BUROS
CENTER FOR TESTING

# TEST REVIEWS

## Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition

# Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition

**Purpose**

Designed "for measuring the intelligence of children."

**Publication Dates**

1949-2012.

**Acronym**

WPPSI-IV.

**Administration**

Individual.

**Price Data, 2012**

$1,120 per complete box kit, including administration and scoring manual (2012, 342 pages), technical and interpretive manual (2012, 272 pages), 3 stimulus books, 25 of each record form (ages 2-6 through 3-11 and ages 4-0 through 7-7), 25 of each response booklet (1, 2, and 3), scoring keys, block set, puzzle set, zoo location set, and 2 ink daubers; $1,170 per complete box kit with 25 web-based score reports; $205 per administration and scoring manual; $205 per technical and interpretive manual; $79 per 25 record forms (ages 2-6 through 3-11); $99 per 25 record forms (ages 4-0 through 7-7); $47 per 25 response booklets (3, Animal Coding); $79 per 25 response booklets (1 and 2, Bug Search and Cancellation); $125 per stimulus book (1, 2, or 3); $3 per ink dauber; $129 per 25 record forms with web-based score report usages (ages 2-6 through 3-11); $149 per 25 record forms with web-based score report usages (ages 4-0 through 7-7).

**Author**

David Wechsler.

**Publisher**

Pearson.

a) 2-6 through 3-11 AGE BAND.

**Population**

Ages 2-6 through 3-11.

**Scores, 9 to 14**

Primary Index Scales: Verbal Comprehension Index (Receptive Vocabulary, Information, Total), Visual Spatial Index (Block Design, Object Assembly, Total), Working Memory Index (Picture Memory, Zoo Locations, Total); Ancillary Index Scales: Vocabulary Acquisition Index (Receptive Vocabulary, Picture Naming, Total), Nonverbal Index (Block Design, Object Assembly, Picture

Memory, Zoo Locations, Total), General Ability Index (Receptive Vocabulary, Information, Picture Naming [supplemental], Block Design, Object Assembly, Total); Full Scale IQ consists of Verbal Comprehension Index (with supplemental Picture Naming), Visual Spatial Index, and Working Memory Index (with supplemental Zoo Locations).

**Time**

[30-45] minutes

b) 4-0 through 7-7 AGE BAND.

**Population**

Ages 4-0 through 7-7.

**Scores, 12 to 25**

Primary Index Scales: Verbal Comprehension Index (Information, Similarities, Total), Visual Spatial Index (Block Design, Object Assembly, Total), Fluid Reasoning Index (Matrix Reasoning, Picture Concepts, Total), Working Memory Index (Picture Memory, Zoo Locations, Total), Processing Speed Index (Bug Search, Cancellation, Total); Ancillary Index Scales: Vocabulary Acquisition Index (Receptive Vocabulary, Picture Naming, Total), Nonverbal Index (Block Design, Object Assembly [supplemental], Matrix Reasoning, Picture Concepts, Picture Memory, Zoo Locations [supplemental] Bug Search, Cancellation [supplemental], Animal Coding [supplemental], Total), General Ability Index (Information, Similarities, Vocabulary [supplemental], Comprehension [supplemental], Block Design, Object Assembly [supplemental], Matrix Reasoning, Picture Concepts [supplemental], Total), Cognitive Proficiency Index (Picture Memory, Zoo Locations, Bug Search, Cancellation, Animal Coding [supplemental], Total); Full Scale IQ consists of Verbal Comprehension Index (with supplemental Vocabulary, Comprehension), Visual Spatial Index (with supplemental Object Assembly), Fluid Reasoning Index (with supplemental Picture Concepts), Working Memory Index (with supplemental Zoo Locations), and Processing Speed Index (with supplemental Cancellation and Animal Coding).

**Time**

[45-60] minutes.

**Cross References**

For reviews by Ronald A. Madle and by Merilee McCurdy and Lynae A. Johnsen, of the previous edition, see 16:267; see also T5:2864 (146 references) and T4:2941 (38 references); for reviews by Bruce A. Bracken and Jeffery P. Braden of an earlier edition, see 11:466 (118 references); for a review by B. J. Freeman, see 9:1356 (33 references); see also T3:2608 (280 references), 8:234 (84 references), and T2:538 (30 references); for reviews by Dorothy H. Eichorn and A. B. Silverstein, and excerpted reviews by C. H. Ammons and by O. A. Oldridge and E. E. Allison, see 7:434 (56 references).

