BUROS
CENTER FOR TESTING

TEST REVIEWS

Woodcock-Johnson® IV

Woodcock-Johnson® IV

Purpose

Designed as a set of "norm-referenced tests for measuring intellectual abilities, academic achievement, and oral language abilities."

Population

Ages 2-90+ years.

Publication Dates

1977-2014.

Acronym

WJ IV®.

Administration

Individual.

Parts, 3

Tests of Achievement, Tests of Cognitive Abilities, Tests of Oral Language.

Price Data, 2014

$2,176.90 per Complete Battery Plus (Achievement Form A or Form B, Cognitive Abilities, Oral Language) with case; $1,838.85 per Complete Kit (Achievement Form A or Form B, Cognitive Abilities) with case; $1,967.90 per Complete Achievement Battery (Forms A, B, and C) with case; $907.35 per Achievement Battery (Form A, B, or C) with case; $1,265 per Cognitive Battery with case; $658.90 per Oral Language kit with case; $1,430 per Oral Language with Cognitive Battery with case; $1,243 per Oral Language with Achievement (Form A) with case; $152.75 per 25 Cognitive Abilities test records with individual score reports; $152.75 per 25 Achievement standard and extended test records and response books with individual score reports (Form A, B, or C); $79 per 25 Oral Language test records with individual score reports; $59 per 25 Cognitive Abilities response books; $59 per Achievement standard and extended response books (Form A, B, or C).

Comments

Cognitive, Achievement, and Oral Language batteries are co-normed and may be used separately or together; tests within each battery may be administered separately or in combinations. Online scoring and reporting available.

Authors

Frederick A. Schrank (tests, online scoring and reporting program), Kevin S. McGrew (tests), Nancy Mather (tests and examiner's manuals), Barbara J. Wendling (examiner's manuals), and David Dailey (online scoring and reporting program).

Publisher

Houghton Mifflin Harcourt.

a) TESTS OF ACHIEVEMENT.

Acronym

WJ IV ACH.

Scores, 42

11 Standard Battery test scores: Letter Word Identification, Applied Problems, Spelling, Passage Comprehension, Calculation, Writing Samples, Word Attack, Oral Reading, Sentence Reading Fluency, Math Facts Fluency, Sentence Writing Fluency; 9 Extended Battery test scores: Reading Recall, Number Matrices, Editing, Word Reading Fluency, Spelling of Sounds, Reading Vocabulary, Science, Social Studies, Humanities; 22 cluster scores: Reading, Broad Reading, Basic Reading Skills, Reading Comprehension, Reading Comprehension–Extended, Reading Fluency, Reading Rate, Mathematics, Broad Mathematics, Math Calculation Skills, Math Problem Solving, Written Language, Broad Written Language, Basic Writing Skills, Written Expression, Brief Achievement, Broad Achievement, Academic Skills, Academic Fluency, Academic Applications, Academic Knowledge, Phoneme-Grapheme Knowledge.

Forms

Standard Battery has 3 parallel forms: A, B, C.

Time

(40) minutes for core set of six tests in Standard Battery; (15-20) minutes for Writing Samples test; (5-10) minutes each for remaining tests.

b) TESTS OF COGNITIVE ABILITIES:

Acronym

WJ IV COG.

Scores, 35

10 Standard Battery test scores: Oral Vocabulary, Number Series, Verbal Attention, Letter-Pattern Matching, Phonological Processing, Story Recall, Visualization, General Information, Concept Formation, Numbers Reversed; 8 Extended Battery test scores: Number-Pattern Matching, Nonword Repetition, Visual-Auditory Learning, Picture Recognition, Analysis-Synthesis, Object-Number Sequencing, Pair Cancellation, Memory for Words; 4 ability scores: General Intellectual Ability, Gf-Gc Composite, Brief Intellectual Ability, Scholastic Aptitudes (Reading Aptitude, Math Aptitude, Writing Aptitude); 7 broad ability clusters: Comprehension-Knowledge, Fluid Reasoning, Short-Term Working Memory, Cognitive Processing Speed, Auditory Processing, Long-Term Retrieval, Visual Processing; 6 narrow ability clusters: Perceptual Speed, Quantitative Reasoning, Auditory Memory Span, Number Facility, Vocabulary, Cognitive Efficiency.

Time

(35) minutes for first seven tests in Standard Battery; (5) minutes for each additional test.

c) TESTS OF ORAL LANGUAGE:

Acronym

WJ IV OL.

Scores, 24

12 test scores: Picture Vocabulary, Oral Comprehension, Segmentation, Rapid Picture Naming, Sentence Repetition, Understanding Directions, Sound Blending, Retrieval Fluency, Sound Awareness, Vocabulario Sobre Dibujos, Comprensión Oral, Comprensión de Indicaciones; 9 cluster scores: Oral Language, Broad Oral Language, Oral Expression, Listening Comprehension, Phonetic Coding, Speed of Lexical Access, Lenguaje Oral, Amplio Lenguaje Oral, Comprensión Auditiva; 3 additional cluster scores can be derived by combining Oral Language tests with certain tests from the Cognitive Abilities battery: Vocabulary, Comprehension-Knowledge—Extended, Auditory Memory Span.

Time

(40) minutes for first eight tests.

Comments

The three Spanish clusters are parallel to three of the English clusters and can be used to compare the examinee's proficiency in English and Spanish.

Cross References

For reviews by Gregory J. Cizek and Jonathan Sandoval of the third edition, see 15:281; see T5:2901 (140 references); for reviews by Jack A. Cummings and by Steven W. Lee and Elaine Flory Stefany of the 1991 edition, see 12:415 (56 references); see also T4:2973 (90 references); for reviews by Jack A. Cummings and Alan S. Kaufman of the 1977 edition, see 9:1387 (6 references); see also T3:2639 (3 references).

