PR) factor (see Figure I.5).

Table I.10 presents sources of variance from the 16 WISC–V primary and secondary subtests according to the bifactor model with four group factors, which are very similar to the SL bifactor results from EFA. Most subtest variance is associated with the general intelligence dimension, and much smaller portions of variance are uniquely associated with the four specific group factors. Omega-hierarchical and omega-subscale coefficients were estimated based on the bifactor results from Table I.10. The ω_h coefficient for general intelligence (.849) was high and sufficient for scale interpretation; however, the ω_s coefficients for the four WISC–V factors (PR, VC, PS, WM) were considerably lower, ranging from .109 (PR) to .516 (PS). Thus, the four WISC–V first-order factors, with the possible exception of PS, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise et al., 2013).

For the 10 WISC–V primary subtests, four- and five-factor models demonstrated good fit to these data. No single four- or five-factor model was superior in terms of ΔCFI > .01 and ΔRMSEA > .015, but AIC values were lowest for the bifactor version that collapsed the FR and VS dimensions (r = .90) into a single (PR) factor (see Figure 21.7). Table I.11 presents sources of variance from the 10 WISC–V primary subtests according to the bifactor model with four group factors, which are very similar to SL bifactor results. Again, most subtest variance is associated with the general intelligence

Table I.10 Sources of Variance in the WISC–V 16 Subtests for the Total Standardization Sample ($N = 2{,}200$) According to a CFA Bifactor Model

	General		Verbal Comprehension		Perceptual Reasoning		Working Memory		Processing Speed			
WISC–V Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
Similarities (SI)	.720	.518	.352	.124							.642	.358
Vocabulary (VC)	.727	.529	.463	.214							.743	.257
Information (IN)	.721	.520	.384	.147							.667	.333
Comprehension (CO)	.625	.391	.325	.106							.496	.504
Block Design (BD)	.639	.408			.382	.146					.554	.446
Visual Puzzles (VP)	.648	.420			.483	.233					.653	.347
Matrix Reasoning (MR)	.641	.411			.137	.019					.430	.570
Figure Weights (FW)	.649	.421			.163	.027					.448	.552
Picture Concepts (PC)	.530	.281			.060	.004					.285	.716
Arithmetic (AR)	.736	.542					.129	.017			.558	.442
Digit Span (DS)	.660	.436					.493	.243			.679	.321
Picture Span (PS)	.548	.300					.296	.088			.388	.612
Letter–Number Sequencing (LN)	.649	.421					.452	.204			.626	.374
Coding (CD)	.366	.134							.626	.392	.526	.474
Symbol Search (SS)	.425	.181							.678	.460	.640	.360
Cancellation (CA)	.189	.036							.368	.135	.171	.829
Total Variance		.372		.037		.027		.034		.062	.532	.468
Common Variance		.699		.070		.050		.065		.116		
		$\omega_h = .849$		$\omega_s = .201$		$\omega_s = .109$		$\omega_s = .181$		$\omega_s = .516$		

Note: b = standardized loading of subtest on factor, S^2 = variance explained in the subtest, h^2 = communality, u^2 = uniqueness, ω_h = omega hierarchical, ω_s = omega subscale.

Table I.11 Sources of Variance in the WISC–V 10 Primary Subtests for the Total Standardization Sample ($N = 2{,}200$) According to a CFA Bifactor Model

	General		Verbal Comprehension		Perceptual Reasoning		Working Memory		Processing Speed			
Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
Similarities (SI)	.693	.480	.441	.194							.675	.325
Vocabulary (VC)	.702	.493	.441	.194							.687	.313
Block Design (BD)	.673	.453			.289	.084					.536	.464
Visual Puzzles (VP)	.677	.458			.498	.248					.706	.294
Matrix Reasoning (MR)	.670	.449			.041	.002					.451	.549
Figure Weights (FW)	.664	.441			.099	.010					.451	.549
Digit Span (DS)	.655	.429					.388	.151			.580	.420
Picture Span (PS)	.549	.301					.388	.151			.452	.548
Coding (CD)	.357	.127							.655	.429	.556	.444
Symbol Search (SS)	.424	.180							.655	.429	.609	.391
Total Variance		.381		.039		.034		.030		.086	.570	.430
Common Variance		.668		.068		.060		.053		.150		
		$\omega_h = .817$		$\omega_s = .231$		$\omega_s = .087$		$\omega_s = .199$		$\omega_s = .543$		

Note: b = standardized loading of subtest on factor, S^2 = variance explained in the subtest, h^2 = communality, u^2 = uniqueness, ω_h = omega hierarchical, ω_s = omega subscale.