# *Review of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition by GARY L. CANIVEZ, Professor of Psychology, Department of Psychology, Eastern Illinois University, Charleston, IL:*

## DESCRIPTION

The Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV) is the latest edition of the individually administered early childhood intelligence test with origins dating back to the Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 1967). The WPPSI-IV continues the traditional design and structure with numerous subtests each providing an estimate of general intelligence consistent with Wechsler's "global capacity" definition of intelligence (Wechsler, 1939, p. 229), but it also incorporates new and innovative methods purporting to measure dimensions consistent with more contemporary explications of intelligence structure. The WPPSI-IV is a major revision and re-standardization with many new and interesting tasks and changes that clinicians likely will appreciate. The WPPSI-IV is divided into two distinct age bands (2 years, 6 months to 3 years, 11 months and 4 years, 0 months to 7 years, 7 months) corresponding to different subtest batteries due to significant cognitive ability and developmental changes during the age range covered.

## DEVELOPMENT

Major revision goals noted in the WPPSI-IV Technical and Interpretive Manual included updating theoretical foundations, increasing developmental appropriateness, increasing user-friendliness, improving psychometric properties, and enhancing clinical utility. To accomplish these goals major changes included deleting subtests, adding new subtests, and modifying subtest content, administration, and scoring. Four WPPSI-III (T8:2911; Wechsler, 2002) subtests (Word Reasoning, Picture Completion, Symbol Search, and Coding) were deleted to reduce redundancy, to decrease emphasis on speed, and to make room for tasks that would better measure working memory and processing speed. Five new subtests were added, including Picture Memory, Zoo Locations, Bug Search, Cancellation, and Animal Coding. Application of contemporary theory was noted in the WPPSI-IV Technical and Interpretive Manual and the resulting structure articulated is a higher-order model with links to Carroll (1993, 2003, 2012), Cattell and Horn (1978), Horn (1991), and Horn and Blankson (2012) in what has come to be known as Cattell-Horn-Carroll (CHC; McGrew, 1997, 2005) theory. Measurement of intelligence from this model is illustrated by narrow ability subtests combining to measure various broad abilities that in turn combine to measure

general intelligence (Spearman, 1904, 1927).

WPPSI-IV stimuli are visually engaging, and several game-like activities that appear relatively easy to administer and score should be helpful in gaining and maintaining child cooperation and participation. To minimize fine motor control effects on processing speed tasks, an ink dauber was incorporated for children to quickly mark objects on response booklets. With this change, children are not required to manipulate a pencil. Core subtests associated with the Full Scale IQ score are administered first, followed by supplemental subtests and then optional subtests.

## TECHNICAL

### *Standardization*

The WPPSI-IV Technical and Interpretive Manual presents extensive and detailed information on the standardization procedures and the normative sample, which included 1,700 children divided into nine age groups. The standardization sample was obtained through stratified proportional sampling across key demographic variables of age, sex, race/ethnicity, parent education level (a proxy for socioeconomic status), and geographic region. Inspection of WPPSI-IV Technical and Interpretive Manual tables indicated a close match to the 2010 U.S. Census data. A list of exclusionary criteria also is presented and included, among other factors, language and communication limitations, visual or hearing impairments, upper extremity disability, and physical or medical conditions or medications that might impact test performance.