Review of the Woodcock-Johnson IV by GARY L. CANIVEZ, Professor of Psychology, Department of Psychology, Eastern Illinois University, Charleston, IL:

DESCRIPTION

The Woodcock–Johnson IV (WJ IV) is a collection of three distinct individually administered test batteries constructed to be consistent with Cattell–Horn–Carroll (CHC) theory of cognitive abilities (Schneider & McGrew, 2012). The three batteries were co–normed and include the Woodcock–Johnson IV Tests of Cognitive Abilities (WJ IV COG), the Woodcock–Johnson IV Tests of Achievement (WJ IV ACH), and the Woodcock–Johnson IV Tests of Oral Language (WJ IV OL). Oral language may be assessed in English and Spanish. The three batteries may be used individually or in any combination. The WJ IV is a major revision of its predecessor, the Woodcock–Johnson III (WJ III; Woodcock, McGrew, & Mather, 2001, 2007), and includes national standardization. Normative data were produced by a nationally representative standardization sample of 7,416 individuals between the ages of 2 and 90+ years. The WJ IV includes an extensive technical manual with the largest compilation of statistical methods, detailed descriptions, and results this reviewer can recall detailing the development and preliminary evaluation of the WJ IV. The WJ IV was designed for use in clinical and educational assessments as well as in research. After administering the measure, the examiner calculates the test raw scores and completes the Test Observation Checklist to then enter information into the online scoring and reporting system. Scoring and analyses are no longer performed by hand. Each of the three examiner's manuals includes training checklists to facilitate development of mastery of administration and scoring prior to clinical use.

DEVELOPMENT

The WJ IV is a major revision of the WJ III, and development and revision goals intended to increase administration and interpretation options using the most up-to-date version of CHC theory and research. New tests (eight) and cluster scores in cognitive, academic achievement, and oral language domains were created (while others were eliminated) with a focus on increasing cognitive complexity (increasing cognitive information processing, demands on memory and attention control, and executive functioning) as well as facilitating measurement of relative strengths and weaknesses across domains. In addition to new tests being created, several tests had items added to the very low or very high difficulty range

(some had both), which lowers the test floor and/or raises the ceiling for better assessment of very low or very high abilities and/or very young children. WJ IV content was reviewed by content experts and also examined for potential bias. Empirical investigation of item bias (differential item functioning [DIF]) was conducted across variables of sex (male and female), race (White and non–White), and ethnicity (Hispanic and Not Hispanic) and reported in the technical manual. Dichotomizing race into White and non-White is problematic as non–invariance might be observed in some racial groups but not others and could be obscured by combining different groups into one "non–White" category. Most items did not demonstrate DIF, and of those that did, most reportedly were not included in publication forms.

The WJ IV COG measures general intelligence hierarchically ordered above broad cognitive abilities (first–order factors), which are measured by several narrow cognitive abilities at the subtest level. As such, the broad cognitive abilities (cluster scores) are inferences or abstractions from the observed test performance on the narrow ability tests, whereas general intelligence (cognitive composites) is an inference from inferences or abstraction from abstractions (Thompson, 2004). The Standard Battery includes 10 tests, and the Extended Battery includes eight additional tests. General intelligence (cognitive composite) is estimated by either a Brief Intellectual Ability (BIA) score (Tests 1–3) that includes one test each of Comprehension–Knowledge (or Gc), Fluid Reasoning (or Gf), and Short–Term Working Memory (or Gwm); a General Intellectual Ability (GIA) score (Tests 1–7) that includes one test each of Comprehension–Knowledge, Fluid Reasoning, Short–Term Working Memory, Processing Speed (or Gs), Auditory Processing (or Ga), Long–Term Retrieval (or Glr), and Visual Processing (or Gv); and/or a Gf–Gc Composite (Tests 1, 2, 8, 9) that includes two tests of Comprehension–Knowledge and two tests of Fluid Reasoning. The core WJ IV COG is represented by the first seven tests from the Standard Battery and yields the GIA. Unlike the BIA where the three tests were equally weighted, the GIA is a weighted composite with tests possessing higher g loadings providing greater influence.

The WJ IV ACH battery includes 20 tests measuring reading (8), mathematics (4), written language (4), science (1), social studies (1), humanities (1), and spelling of sounds (1) that are combined to produce six reading, four mathematics, four writing, and six cross–domain cluster scores. The most frequently used achievement tests are the 11 Standard Battery tests. There are three parallel forms of WJ IV ACH tests (A, B, C) to minimize item exposure in repeated administrations.

The WJ IV OL battery includes 12 tests that were reported to be ordered to maximize interpretation options with minimum testing and to allow for assessment of strengths and weaknesses in oral language and cognitive-linguistic skills. There are nine Oral Language cluster scores and two Oral Language and Cognitive scores. The WJ IV OL tests of Oral Language and Listening Comprehension clusters were adapted for Spanish administration using parallel form development, not translation. Administering tests in English and Spanish allows for direct comparison of English and Spanish language skills that inform language dominance and proficiency.

TECHNICAL

Standardization

The technical manual provides detailed information about the development of WJ IV norms, data for which were collected over a 2-year period from 46 U.S. states and the District of Columbia. The standardization sample included 7,416 individuals (664 preschool children [ages 2–5]; 3,891 students in kindergarten through grade 12; 775 undergraduate and graduate students; and 2,086 adults). Stratified random sampling was employed and stratification variables included geographic region, sex, country of birth, race, ethnicity, community size, parent education level, school type, college type, educational attainment (adults), employment status (adults), and occupational level (adults). Close approximations to 2010 U.S. Census estimates were obtained and where differences were observed, partial subject weighting was used.

Because the WJ IV includes 50 tests, it was reportedly not practical to administer all tests to each participant in the standardization sample. Therefore, a method of planned missing data ("partial matrix sampling" [technical manual, p. 68]) was employed whereby the tests were divided into four blocks that included 15 to 19 tests each. One block consisted of a common set of tests to allow for linking. Following data collection the use of maximum likelihood estimation with multiple imputation created data (scores) for each standardization participant on tests that they had not themselves taken. The final data augmentation multiple imputation produced 10 data sets (1,000 iterations each), and one of the 10 data sets was randomly selected for WJ IV norms and statistical analyses. These scores were then used in norm development for each test. What is unknown is how adequate the 1,000 iterations were relative to the amount of missing information present in the partial matrix sampling procedure (Graham, Olchowski, & Gilreath, 2007).

Following generation of complete standardization sample data, "individual examinee weights were applied during the WJ IV norming data construction process to insure test, cluster, and difference score norms were based on a sample with characteristics proportional to the U.S. population distribution" (technical manual, p. 71). The method of calculation of percentile ranks and standard score norms used a procedure of "maintaining the real-world skew of score distributions" (technical manual, p. 83). As with all tests of cognitive abilities, it can be argued that skew (and kurtosis) are a function of sampling error and should be removed if the assumption is that those abilities are normally distributed in the population of interest. A variety of scores are provided via the online scoring and reporting system including W scale (an item response theory score), deviation IQ, z scores, T scores, stanines, normal curve equivalents, percentile ranks, relative proficiency indexes, developmental zones, and cognitive–academic language proficiency levels. Users may also select age- or grade-based norms depending on

the more appropriate reference group for interpretation.