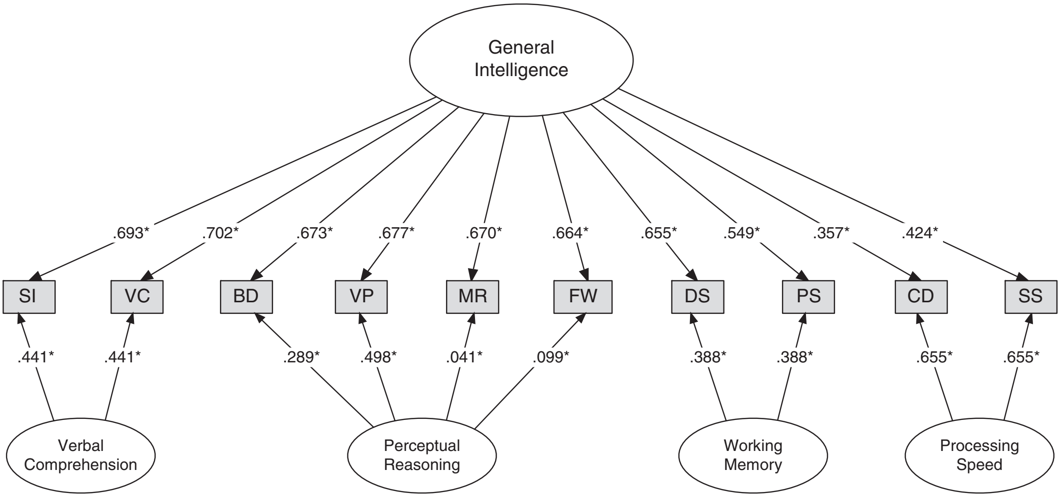

Figure I.6 Bifactor measurement model (4a Bifactor), with standardized coefficients, for WISC–V standardization sample (*N* = 2,200) 10 Primary Subtests. SI = Similarities, VC = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search. *$p < .05$.

dimension and smaller portions of variance are uniquely associated with the specific group factors. Omega-hierarchical and omega-subscale coefficients were estimated based on the bifactor results from Table I.11. The ω_h coefficient for general intelligence (.817) was high and sufficient for scale interpretation; however, the ω_s coefficients for the four WISC–V factors (PR, VC, PS, WM) were considerably lower, ranging from .087 (PR) to .543 (PS). Thus, the four WISC–V first-order factors, with the possible exception of PS, likely possess too little true score variance for clinicians to interpret (Reise, 2012; Reise et al., 2013).

REFERENCES

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

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

Aspel, A. D., Willis, W. G., & Faust, D. (1998). School psychologists' diagnostic decision-making processes: Objective-subjective discrepancies. *Journal of School Psychology, 36,* 137-149.

Babad, E. Y., Mann, M., & Mar-Hayim, M. (1975). Bias in scoring the WISC subtests. *Journal of Consulting and Clinical Psychology, 43,* 268. doi:10.1037/h0076368

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7,* 461-483. doi:10.1207/S153280 07SEM0703_6

Borsuk, E. R., Watkins, M. W., & Canivez, G. L. (2006). Long-term stability of membership in a Wechsler Intelligence Scale for Children-Third Edition (WISC-III) subtest core profile taxonomy. *Journal of Psychoeducational Assessment, 24,* 52-68. doi:10.1177/0734 282905285225

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). New York, NY: Guilford Press.

Braden, J. P., & Shaw, S. R. (2009). Intervention validity of cognitive assessment: Knowns, unknowables, and unknowns. *Assessment for Effective Intervention, 34,* 106-115. doi:10.1177/1534508407313013

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

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

Buros, O. K. (1965). *The sixth mental measurements yearbook*. Highland Park, NJ: Gryphon Press.