Subtest scaled scores (M = 10, SD = 3, range = 1 to 19) for each of 19 age groups were derived from an "inferential norming" (manual, p. 45) procedure. This procedure used raw score means, standard deviations, and skewness estimates that were examined from 1st through 4th order polynomial regressions. The best fitting model was selected with comparison to theoretical distributions and growth curve patterns that produced percentiles for each raw score. Minor irregularities were reportedly corrected through smoothing, but the method (statistical vs. hand/visual) was not specified. Composite scores for children ages 2:6 to 3:11 include the Full Scale IQ (FSIQ); primary index scores of Verbal Comprehension (VCI), Visual Spatial (VSI), and Working Memory (WMI); and Ancillary Index scores of Vocabulary Acquisition (VAI), Nonverbal (NVI), and General Ability (GAI). Composite scores for children ages 4:0 to 7:7 include the Full Scale IQ (FSIQ); primary index scores of Verbal Comprehension (VCI), Visual Spatial (VSI), Fluid Reasoning (FRI), Working Memory (WMI), and Processing Speed (PSI); and ancillary index scores of Vocabulary Acquisition (VAI), Nonverbal (NVI), General Ability (GAI), and Cognitive Proficiency (CPI). Composite scores (M = 100, SD = 15) were derived from sums of scaled scores from appropriately included subtests, and standard score distributions were visually smoothed to eliminate irregularities and ensure they were approximately normal. Some composite scores range from 45–155; others range from

40–160. Overall, a wide range (7.3–8.0 SDs) is represented.

Interpretation methods are delineated including reporting and describing the FSIQ and index scores and then evaluating index score strengths and weaknesses, index score pairwise comparisons, ipsative subtest strengths and weaknesses, and pairwise subtest comparisons, long a part of clinical tradition. Ability-achievement discrepancy analysis is discussed in assessment of specific learning disability, as is the pattern of strengths and weaknesses (PSW). No statistical analyses of reliability, validity, or diagnostic utility of these strengths or weaknesses, pairwise comparisons, or PSWs were provided to judge the value of such comparisons. Notwithstanding the numerous methods for clinical interpretation promoted in the Technical and Interpretive Manual and by others in textbooks and clinical guidebooks, there was a selective reporting of literature. Relevant empirical literature on the overwhelming shortcomings and failures of these ipsative and pairwise comparison methods to inform clinical practice (see Canivez, 2013a; Macmann & Barnett, 1997; Watkins, 2003; Watkins, Glutting, & Youngstrom, 2005) was notably absent and should be duly considered. Specific psychometric evidence for the WPPSI-IV interpretations demands strong empirical evidence for application to individuals but is absent.

***Reliability***

Three methods of estimating reliability of WPPSI-IV scores are reported in the Technical and Interpretive Manual: internal consistency, test-retest stability, and interscorer agreement. Internal consistency estimates were produced by Spearman-Brown corrected split-half coefficients for all subtests except Bug Search, Cancellation, and Animal Coding, as these are speeded tests where test-retest stability served as the reliability estimate. A table in the Technical and Interpretive Manual presents subtest, process, and composite score reliability coefficients for each of the nine age groups as well as average reliability coefficients across the age groups. Internal consistency coefficients across the nine age groups were .95–.96 for the FSIQ and ranged from .85 to .96 for index scores and from .71 to .95 for the subtest scores. Standard errors of measurement based on the reliability coefficients are also included in the Technical and Interpretive Manual but should be considered best-case estimates because they do not consider other major sources of error such as long-term temporal stability, administration errors, or scoring errors (Hanna, Bradley, & Holen, 1981) that influence test scores in clinical assessments. Short-term test-retest stability of WPPSI-IV scores was examined with a sample of 172 children in three age bands with retest intervals of 7–48 days (M = 23 days). Stability coefficients (uncorrected) for all ages were .88 for the FSIQ, .78–.88 for index scores, and .69–.81 for the subtests. Corrected coefficients were slightly higher. Mean differences across the retest interval were mostly small but reflected some practice effects as typically observed. All standardization sample record forms were double scored by two independent scorers and, as noted in the Technical and Interpretive Manual, the overall interscorer agreement was very high (.98–.99) due to most subtests containing simple and objective scoring. A random sample of 60 cases

was used to examine interscorer agreement for Information, Similarities, Vocabulary, Comprehension, and Picture Naming subtests where judgment is involved. Resultant intraclass correlations ranged from .96 to .99, reflecting very strong agreement in application of scoring criteria.

Estimated true score confidence intervals (90% and 95%) based on the standard errors of measurement are provided in the Administration and Scoring Manual, but the Technical and Interpretive Manual noted that there may be a preference for using the obtained score confidence interval. In fact, when the assessment question is concerned with estimating the true score of the individual at the time of the evaluation (rather than the long-term estimate), the obtained score confidence interval is appropriate (Glutting, McDermott, & Stanley, 1987; Sattler, 2008). Although a table of obtained score confidence intervals was not included in the manual, the formula for its calculation and an example were provided. For most WPPSI-IV scores, both confidence intervals will be quite close due to the generally high reliability indices.