Reliability

Internal consistency reliabilities for untimed tests and dichotomously scored items were estimated with the split–half method, whereas tests with subtests and cluster reliabilities were estimated with Mosier's (1943) formula for unweighted composites. Reliability for speeded tests was estimated using the test–retest method with a 1-day retest interval, and correlations were corrected for range restriction. The technical manual presents reliability coefficients for all test, cluster, and composite scores for each age 2–19 and the seven adult age groups. Median reliability coefficients were uniformly high: 38 of 39 were .80 or higher, and 17 were .90 or higher. Test–retest correlations for speeded tests were mostly in the .80 to .90 range. Cluster scores include two or more tests and as such produce higher reliability estimates as predicted by true score theory. Due to the higher precision in measurement it was recommended in the WJ IV technical manual that cluster scores (rather than subtest scores) be used in interpretation, especially when used in individual decision-making. It is clear that the approaches used by the test authors meet the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014); yet, one issue requiring attention from developers of the WJ IV and many other similar test batteries is that test scores conflate different sources of variance when hierarchically ordered, and model-based reliability coefficients (omega hierarchical) could be assessed to determine unique true score variance captured by the different scores (Brunner, Nagy, & Wilhelm, 2012; Canivez & Watkins, 2016; McDonald, 1999; Reise, 2012; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006). This factor is also critical in judging whether latent constructs are sufficiently precise for interpretation beyond the general intelligence estimate.

Because the WJ IV ACH includes three parallel forms, alternate forms equivalence was examined. Item difficulty graphs illustrated virtually identical difficulty of items across the three forms except for Test 8 (Oral Reading), which showed greater variability. Item content was also quite similar among the forms. Test characteristic curves also show very close correspondence between the forms.

Validity

Consistent with the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999, 2014), evidence for validity was structured around areas of test content, response processes, internal structure, and relations with other variables. WJ IV test content was from the outset selected to be consistent with the most recent articulation of CHC theory (Schneider & McGrew, 2012), and content of

the WJ IV ACH and WJ IV OL also included content specified in federal legislation. A change in the WJ IV from the WJ III was to include in the cluster scores "hybrid broad plus narrow" tests (technical manual, p. 120) rather than attempting to construct more pure CHC clusters. However, this reviewer wonders whether such a method negatively affected structural validity estimates. Classification of new WJ IV tests to the CHC classifications were provided by the test authors. Some evidence for validity was based on developmental changes across the life span through the use of cross–sectionally derived trends that paralleled expected developmental growth curves.

Validity based on internal structure utilized multidimensional scaling, the more traditional factor-analytic methods (exploratory factor analysis [EFA] and confirmatory factor analysis [CFA]), and cluster analysis. Full results of analyses were provided in the technical manual appendices. Full correlation matrices for age groups are also included in the technical manual appendices. EFA was reportedly problematic in that principal axis factor (PAF) extraction with oblique rotation (the appropriate method of EFA to examine the latent structure of correlated dimensions) apparently would not converge (although this concern may have been due to attempting to extract too many factors) and the test authors suggested the extensive coverage of WJ IV domains produced high multicolinearity and use of PAF with oblique rotation would require removing tests until properly estimated. Instead, test developers used principal components analysis (PCA) with varimax rotation (forcing the extracted factors to be uncorrelated) as a "practical decision" (technical manual, p. 156) and noted that oblique solutions were tested in CFA. Determination of the number of factors to extract was apparently influenced by the scree test (Cattell, 1966) and theory, although two alternate methods—parallel analysis (Horn, 1965) and minimum average partials (Velicer, 1976; Velicer, Eaton, & Fava, 2000)—are typically considered the most accurate approaches to guard against overextraction (Frazier & Youngstrom, 2007), and apparently were not used in these analyses. Results for 8, 9, and 10 extracted components were presented in the technical manual but of limited value given the extraction method (PCA) and the orthogonal rotation.

CFA results reported in the technical manual did not test or compare very many models (Model 1: single g–factor; Model 2: 9 broad CHC higher–order model; and Model 3: broad plus narrow CHC higher–order factor model). It seemed odd that Woodcock's Cognitive Performance Model (CPM; Woodcock, 1997) was not mentioned in the technical manual nor examined in CFA models despite support provided by Taub and McGrew (2014) with the WJ III. CFA results presented in the technical manual showed for each age group that the best fitting initial models and cross–validation models did not have comparative fit indexes (CFI; .603–.713) or Tucker–Lewis Indexes (TLI; .607–.684) that even approached the levels considered adequate (≥ .90; Hu & Bentler, 1999) and neither were the root mean square errors of approximation (RMSEA; .115–.123) (≤ .08; Hu & Bentler, 1999). So while the proposed final cross–validated models were the best among the models tested, they certainly were not well fitting models. However, the technical manual noted that WJ IV data were multivariate non–normal. Given multivariate

non–normal data, maximum likelihood (ML) estimates should not be trusted, and robust ML estimates should be used instead (Byrne, 2006, 2012). AMOS (Arbuckle, 2012), the structural equation modeling program apparently used for the WJ IV CFA, does not include robust estimators and thus does not handle multivariate non-normal data well. A note in the final two pages of the technical manual reported the use of Mplus (Muthén & Muthén, 1998-2011) to examine results from robust estimators. The Mplus MLM (robust estimator) produced CFI = .638, TLI = .610, and RMSEA = .116 for the final ages 9–13 MD model, which were quite close to the ML estimates previously reported and indicative of a model that did not fit well. Reported Mplus MLMV estimates were equally poor and seriously challenge the merits of the proposed WJ IV structure. Given superiority of Mplus and its use in examining robust estimators this reviewer wondered why AMOS was used at all.

Evidence of validity based on relations with other variables also was reported. The WJ IV COG was compared with six different contemporary intelligence tests with small to moderately large samples (Ns = 50-177), and the correlations between general intelligence scores were generally high, most in the .80s, and indicated concurrent measurement of general intelligence. Correlations at the factor and cluster level were generally lower. The WJ IV OL was compared with three other tests of language functioning with four small samples (Ns = 50-56), and most correlations were in the .60 to .70 range. The WJ IV ACH was compared to two different achievement batteries and a test of oral and written language with small samples (Ns = 49-51); composite score correlations were mostly in the .60 to .70 range within the separate reading, mathematics, and writing areas, but some were in the .80 and .90 range.