Canivez, G. L. (2008). Orthogonal higher–order factor structure of the Stanford–Binet Intelligence Scales for children and adolescents. *School Psychology Quarterly, 23,* 533–541. doi: 10.1037/a0012884

Canivez, G. L. (2010). Review of the Wechsler Adult Intelligence Test–Fourth Edition. In R. A. Spies, J. F. Carlson, and K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 684–688). Lincoln, NE: Buros Institute of Mental Measurements.

Canivez, G. L. (2011). Hierarchical factor structure of the Cognitive Assessment System: Variance partitions from the Schmid–Leiman (1957) procedure. *School Psychology Quarterly, 26,* 305-317. doi:10.1037/a0025973

Canivez, G. L. (2013a). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean, (Eds.), *The Oxford Handbook of Child Psychological Assessments* (pp. 84–112). New York, NY: Oxford University Press.

Canivez, G. L. (2013b). Incremental validity of WAIS–IV factor index scores: Relationships with WIAT–II and WIAT-III subtest and composite scores. *Psychological Assessment, 25,* 484-495. doi:10.1037/a0032092

Canivez, G. L. (2014a). Review of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The nineteenth mental measurements yearbook* (pp. 732-737). Lincoln, NE: Buros Institute of Mental Measurements.

Canivez, G. L. (2014b). Construct validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. *School Psychology Quarterly, 29,* 38-51. doi:10.1037/spq0000032

Canivez, G. L. (in press). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements*. Gottingen, Germany: Hogrefe.

Canivez, G. L., Konold, T. R., Collins, J. M., & Wilson, G. (2009). Construct validity of the Wechsler Abbreviated Scale of Intelligence and Wide Range Intelligence Test:

Convergent and structural validity. *School Psychology Quarterly, 24,* 252–265. doi: 10.1037/a0018030

Canivez, G. L., & Kush, J. C. (2013). WISC–IV and WAIS–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:10.1177/0734282913478036

Canivez, G. L., Neitzel, R., & Martin, B. E. (2005). Construct validity of the Kaufman Brief Intelligence Test, Wechsler Intelligence Scale for Children-Third Edition, and Adjustment Scales for Children and Adolescents. *Journal of Psychoeducational Assessment, 23,* 15-34. doi:10.1177/073428290502300102

Canivez, G. L., & Watkins, M. W. (2010a). Investigation of the factor structure of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): Exploratory and higher-order factor analyses. *Psychological Assessment, 22,* 827–836. doi:10.1037/a0020429

Canivez, G. L., & Watkins, M. W. (2010b). Exploratory and higher-order factor analyses of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS–IV) adolescent subsample. *School Psychology Quarterly, 25*, 223-235. doi:10.1037/a0022046

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

Carroll, J. B. (1993). *Human cognitive abilities.* Cambridge, England: Cambridge University Press.

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

Carroll, J. B. (1998). Human cognitive abilities: A critique. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 5-23). Mahwah, NJ: Erlbaum.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York, NY: Pergamon.

Carroll, J. B. (2012). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 883-890). New York, NY: Guilford.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276. doi:10.1207/s15327906mbr0102_10

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15,* 139-164. doi:10.1111/j.1745-3984.1978.tb00065.x

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,* 219-251. doi:10.1111/j.1467-6494.2011.00739.x

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255. doi:10.1207/S15328007SEM0902_5

Cornoldi, C., Orsini, A., Cianci, L., Giofrè, D., & Pezzuti, L. (2013). Intelligence and working memory control: Evidence from the WISC-IV administraation to Italian children. *Learning and Individual Differences, 26,* 9-14. doi:10.1016/j.lindif.2013.04.005

Cronbach, L. J., & Snow, R E. (1977). *Aptitudes and instructional methods*. New York, NY: Wiley.