### *Validity*

Consistent with Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), evidence for validity was structured around areas of test content, response processes, internal structure, relations with other variables, and consequences of testing. As anticipated, subtest intercorrelations were all positive for both age bands and reflected Spearman's (1904) positive manifold and measurement of general intelligence (g). Factor index score intercorrelations were also moderately high for both age bands as observed in other Wechsler scales and intelligence tests in general (Canivez, 2013a). Results from exploratory factor analyses (EFA) were not reported in the WPPSI-IV Technical and Interpretive Manual despite the fact that four former subtests were deleted and five new ones were added. Confirmatory factor analysis (CFA) preference was argued and evidence for the internal structure relied exclusively on CFA. CFA was examined separately for the two age bands due to different subtest compositions. One issue present in CFA modeling was the fact that with only two subtests representing Visual Spatial (VS) and Working Memory (WM) dimensions in the 2:6 to 3:11 age group and only two subtests representing VS, Fluid Reasoning (FR), and WM dimensions in the 4:0–7:7 age group, models appear only just identified. Including one additional subtest to each of these latent dimensions would provide for overidentification recommended in CFA. Based on the theoretical models reported for the 2:6 to 3:11 age band, results from CFA with all subtests supported the higher-order model with a general intelligence dimension and three broad first-order dimensions (Verbal Comprehension [VC], VS, and WM). Based on the theoretical models reported for the 4:0 to 7:7 age band, results from CFA with all subtests supported the higher-order model with a general intelligence dimension and four (traditional Wechsler model like the WISC-IV [Wechsler, 2003] or WAIS-IV [Wechsler, 2008a]) or five (CHC

inspired) broad first-order dimensions. Large differences between these two higher-order models were not apparent. To examine the latent structure of the WPPSI-IV further, CFA was used with only the primary index subtests for the two age bands. The respective higher-order g with first-order VC, VS, and WM dimensions (2:6–3:11) or first-order VC, VS, FR, WM, and Processing Speed (PS) dimensions (4:0–7:7) both produced well fitting models according to contemporary standards for goodness-of-fit statistics (Hu & Bentler, 1998, 1999).

Numerous small sample comparative studies of the WPPSI-IV in relation to other intelligence or neuropsychological tests were reported in the Technical and Interpretive Manual, and preliminary evidence for convergent and divergent validity was present. Correlations between subtest and composite scores from the WPPSI-IV and the Wechsler Individual Achievement Test-Third Edition (WIAT-III; 18:153; Pearson, 2009) were presented for generally small samples of children ages 4:0–7:6. Zero-order correlations between the WPPSI-IV FSIQ and WIAT-III subtest and composite scores were similar to those obtained with older children and adolescents who were administered the WISC-IV. Incremental validity (Hunsley, 2003; Hunsley & Meyer, 2003) of factor index scores was not reported. Small sample special groups and matched controls were compared to test for distinct group differences. Special groups included individuals identified with giftedness, mild and moderate intellectual disability, developmental delay-cognitive, developmental risk factors, attention-deficit/hyperactivity disorder, autistic disorder, and Asperger's disorder as well as some others. Mean differences were typically what one would expect with groups of individuals with giftedness scoring appreciably higher and children with disabilities or risk factors scoring lower than the matched control group. However, distinct group differences are a necessary but not sufficient condition for diagnostic utility and individual clinical application, and analyses examining diagnostic efficiency statistics (Kessell & Zimmerman, 1993) and/or receiver operator characteristic curves (Swets, 1996; Treat & Vicken, 2012) will need to be examined.