Group differences were assessed by examining the performance of small clinical samples (gifted, intellectual disability/mental retardation, learning disabilities [reading, mathematics, and writing], language delay, attention-deficit/hyperactivity disorder, head injury, and autism spectrum disorder) using only specific WJ IV tests, although not compared to matched samples of non-clinical groups. Inclusion criteria were specified in the technical manual, and it was noted that variation in classification or diagnostic criteria within each group was likely as children were previously classified. Results were generally in expected directions although the sample of gifted children had mean performance only approximately 1 standard deviation above the mean, which was somewhat lower than expected. Children with intellectual disabilities had mean performance well below 2 standard deviations below the mean. Given the small samples such results should be considered preliminary, and additional research on clinical group performance is needed.

COMMENTARY

The WJ IV technical manual included an enormous amount of information that contained some extremely useful descriptions and explanations of statistical methods employed. Such detail is welcome indeed although there are numerous analyses and summary statistics notably absent that would provide

test users extremely valuable and even necessary information about the relative contributions of various scores and comparisons that are supposed to guide clinical interpretation. Unfortunately, such absences are an all too common occurrence in test technical manuals, and users must await peer-reviewed research to provide such information to guide interpretations. Further, there appears to be little specific designation in the technical manual as to which statistical analysis program(s) was/were used to analyze data, which is critical because different programs may not use the same algorithms and can produce different results. Various references are included so one may infer which program may have been used in some situations, but the information should be more explicit. In the case of the CFA it was indicated in the note on the last two technical manual pages that Mplus was used and compared to results from AMOS, but by then results had already been presented and Mplus results likely deemphasized by their location.

In reviewing the literature on studies conducted on the WJ III, the authors of the test manual made no mention of recent publications challenging the structure of the WJ III that preceded publication of the WJ IV or were published near the same time (Dombrowski 2013, 2014a, 2014b; Dombrowski & Watkins, 2013). These studies raise serious questions about WJ III overfactoring and limited unique variance in WJ III broad ability clusters, a problem also likely to exist with the WJ IV, and that could have been reported in the WJ IV technical manual. Extraction (EFA) or specification (CFA) of fewer latent WJ IV first-order factors as suggested by these WJ III studies, as well as the poor EFA results and poor CFA model fit statistics reported in the WJ IV technical manual deserved attention. Recent research illustrating subsequent limitations of WJ III broad ability factors' incremental validity beyond the general intelligence factor (McGill, 2015; McGill & Busse, 2015) also confront the WJ IV and will need to be addressed in future research.

Gorsuch (1983) noted the complementary nature of EFA and CFA procedures and greater confidence when both converge on the same structure, but because the test developers used entirely different and incompatible methods for EFA and CFA, there can be no real examination of such possible convergence. The WJ IV technical manual noted the use of principal components analysis (considered by many [c.f., Fabrigar, Wegener, MacCallum, & Strahan, 1999; Gorsuch, 1983; Widaman, 1993] not to be “factor analysis”) rather than principal-factors analysis and used varimax (orthogonal) rotation (forcing extracted factors to be uncorrelated) rather than an oblique rotation, which this reviewer considers more appropriate in this instance. Such an analysis is considerably different from the higher–order models presented in the CFA, which implies a second–order dimension that explains the first–order factor correlations. The WJ IV technical manual noted that issues of the high degree of multicolinarity because “of the extensive coverage of certain domains in the WJ IV battery, can produce convergence problems when attempting the principal axis factor extraction method followed by oblique factor rotation” (technical manual, p. 156), so it must be assumed such non-convergence was produced. This non-convergence

may be an issue of attempting EFA with the entire set of tests from the COG, ACH, and OL batteries. This reviewer would have found the analyses more useful to have separate EFAs for each of the three batteries with the proper PFA and oblique rotation. [Editor's note: The test authors report that they initially followed this procedure of factoring several test batteries together at the recommendation of Jack Carroll. They believe that the results of the combined factor analysis has advantages, including eliminating many nuisance factors.] Another reason for the lack of convergence is likely that too many factors were selected to be extracted.

Despite invoking Carroll's work there was a failure to utilize and disclose information regarding techniques he insisted upon (Schmid & Leiman, 1957) that clarifies how much reliable subtest variance is due to the higher–order factors and that which is due to the lower–order factors (cluster scores), although that would require oblique rotation in EFA (Carroll, 1993, 1995). Perhaps subsequent research can examine the three WJ IV batteries separately.

The test developers did not use parallel analysis (Horn, 1965) or minimum average partials (Velicer, 1976; Velicer et al., 2000), two of the best methods for determining how many factors to extract and retain, but even the scree plots in the technical manual suggest four, rather than 8 to 10 or more factors to this reviewer. Although it is true that parallel analysis tends to suggest underfactoring (taking fewer factors) in the presence of a strong general factor such as in measuring intelligence (Crawford et al., 2010), understanding how well the additional extracted factors perform must be assessed by examining how much unique true score variance they measure apart from general intelligence and the other group factors. No variance estimates or model-based reliabilities are presented in the technical manual for users to determine adequacy of the factor-based scores, so abundant caution should be exercised in interpretation of lower–order scores until such research is produced. Model-based reliability coefficients (omega) should be provided as recommended by numerous psychometric experts (Brunner et al., 2012; Canivez & Watkins, 2016; Gignac & Watkins, 2013; McDonald, 1999; Reise, 2012; Reise, Bonifay, & Haviland, 2013; Zinbarg et al., 2005; Zinbarg et al., 2006). Omega-hierarchical would produce an estimate of how much true score variance is uniquely captured by general intelligence without the influence of the lower–order group factors (clusters), while omega–subscale would provide an estimate of how much true score variance is uniquely captured by the group factor (cluster) with the effects of general intelligence and all other group factors removed (Brunner et al., 2012; Reise, 2012). Without these estimates users of the WJ IV have no way to adequately determine how important the various factor-based cluster scores are and what interpretive value they might provide. Also, internal consistency estimates and resulting standard errors of measurement are based on methods that may be incorrect (Raykov, 1997) due to the hierarchical nature of measurement and the conflated sources of variance from general intelligence and broad abilities (cluster scores).