Devena, S. E., & Watkins, M. W. (2012). Diagnostic utility of WISC-IV general abilities index and cognitive proficiency index reference scores among children with ADHD. *Journal of Applied School Psychology, 28,* 133-154. doi:10.1080/15377903.2012.669743

Dombrowski, S. C., & Watkins, M. W. (2013). Exploratory and higher order factor analysis of the WJ-III full test battery: A school aged analysis. *Psychological Assessment, 25,* 442-455. doi:10.1037/a0031335

Dombrowski, S. C., Watkins, M. W., & Brogan, M. J. (2009). An exploratory investigation of the factor structure of the Reynolds Intellectual Assessment Scales (RIAS). *Journal of Psychoeducational Assessment, 27,* 494-507. doi:10.1177/07342829093 33179

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299. doi:10.1037/1082-989X.4.3.272

Faust, D. (1989). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences & the Law, 7,* 469-483. doi:10.1002/bsl.2370070405

Faust, D. (2007). Some global and specific thoughts about some global and specific issues. *Applied Neuropsychology, 14,* 26-36. doi:10.1080/09084280701280411

Fava, J. L., & Velicer, W. F. (1992) The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 21,* 387-415. doi:10.1207/s15327906 mbr2703_5

Fina, A. D., Sánchez-Escobedo, P., & Hollingworth, L. (2012). Annotations on Mexico's WISC-IV: A validity study. *Applied Neuropsychology: Child, 1,* 6-17. doi: I0. 1080/21622965.20 I2.665771

Flanagan, D. P., Ortiz, S. O., & Alphonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). New York, NY: Wiley.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39,* 414-423. doi:10.1037/0735-7028.39.4.414

Floyd, R. G., Reynolds, M. R., Farmer, R. L., & Kranzler, J. H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis. *School Psychology Review, 42,* 383-401.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35,* 169−182. doi:10.1016/j.intell.2006. 07.002

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,* 131-139. doi:10.1080/ 09084280802084010

Fuchs, D., & Fuchs, L. S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research, 56,* 243-262.

Gignac, G. E. (2005). Revisiting the factor structure of the WAIS-R: Insights through nested factor modeling. *Assessment, 12,* 320-329. doi:10.1177/1073191105278118

Gignac, G. E. (2006). The WAIS-III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences, 27,* 73-86. doi:10.1027/1614-0001.27.2.73

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: *g* as superordinate or breadth factor? *Psychology Science Quarterly, 50,* 21-43.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48,* 639-662. doi:10.1080/00273171.2013.804398

Glutting, J. J., & McDermott, P. A. (1990). Principles and problems in learning potential. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 296-347). New York, NY: Guilford.

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement, 47,* 607-614. doi:10.1177/ 001316448704700307

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC–IV in estimating reading and math achievement on the WIAI–II. *Journal of Special Education, 40,* 103–114. doi:10.1177/0022466906 0400020101

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,* 295-301. doi:10.1037/1040-3590.9.3.295

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment, 23,* 143–152. doi:10.1037/a0021230

Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2013). Further insights on the French WISC-IV factor structure through Bayesian structural equation modeling (BSEM). *Psychological Assessment, 25,* 496-508. doi:10.1037/a0030676

Good, R. H., Vollmer, M., Creek, R. J., Katz, L. I., & Chowdhri, S. (1993). Treatment utility of the Kaufman Assessment Battery for Children: Effects of matching instruction and student processing strength. *School Psychology Review, 22,* 8-26.

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

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68, 532*-560. doi:10.1207/s15327752jpa6803_5

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly, 12,* 249-267. doi:10.1037/h0088961

Hale, J. B., Fiorello, C. A., Miller, J. A., Wenrich, K., Teodori, A., & Henzel, J. N. (2008). WISC-IV interpretation for specific learning disabilities and intervention: A cognitive hypothesis testing approach. In A. Prifitera, D. H. Saklofske, & E. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (2nd ed., pp. 109–171). New York, NY: Elsevier.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology, 19,* 370-376. doi:10.1016/0022-4405(81)90031-5

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment, 15,* 456-466. doi:10.1037/1040-3590.15.4.456

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54. doi:10.1007/BF02287965

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185. doi:10.1007/BF02289447

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock-Johnson technical manual* (Rev. ed., pp. 197-232). Itasca, IL: Riverside.

Horn, J. L., & Blankson, A. N. (2012). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 73-98). New York, NY: Guilford.