Another important limitation is that only higher–order structural models were examined although

research with the WJ III has also supported a rival bifactor model (Dombrowski, 2014a, 2014b) as an alternative structural model. Higher–order measurement models conceive of general intelligence (g) as a superordinate construct (Gignac, 2008) that has influences on subtests fully mediated through the first–order factors. Rival alternate bifactor (Holzinger & Swineford, 1937)/nested factor (Gustafsson & Balke, 1993; Keith, 2005)/direct hierarchical (Gignac, 2005, 2006, 2008) models that conceive of general intelligence as a breadth factor could have been tested and often are equally or better fitting. Gignac and others (i.e., Brunner et al., 2012; Canivez, 2016; Reise, 2012; Watkins, 2010) have made arguments regarding superiority of the bifactor model in that the general factor has direct subtest influences, is easy to interpret, both general and specific influences on subtests can be simultaneously examined, and it evaluates the psychometric properties necessary for determining scoring and interpretation of subscales (Canivez, 2016; Reise, 2012). While others hold that model comparisons should rely on more than statistical model fit, it was nevertheless disappointing for this reviewer to see that rival bifactor models were not tested against the higher–order WJ IV structural models. Also absent was assessment of factor invariance across sex, age, and race/ethnicity, evidence of fairness that is needed. It may be noted that the bifactor vs. hierarchical factor analysis debate is a continuing controversy, and no clear-cut or definite answer is yet determined. If hierarchical models are used, however, decomposed variance estimates for the higher- and lower-order factors should be presented in order to understand relative contributions.

Zero-order correlations were provided between Cognitive, Achievement, and Oral Language tests; however, because the first–order factor (cluster) scores conflate variance due to the higher–order factor (g) and the first–order cluster (broad ability) score, it is unknown how much of the correlation is due to g and how much is due to the broad ability (factor/cluster) scores. The absence of incremental validity examination (Hunsley, 2003; Hunsley & Meyer, 2003) via hierarchical multiple regression analyses or latent factor predictions of academic achievement or oral language scores with disclosure of how much variance is uniquely due to the higher–order g factor (GAI or Gf–Gc) versus the first–order factors (cluster/broad ability) is particularly troubling because users of the WJ IV have no ability to judge the relative merits of the general intelligence versus cluster based scores (c.f., Canivez, 2013; Glutting, Watkins, Konold, & McDermott, 2006; McGill, 2013, 2015; McGill & Busse, 2015). Other assessments of the utility of strengths and weaknesses generated by the WJ IV also need to be assessed to determine whether such strengths and weaknesses are sufficiently stable over time, whether they correctly identify independently classified diagnostic groups with which they should be associated (c.f., Swets, 1996; Treat & Viken, 2012), and whether such strengths and weaknesses lead to differential instructional benefits. Such reliability, validity, and diagnostic utility research is not included in the technical manual and must be examined through subsequent peer reviewed research.

In discussing variations of test association with latent factors, the test authors noted that most WJ IV OL

and ACH clusters were designed to measure "practical (school curriculum content distinctions), functional, or legal (e.g., federal SLD guidelines)" (technical manual, p. 175) aspects and were not designed as pure CHC factors. Perhaps this approach is a partial reason the CFA fit statistics did not indicate well fitting models (ML or robust MLM estimates). As such it may have been useful to provide separate EFA and CFA for the three batteries to understand their separate measurement characteristics.

SUMMARY

The WJ IV is the latest edition of a major test of intelligence (WJ IV COG), achievement (WJ IV ACH), and oral language (WJ IV OL). The three distinct batteries were co-normed using a large, demographically representative standardization sample. Examiner training checklists may help to facilitate development of proficiency of administration and scoring. Preliminary results show good reliability estimates, particularly for the general intelligence composite scores (BIA, GIA, Gf–Gc) and cluster scores. The WJ IV appears a good measure of general intelligence and provides useful measures of academic achievement, which may well be how the WJ IV will be primarily used. However, given the omissions and problems enumerated in this review and its commentary, there appear to be questions still needing answers about the measure, despite the appearance for completeness of the technical manual. Shortcomings in EFA and CFA methods presented in the technical manual indicate further research needs to be done to examine the latent factor structure of the three batteries separately as well as examining the possibility (likelihood) of fewer latent factors than proposed. Users of the WJ IV should also be provided with decomposed variance estimates in order to see how much variance is captured by the higher–order general intelligence dimension and what remains at the first–order factor level, in addition to being provided with estimates of latent factor reliabilities to adequately judge the merits of the provided scores. At present there is much information contained in the technical manual but much more needed to be provided, and major limitations and challenges exist as noted in this review. Clinical interpretation of the WJ IV COG beyond the general intelligence scores should be done with caution until further research supports various scores and comparisons. Users of the WJ IV must consider such future information in order to "(a) know what their tests can do and (b) act accordingly" (Weiner, 1989, p. 829).

REVIEWER'S 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.

Arbuckle, J. L. (2012). IBM SPSS AMOS 21 users guide. Crawfordville, FL: Amos Development.

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

Byrne, B. M. (2006). Structural equation modeling with EQS: Basic concepts, applications, and programming (2nd ed.). New York, NY: Routledge.

Byrne, B. M. (2012). Structural equation modeling with Mplus: Basic concepts, applications, and programming. New York, NY: Routledge.

Canivez, G. L. (2013). 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. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation. In K. Schweizer & C. DiStefano (Eds.), Principles and methods of test construction: Standards and recent advancements (pp. 247–271). Gottingen, Germany: Hogrefe.

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

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.

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

Crawford, A. V., Green, S. B., Levy, R., Lo, W.–J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. Educational and Psychological Measurement, 70, 885–901. doi:10.1177/0013164410379332

Dombrowski, S. C. (2013). Investigating the structure of the WJ-III Cognitive at school age. School

Psychology Quarterly, 28, 154–169. doi:10.1037/spq0000010

Dombrowski, S. C. (2014a). Exploratory bifactor analysis of the WJ-III Cognitive in adulthood via the Schmid–Leiman procedure. Journal of Psychoeducational Assessment, 32, 330–341. doi:10.1177/0734282913508243

Dombrowski, S. C. (2014b). Investigating the structure of the WJ-III Cognitive in early school age through two exploratory bifactor analysis procedures. Journal of Psychoeducational Assessment, 32, 483–494. doi:10.1177/0734282914530838

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

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

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

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., 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 WIAT–II. Journal of Special Education, 40, 103–114. doi:10.1177/00224669060400020101

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

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8, 206–213.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28, 407-434. doi:10.1207/s15327906mbr2804_2

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.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6, 1-55. doi:10.1080/10705519909540118

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. Psychological Assessment, 15, 443–445.

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

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (2nd ed., pp. 581–614). New York, NY: Guilford.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, 519-530.

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

McGill, R. J. (2013). Beyond g: Assessing the incremental validity of Cattell-Horn-Carroll (CHC) broad ability factors on the Woodcock-Johnson III Tests of Cognitive Abilities (Doctoral dissertation). ProQuest Dissertations and Theses, 238. (UMI No. 3621595)

McGill, R. J. (2015). Incremental criterion validity of the WJ–III COG clinical clusters: Marginal predictive effects beyond the general factor. Canadian Journal of School Psychology, 30, 51–63.

doi:10.1177/0829573514553926

McGill, R. J., & Busse, R. T. (2015). Incremental validity of the WJ–III COG: Limited predictive effects beyond the GIA–E. School Psychology Quarterly, 30, 353–365. doi:10.1037/spq0000094

Mosier, C. I. (1943). On the reliability of a weighted composite. Psychometrika, 8, 161–168. doi:10.1007/BF02288700

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus user's guide (6th ed.). Los Angeles, CA: Authors.

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

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

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

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (3rd ed., pp. 99–144). New York, NY: Guilford.

Schrank, F. A., & Dailey, D. (2014). Woodcock–Johnson online scoring and reporting [Online format]. Rolling Meadows, IL: Riverside.

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

Taub, G. E., & McGrew, K. S. (2014). The Woodcock–Johnson Tests of Cogntive Abilities III's Cognitive Performance Model: Empirical support for intermediate factors within CHC theory. Journal of Psychoeducational Assessment, 32, 187–201. doi: 10.1177/0734282913504808

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and

applications. Washington, DC: American Psychological Association.

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: Vol. 1. Foundations, planning, measures, and psychometrics (pp. 723–744). Washington, DC: American Psychological Association.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41, 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: Honoring Douglas N. Jackson at seventy (pp. 41–71). Norwell, MA: Kluwer Academic.

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

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. Journal of Personality Assessment, 53, 827-831.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? Multivariate Behavioral Research, 28, 263–311.

Woodcock, R. W. (1997). The Woodcock–Johnson Tests of Cognitive Ability–Revised. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 373–402). New York, NY: Guilford.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock–Johnson III. Rolling Meadows, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). Woodcock–Johnson III Normative Update. Rolling Meadows, IL: Riverside.

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 [omega hierarchical]. Applied Psychological Measurement, 30, 121–144. doi:10.1177/0146621605278814

Cite this review

Canivez, G. L. (2017). [Test review of Woodcock-Johnson® IV]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook*. Retrieved from http://marketplace.unl.edu/buros/

REVIEW 2 OF 2

Review of the Woodcock-Johnson IV by RONALD A. MADLE, Licensed Psychologist, Independent Practice, Mifflinburg, PA:

DESCRIPTION

The Woodcock-Johnson IV (WJ IV) consists of 50 individual tests arranged into three batteries that measure broad and narrow cognitive abilities including general intelligence, oral language characteristics, and academic skills and knowledge. It has been developed based mainly on contemporary Cattell-Horn-Carroll (CHC) theory.

Materials include five test easels in a carrying case and an overall technical manual. For each of the three batteries there are separate test booklets, response records, scoring guides, an audio CD, an examiner's manual, and an examiner's training guide.

The examiner's manuals include detailed test and cluster descriptions, administration and scoring directions, accommodations for impaired individuals, and explanations of obtained scores and interpretive procedures.

Administration is fundamentally unchanged from the Woodcock-Johnson III (WJ III; Woodcock, McGrew & Mather, 2001) with self-standing easels containing items and administration directions for presenting the items. Because all tests often are not administered, selective testing tables help determine which tests to administer to obtain desired cluster scores.

Administration of some tests is based on time or on item blocks, but most use a procedure in which testing is continued until a specified ceiling is reached following establishment of a basal. Most answers are given orally by the examiner or written into the response booklet. The examiner uses test booklets,

which include sections for identifying information and testing observations, to record answers. Administration time is about 35 to 40 minutes for the basic versions of each battery, but can take considerably longer if most tests are administered. Raw scores are transferred to a computer program for scoring.

The WJ IV Tests of Cognitive Abilities (WJ IV COG) measure broad and narrow cognitive abilities in addition to general intelligence. There is a Standard Battery, which can be used alone, and an Extended Battery. Individual tests that make up each battery (as well as the remaining batteries) are listed in the pre-matter to this review. Five tests and three clusters (Gf-Gc where Gf refers to Fluid Reasoning ability and Gc refers to Comprehension-Knowledge, Short-Term Working Memory-Extended, and Number Facility) have been added to this edition. Other tests have been rearranged, renamed, or eliminated.

The Standard Battery provides three estimates of overall ability: General Intellectual Ability (GIA), Brief Intellectual Ability (BIA), and Gf-Gc Composite). The Gf-Gc Composite can be valuable if the GIA might have been adversely affected by weaknesses in cognitive processing, memory, or speeded abilities, as is often the case in learning disability evaluations.

The other WJ IV COG broad abilities include Comprehension-Knowledge, Fluid Reasoning, Short-Term Working Memory, Cognitive Processing Speed, Auditory Processing, Long-Term Retrieval, and Visual Processing. Several other useful clusters (Cognitive Efficiency, Perceptual Speed, Quantitative Reasoning, Number Facility, Auditory Memory Span, and Vocabulary) have been incorporated. Finally, the WJ IV COG again includes six scholastic aptitude clusters, based on the four cognitive tests in each skill area best predicting near-term achievement.

The WJ IV Tests of Oral Language (WJ IV OL) provides cluster scores for Oral Language, Broad Oral Language, Oral Expression, Listening Comprehension, Phonetic Coding, and Speed of Lexical Access. Although the battery is new, most tests were previously included in other sections of the WJ III. Only Segmentation, which involves breaking words into parts and phonemes, is truly new.

Three of the nine English tests have parallel Spanish tests and corresponding cluster scores (Lenguaje oral, Amplio lenguaje oral, and Comprehensión auditiva). Administration can be performed by a bilingual examiner or by using a primary/ancillary examiner approach. Finally, Cognitive-Academic Language Proficiency scores can be useful for examining appropriateness of testing bilingual students in English.