Hu, L. -T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 5,* 1-55. doi:10.1080/10705519909540118

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362. doi:10.1037/0033-2909.112.2.351

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment, 15,* 443–445. doi:10.1037/1040-3590.15.4.443

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29-51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446–455. doi:10.1037/1040-3590.15.4.446

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76,* 537-549. doi:10.1007/s11336-011-9218-4

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence, 36,* 81-95. doi:10.1016/j.intell.2007.06.001

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151. doi: 10.1177/001316446002000116

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36. doi:10.1007/BF02291575

Kamphaus, R. W., Reynolds, C. R., & Vogel, K. K. (2009). Intelligence testing. In J. L. Matson, F. Andrasik, & M. L. Matson (Eds.), *Assessing childhood psychopathology and* *developmental disabilities* (pp. 91-115). New York, NY: Springer.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A history of intelligence test interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 56-70). New York, NY: Guilford.

Kaufman, A. S., Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2006). Review of the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV). *Journal of Psychoeducational Assessment, 24,* 278-295. doi:10.1177/0734282906288389

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement* (3rd ed.). Bloomington, MN: NCS Pearson.

Kavale, K., & Mattson, P. D. (1983). "One jumped off the balance beam": Meta-analysis of perceptual-motor training. *Journal of Learning Disabilities, 16,* 165-173. doi:10.1177/002221948301600307

Kearns, D. M., & Fuchs, D. (2013). Does cognitively focused instruction improve the academic performance of low-achieving students? *Exceptional Children, 79,* 263-290.

Keith, T. Z. (1997). What does the WISC-III measure? A reply to Carroll and Kranzler. *School Psychology Quarterly, 12,* 117–118. doi:10.1037/h0088953

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

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self–administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399. doi:10.1037/1040-3590.5.4.395

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Kramer, J. J., Henning-Stout, M., Ullman, D. P., & Schellenberg, R. P. (1987). The viability of scatter analysis on the WISC-R and the SBIS: Examining a vestige. *Journal of Psychoeducational Assessment, 5,* 37-47. doi:10.1177/073428298700500105

Kranzler, J. H. (1997). What does the WISC-III measure? Comments on the relationship between intelligence, working memory capacity, and information processing speed and efficiency. *School Psychology Quarterly, 12,* 110–116. doi:10.1037/h0088952

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

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology, 3,* 55-81. doi:10.1300/J151v03n03_04

Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164-180). New York, NY: Guilford.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50,* 7-36. doi:10.1016/j.jsp.2011.09.006

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in sem and macs models. *Structural Equation Modeling, 13,* 59-72. doi:10.1207/s15328007sem1301_3

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "Intelligent Testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197-234. doi:10.1037/h0088959

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290-302. doi:10.1177/073428299000800307

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. *Journal of Special Education, 25,* 504-526. doi:10.1177/002246699202500407

McDermott, P. A., Marston, N. C., & Stott, D. H. (1993). *Adjustment Scales for Children and Adolescents*. Philadelphia, PA: Edumetric and Clinical Science.

McDermott, P. A., Watkins, M. W., & Rhoad, A. (2014). Whose IQ is it?–Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26,* 207-214. doi:10.1037/a0034832

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675-686. doi:10. 1177/1745691610388766

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York, NY: Guilford.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-181). New York, NY: Guilford.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216. doi:10.1037/h0048070

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749. doi:10.1037/0003-066X.50.9.741

Miciak, J., Fletcher, J. M., Vaughn, S., Stuebing, K. K., & Tolar, T. D. (2014). Patterns of cognitive strengths and weaknesses: Identification rates, agreement, and validity for learning disability identification. *School Psychology Quarterly, 29,* 21-37. doi:10.1037/spq0000037

Millsap, R. E. (2007). Structural equation modeling made difficult. P*ersonality and Individual Differences, 42,* 875-881. doi:10.1016/j.paid.2006.09.021

Moon, G. W., Blakey, W. A., Gorsuch, R. L., & Fantuzzo, J. W. (1991). Frequent WAIS-R administration errors: An ignored source of inaccurate measurement. *Professional Psychology: Research and Practice, 22,* 256–258. doi:10.1037/0735-7028.22.3.256

Muthén, B. O., & Muthén, L. K. (2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and achievement: Just how correlated are they? *Journal of Psychoeducational Assessment, 21,* 244−260. doi:10.1177/073428290302100 302

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System. Interpretive handbook.* Rolling Meadows, IL: Riverside.