The WJ IV Tests of Achievement (WJ IV ACH) consist of Standard and Extended Batteries, with a total of 20 tests. Various aspects of central academic skills—reading, mathematics, and writing, as well as academic knowledge–are measured. As with the WJ IV COG, a number of tests were rearranged and

revised and three new tests (Oral Reading, Reading Recall, and Word Reading Fluency) have been added. Several tests (Picture Vocabulary, Oral Comprehension, Understanding Directions, and Sound Awareness) were moved to the WJ IV OL. There now are three forms of the Standard battery for the WJ IV ACH to facilitate retesting without undue item exposure. The single Extended Battery is used with all three forms.

In addition to Broad Achievement and Brief Achievement scores, a number of clusters are included for reading, mathematics, and written language (e.g., Broad Reading, Broad Math). Basic skills clusters (e.g., Basic Reading Skills, Reading Fluency, Reading Rate, Math Calculation Skills, Basic Writing Skills) as well as applied skills clusters (e.g., Reading Comprehension, Math Problem Solving, Written Expression) are presented. Finally, there is an Academic Knowledge composite, derived from tests of Science, Social Studies, and Humanities, which can be examined in the four areas of Knowledge, Skills, Fluency, and Applications. There is a cluster for Phoneme-Grapheme Knowledge as well.

Another change is the use of an online scoring and reporting program. After an account has been established, multiple examiners can use the program with scoring credits being gained when test booklets are purchased. After locking in a test administration date, a varied set of derived scores (e.g., standard scores, confidence intervals, percentile ranks, relative performance indices [RPIs], and age- and grade-equivalents) can be obtained. The various reports (e.g., score report, parent report, profiles, class rosters), which can be printed in English or Spanish, have several scoring and score display options. The online scoring center also includes printable report and score interpretation guides for each battery.

Two types of comparative scores are presented. One type describes a pattern of intra-individual strengths and weaknesses; the second shows when obtained scores are outside a range of predicted scores. These scores can be obtained using various predictors, including the GIA and the Gf-Gc Composite, scholastic aptitude clusters, or even oral language scores.

DEVELOPMENT

The overall WJ IV revision goals presented in the technical manual include making advances in exploring individual strengths and weaknesses, complementing response to intervention models, reframing aptitude/ability achievement comparisons, pushing the boundaries of practical uses of CHC theory, and making improvements in ease and flexibility of test use.

Overall, eight new tests were added, and 14 tests had new items developed and included. These changes facilitated extending the test floors and ceilings as well as helped with item selection for the new WJ IV ACH form. All went through extensive pilot testing and were reviewed for content, sensitivity,

and bias. Items were calibrated using a Rasch model before completing tryout administration to 100 to 500 persons per test.

TECHINCAL

Standardization

A stratified random sampling procedure controlled for region, sex, race, ethnicity, birth country, community type, parent education, school/college type, educational attainment, employment status, and occupational level. Data presented suggest a good approximation to the 2010 U.S. Census population projections with only a few minor deviations.

Professional examiners who received additional training collected all standardization data. A multiple matrix sampling design was used to avoid having to administer the large number of tests to each examinee. Data collection covered the period from December 2009 through January 2012. Data were examined twice to make minor adjustments. When data were complete, item bias analyses (gender, white vs. non-white, and Hispanic vs. non-Hispanic) were completed, and flagged items were removed whenever possible.

The final data were obtained from 7,416 individuals, ranging from age 2 to age 90+, who came from varied settings across 46 states and the District of Columbia. Stated subsamples included 664 preschool children; 3,891 children in kindergarten through 12th grade; 775 college undergraduate and graduate students; and 2,086 adults. The dense sampling at school age was used because this age range is the period of greatest expected change. Items for the Spanish oral language tests also were calibrated with a sample of 1,413 native Spanish-speaking individuals.

U.S. Census-based subject weights were used to construct norms to account for minor sample variations. Various cluster scores were developed, including the GIA, which involved differentially weighting the seven tests comprising that cluster according to their relative contributions. All derived scores (e.g., standard scores, percentile ranks, RPIs, confidence intervals, age- and grade-equivalents) were then calculated.

The technical manual provides the age distribution of the norm sample at 1-year increments up to age 19. Sample sizes ranged from 173 to 336 with a mean group size of 278.67. In the seven adult groups the average size was 342.86 per group. Reported average sample sizes for Grades K-12 and 13-17+ were 299.31 and 155.00 per group, respectively.

Reliability

The median internal consistency coefficient (.97, with median coefficients of .95 to .97 for different age groups) for the WJ IV COG GIA score is very strong. The Gf-Gc reliability coefficients range from .94 to .98 (median = .96), while the BIA reliability coefficient median is .94 (.92-.95). Median coefficients for the various cognitive clusters range from .86 (Visual Processing) to .91 (Short-Term Working Memory). Oral Language reliability coefficients were similarly strong, ranging from .89 for Oral Expression to .92 for Broad Oral Language. Academic cluster median reliability coefficients were very strong, ranging from .92 to .97, with a Broad Achievement score median reliability coefficient of .99.

Test-retest studies were completed for speeded tests only, and median reliability coefficients ranged from .88 to .92 for three different age groups. The equivalence of item difficulties and item content for the achievement forms was examined and found to be strong.

Validity

The WJ IV technical manual includes diverse types of validity evidence. Content and construct validity evidence was supported using CHC research and theory, item examination, and test and cluster content descriptions. Extensive cluster analyses and exploratory and confirmatory factor analyses were completed. The final structural models supported the WJ IV battery organization as well as test and cluster score interpretation. These studies are discussed extensively in the technical manual.

Descriptive statistics showed that tests and clusters vary developmentally in manners that are consistent with accepted life span changes. For example, Comprehension-Knowledge and similar abilities increased until adult ages and then declined minimally with age, while other abilities showed greater declines after early adulthood into old age. The results were quite consistent with life span research on intellectual abilities (Baltes, Staudinger, & Lindenberger, 1999).

Tests and cluster scores for nine clinical groups were examined and provided more support for cluster validity. The clearest differences were between groups known to diverge in overall intelligence, such as scores for the intellectually disabled being lowest, scores for gifted individuals being highest, and scores for other groups being intermediate. In addition, groups of individuals with learning disabilities showed the weakest achievement consistent with their area of classification.

Fifteen studies generally provided strong concurrent evidence of validity of WJ IV scores. Scores from the WJ IV COG were correlated with scores from six major intelligence tests. The GIA (M = 107.2) was similar to the Full Scale IQ (M = 106.7) on the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV; Wechsler, 2003), with a .86 correlation. The BIA and Gf-Gc showed similar correlations (.83) but slightly lower mean scores (105.2 and 104.8). The Wechsler Adult Intelligence Scale—Fourth

Edition (WAIS-IV; Wechsler, 2008) Full Scale IQ (M = 107.1) similarly showed a correlation of .84 with the GIA (M = 104.3).

For the CHC-based Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004a), the GIA (M = 99.5) showed a moderately strong correlation (.77) with the Fluid-Crystallized Index (M = 100.3). The KABC-II Mental Processing Index, which has a different composition, was not as well correlated (.72) with the WJ COG measure. Finally, the Stanford-Binet Intelligence Scales, Fifth Edition (SB-5; Roid, 2003) (M = 100.0) correlated strongly (.80) with the GIA (M = 97.8).

Studies with two preschool-age tests—the Differential Ability Scales—Second Edition (DAS-II; Elliot, 2007) and the Wechsler Preschool and Primary Scale of Intelligence—Third Edition (WPPSI-III; Wechsler, 2002) could not examine full scale scores because the WJ IV COG does not report a GAI. These tests were examined at the cluster and test levels, as were the earlier tests. In all but one case the correlation matrices showed reasonably strong convergent-divergent validity evidence for the WJ IV COG. The exception was with the SB-5, which is consistent with other studies that have failed to replicate the SB-5's CHC structure (e.g., DiStefano & Dombrowski, 2006).

Five studies investigated the WJ IV ACH in relation to major achievement tests. Domain scores from the Kaufman Test of Educational Achievement, Second Edition (KTEA-II, Kaufman & Kaufman, 2004b) and the Wechsler Individual Achievement Test—Third Edition (WIAT-III; Wechsler, 2009) were consistent with WJ-IV scores and demonstrated moderate to strong convergent and divergent validity. Some scales, notably written expression, showed only moderate correlations (.52 to .62). WJ IV writing clusters showed moderate to strong (.65 to .75) correlations with the Written Expression score on the Oral and Written Language Scales: Written Expression (OWLS Written; Carrow-Woolfolk, 1996). Overall the WJ IV Broad Achievement correlations with these tests were strong (.85 to .91).

The WJ IV OL was compared to parts of the KTEA-II, and WIAT-III, as well as with several oral language tests (e.g., Peabody Picture Vocabulary Test, Fourth Edition [PPVT-4, Dunn & Dunn, 2007] and Comprehensive Assessment of Spoken Language [CASL, Carrow-Woolfolk, 1999]). Oral Language cluster correlations with the global oral language scores from these tests were in the .60 to .85 range, providing moderate to strong validity evidence as a measure of receptive and expressive oral language. Low to moderate correlations for the WJ IV OL Phonetic Coding and Speed of Lexical Access clusters, however, suggested they represent abilities perhaps not present on the other language tests. COMMENTARY AND SUMMARY. The WJ IV is likely to further the widely held opinion that the WJ is the best instrument available to measure CHC-based constructs. It is a carefully thought-out test that has been developed and standardized following contemporary test development standards. Substantial data support its reliability and validity, as well as its structural integrity. Although only limited changes are

evident at a surface level, significant advancements taken from CHC theory and research have been incorporated.

Some particularly strong features include the addition of the oral language battery, the various components aimed at test user training (e.g., examiner's training guides), attention to accommodations for impaired and developmentally young individuals, and prediction of near-term achievement using scholastic ability clusters. Finally, the new online scoring system has great potential but could lead to problems in some rural areas where Internet access may be slower and less reliable.

No test is perfect and a concern was noted. This issue is the difficulty in examining test and cluster floors, ceilings, and item gradients, as the norms are embedded in the scoring program. Limited inspection suggests, as with many tests, that use at the age extremes might provide unclear results. For example, when a sample examinee obtained only one correct item on the first seven WJ IV COG tests, the resulting standard scores at ages 3, 4, and 5 yielded only four (out of 21) standard scores less than 70. Many scores were average or below average. Adequate floors should give standard scores of 69 or lower in this situation (Bracken & Walker, 1997). As a result, other tests covering specifically the preschool age span, such as the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV; Wechsler, 2012) or the KABC-II, appear more psychometrically appropriate for preschool children, especially when assessing for developmental delays. [Editor's note: The test publisher advises that an early childhood version of the WJ IV, the Woodcock-Johnson IV Tests of Early Cognitive and Academic Development, is also available.]

REVIEWER'S REFERENCES

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. Annual Review of Psychology, 50, 471-507.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests with preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary Intellectual Assessment: Theories, Tests, and Issues (pp. 484-502). New York, NY: Guilford.

Carrow-Woolfolk, E. (1996). Oral and Written Language Scales: Written Expression. Torrance, CA: Western Psychological Services.

Carrow-Woolfolk, E. (1999). Comprehensive Assessment of Spoken Language. Torrance, CA: Western Psychological Services.

DiStefano, C., & Dombrowski, S. C. (2006). Investigating the theoretical structure of the Stanford-Binet—

Fifth Edition. Journal of Psychoeducational Assessment, 24, 123-136.

Dunn, L. M., & Dunn, D. M. (2007). Peabody Picture Vocabulary Test, Fourth Edition. San Antonio, TX: Pearson.

Elliot, C. D. (2007). Differential Ability Scales–Second Edition. San Antonio, TX: Pearson.

Kaufman, A. S., & Kaufman, N. L. (2004a). Kaufman Assessment Battery for Children, Second Edition. San Antonio, TX: Pearson.

Kaufman, A. S., & Kaufman, N. L. (2004b). Kaufman Test of Educational Achievement, Second Edition. San Antonio, TX: Pearson.

Roid, G. H. (2003). Stanford-Binet Intelligence Scales, Fifth Edition. Austin, TX: PRO-ED.

Wechsler, D. (2002). Wechsler Preschool and Primary Scale of Intelligence–Third Edition. San Antonio, TX: Pearson.

Wechsler, D. (2003). Wechsler Intelligence Scale for Children–Fourth Edition. San Antonio, TX: Pearson.

Wechsler, D. (2008). Wechsler Adult Intelligence Scale–Fourth Edition. San Antonio, TX: Pearson.

Wechsler, D. (2009). Wechsler Individual Achievement Test–Third Edition. San Antonio, TX: Pearson.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III. Rolling Meadows, IL: Riverside Publishing.

Cite this review

Madle, R. A. (2017). [Test review of Woodcock-Johnson® IV]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook*. Retrieved from http://marketplace.unl.edu/buros/