Nasser, F., Benson, J., & Wisenbaker, J. (2002). The performance of regression-based variations of the visual scree for determining the number of common factors. *Educational and Psychological Measurement, 62,* 397-419. doi:10.1177/00164402062003001

Nelson, J. M., & Canivez, G. L. (2012). Examination of the structural, convergent, and incremental validity of the Reynolds Intellectual Assessment Scales (RIAS) with a clinical sample. *Psychological Assessment, 24,* 129-140. doi:10.1037/a0024878

Nelson, J. M, Canivez, G. L, Lindstrom, W., & Hatt, C. (2007). Higher-order exploratory factor analysis of the Reynolds Intellectual Assessment Scales with a referred sample. *Journal of School Psychology, 45*, 439-456. doi: 10.1016/j.jsp.2007.03.003

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) with a clinical sample. *Psychological Assessment, 25,* 618-630. doi:10.1037/a0032086

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175-220. doi:10.1037/1089-2680.2.2.175

Oakland, T., Lee, S. W., & Axelrad, K. M. (1975). Examiner differences on actual WISC protocols. *Journal of School Psychology, 13,* 227–233. doi:10.1016/0022-4405(75)90005-9

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7,* 557-595. doi:10.1207/S15328007SEM0704_3

Parkin, J. R., & Beaujean, A. A. (2012). The effects of Wechsler Intelligence Scale for Children-Fourth Edition cognitive abilities on math achievement. *Journal of School Psychology, 50,* 113-128. doi:10.1016/ j.jsp.2011.08.003

Pearson. (2009). *Wechsler Individual Achievement Test-Third Edition*. San Antonio, TX: Author.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329-353. doi:10.1207/s15327906mbr3204_2

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518-524. doi:10.1037/0021-9010.79.4.518

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667-696. doi: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,* 129-140. doi:10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71,* 684-711. doi:10.1177/0013164410378690

Reschly, D. J. (1997). Diagnostic and treatment utility of intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 437-456). New York, NY: Guilford.

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). New York, NY: Oxford University Press.

Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, pp. 67-93). Hoboken, NJ: Wiley.

Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the chc taxonomy. *Journal of School Psychology, 51,* 535-555. doi: 10.1016/j.jsp.2013.02.003

Ruscio, J., & Stern, A. R. (2006). The consistency and accuracy of holistic judgment. *Scientific Review of Mental Health Practice, 4,* 52-65.

Ryan, J. J., Kreiner, D. S., & Burton, D. B. (2002). Does high scatter affect the predictive validity of WAIS-III IQs? *Applied Neuropsychology, 9,* 173-178. doi:10.1207/S15324826AN0903_5

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

Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assessment, 31,* 186-201. doi:10.1177/0734282913478046

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. *Measurement and Evaluation in Counseling and Development, 25,* 156–161.

Spearman, C. (1927). *The abilities of man*. New York, NY: Cambridge University Press.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23,* 63-86. doi:10.1080/08957340903423651

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237-244. doi:10.\037/11040-3590.12.3.237

Stuebing, K. K., Fletcher, J. M., Branum-Martin, L., & Francis, D. J. (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review, 41,* 3-22.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychological diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.

Szarko, J. E., Brown, A. J., & Watkins, M. W. (2013). Examiner familiarity effects for children with autism spectrum disorders. *Journal of Applied School Psychology, 29,* 37-51. doi:10.1080/15377903.2013.751475

Taub, G. E. (2014). An empirical investigation comparing the WISC-III and WISC-IV: Implications for the Atkins criterion. *Open Access Journal of Forensic Psychology, 6,* 1-16.

Terman, L. M. (1918). Errors in scoring Binet tests. *Psychological Clinic, 12,* 33–39.

Treat, T. A. & Viken, R. J. (2012). Measuring test performance with signal detection theory techniques. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.); *Handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (Vol. 1, pp. 723–744). Washington, DC: American Psychological Association.

Velicer, W. F. (1976). Determining the number of components form the matrix of partial correlations. *Psychometrika, 31,* 321-327. doi:10.1007/BF02293557

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A view and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: A festschrift to Douglas Jackson at seventy* (pp. 41–71). Norwell, MA: Kluwer Academic.

Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist, 20,* 723-733. doi:10.1037/h0021472

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, 2nd ed., pp. 50-81). Hoboken, NJ: Wiley.

Watkins, M. W. (1999). Diagnostic utility of WISC-III subtest variability among students with learning disabilities. *Canadian Journal of School Psychology, 15,* 11-20. doi:10.1177/082957359901500102

Watkins, M. W. (2000). *Monte Carlo PCA for Parallel Analysis* [Computer Software]. State College, PA: Ed & Psych Associates.

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

Watkins, M. W. (2004). *MacOrtho.* [Computer Software]. State College, PA: Ed & Psych Associates.

Watkins, M. W. (2005). Diagnostic validity of Wechsler subtest scatter. *Learning Disabilities: A Contemporary Journal, 3,* 20-29.

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

Watkins, M. W. (2007). *SEscree* [Computer software]. State College, PA: Ed & Psych Associates.

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). Hoboken, NJ: Wiley.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children - Fourth Edition among a national sample of referred students. *Psychological Assessment, 22*, 782-787. doi:10.1037/a0020043

Watkins, M. W. (2013). *Omega*. [Computer software]. Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler Preschool and Primary Scale of Intelligence-Fourth edition. *School Psychology Quarterly, 29,* 52-63. doi:10.1037/spq0000038

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite strengths and weaknesses. *Psychological Assessment, 16,* 133-138. doi:10.1037/1040–3590.16.2.133

Watkins, M. W., Canivez, G. L., James, T., James, K., & Good, R. (2013). Construct validity of the WISC-IV^{UK} with a large referred Irish sample. *International Journal of School & Educational Psychology, 1,* 102-111. doi:10.1080/21683603.2013.794439

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,* 13-20. doi:10.1080/09084280701280353

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251-268). New York, NY: Guilford.

Watkins, M. W., & Ravert, C. M. (2013). Subtests, factors, and constructs: What is being measured by tests of intelligence? In J. C. Kush (Ed.), *Intelligence quotient: Testing, role of genetics and the environment and social outcomes* (pp. 55-68). New York, NY: Nova Science.

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

Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1949). *Wechsler intelligence scale for children.* New York, NY: The Psychological Corporation.

Wechsler, D. (1981). *Manual for the Wechsler Intelligence Scale for Children-Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition*. San Antonio, TX: The Psychological Corporation.

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

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale—Fourth Edition*. San Antonio, TX:
NCS Pearson.

Wechsler, D. (2012a). *Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition.* San Antonio, TX: NCS Pearson.

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children-Fifth Edition.* San Antonio, TX: NCS Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children-Fifth Edition administration and scoring manual.* San Antonio, TX: NCS Pearson.

Wechsler, D. (2014c). *Wechsler Intelligence Scale for Children-Fifth Edition administration and scoring manual supplement.* San Antonio, TX: NCS Pearson.

Wechsler, D. (2014d). *Wechsler Intelligence Scale for Children-Fifth Edition technical and interpretive manual.* San Antonio, TX: NCS Pearson.

Wechsler, D. (2014e). *Technical and interpretive manual supplement: Special group validity studies with other measures and additional tables*. San Antonio, TX: NCS Pearson.

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). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 94-113. doi:10.1177/0734282913478030

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

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and over-extraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1,* 354-365. doi:10.1037/1082-989X.1.4.354

Yuan, K. –H., & Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika, 70,* 791-798. doi:10.1007/s11336-001-0930-910.1007/s11336-0010930-9

Yung, Y. –F., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64,* 113-128. doi:10.1007/BF02294531

Zhu, J., & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29,* 570-580. doi:10.1177/0734282910396323

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123–133. doi:10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement, 30,* 121–144. doi:10.1177/0146621605278814

Zirkel, P. A. (2013). The Hale position for a "third method" for specific learning disabilities identification: A legal analysis. *Learning Disability Quarterly, 36,* 93-96. doi:10.1177/0731948713477850

Zirkel, P. A. (2014). The legal quality of articles published in school psychology journals: An initial report card. *School Psychology Review, 43,* 318-339.

Zoski, K. W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement, 56,* 443-451. doi:10.1177/0013164496056003006

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 117,* 253–269. doi:10.1037/0033-2909.99.3.432