## COMMENTARY

In the Forward of the WPPSI-IV Administration and Scoring Manual, Alan Kaufman and Nadeen Kaufman quoted the WPPSI-IV research director as saying, "We wanted to give the clinician more insight with broader construct coverage, but not at the expense of him, her, or the child" (p. x). In their opinion, this goal was achieved. While there are many notable changes and improvements including new engaging subtests that are geared to early childhood, ease of administration and scoring, the outstanding large and representative standardization sample, strong estimates of score reliability, and some preliminary evidence for validity, there are a number of important and disconcerting omissions in the WPPSI-IV Technical and Interpretive Manual that should have been included and must be addressed in the future. Some omissions are the same as those pointed out in a review of the WAIS-IV (Canivez, 2010).

With the deletion of four WPPSI-III subtests and the addition of five new subtests, this reviewer considered

it disappointing to see, as with the WAIS-IV (Wechsler, 2008b), that EFA results were not reported. While CFA is informative and tests theory, EFA and CFA answer different questions, and Gorsuch (1983) noted the complementary nature of the two approaches and general confidence in the latent structure when both were in agreement. It would have been useful to have seen evidence for factor extraction based on Cattell's scree test (Cattell, 1966), Horn's parallel analysis (Horn, 1965), and minimum average partials (Velicer, 1976), which help to guard against overextraction observed by Frazier and Youngstrom (2007). Further, presentation of oblique rotation factor pattern and structure coefficients for subtests would illustrate the degree to which subtests were uniquely associated with their theoretically postulated dimension. Presentation of variance accounted for by extracted factors would also assist clinicians in judging the merits of the factors. Absence of EFA results indicates that WPPSI-IV data were not allowed to speak for themselves. Failure to present Schmid and Leiman (1957) apportioned subtest variance to the general intelligence dimension and to the three, four, or five first-order dimensions (as insisted by Carroll, 1995) does not allow WPPSI-IV users the opportunity to judge for themselves whether sufficient variance is captured by index score factors for interpretation beyond the FSIQ.

Schmid and Leiman decomposed variance estimates were also absent from the CFA presentation. Such analysis results would have illustrated what subtest variance is associated with the higher-order g factor and what remains in the first-order factors. This decomposition is necessary to understand the higher-order dimensionality better. Also disappointing was the exclusive presentation of only higher-order measurement models that conceive general intelligence 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 should have been tested and often are equally or better fitting. Gignac and others (i.e., Brunner, Nagy, & Wilhelm, 2012; Reise, 2012; Watkins, 2010) have made compelling arguments regarding superiority of the direct hierarchical (bifactor/nested factors) model in that the general factor having direct subtest influences is easy to interpret, both general and specific influences on subtests can be simultaneously examined, and psychometric properties necessary for determining scoring and interpretation of subscales can be evaluated. The direct hierarchical model can also be considered a more parsimonious model (Gignac, 2006). Also missing is estimation of latent factors reliability from CFA by calculating coefficient omega and omega hierarchical (Brunner et al., 2012; McDonald, 1999; Reise, 2012; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006), which is also critical in judging whether latent constructs are sufficiently precise for interpretation beyond the g estimate (FSIQ).

Absent also were hierarchical multiple regression analyses examining the relationships between WPPSI-IV index scores and WIAT-III scores after the effects of the FSIQ were accounted for to test the incremental validity of factor index scores as illustrated by Glutting, Watkins, Konold, and McDermott (2006) with the

WISC-IV or Canivez (2013b) with the WAIS-IV. Understanding the relative contribution of FSIQ and index scores in predicting achievement is necessary in determining interpretive weight of FSIQ and lower-order index scores.

## SUMMARY

While the WPPSI-IV is an outstanding revision overall with many positive attributes and much of the WPPSI-IV Technical and Interpretive Manual can be considered a model for presentation of important information, like the WAIS-IV Technical and Interpretive Manual (Wechsler, 2008b), it falls short of fully disclosing critical statistical analyses and results necessary for clinicians to judge the adequacy of provided scores and interpretation methods. Such information is readily available. This reviewer believes that there can be no justifiable rationale for not including key pieces of information such as proportion of variance apportioned to the FSIQ, factor indexes, and subtests from EFA with Schmid and Leiman (1957) transformation, proportions of variance explained in subtests by the latent g dimension and latent first-order factors from CFA, or factor index score incremental prediction of WIAT-III achievement scores beyond the FSIQ. These results are critical in guiding clinicians in their selection and utilization of the WPPSI-IV. As such clinicians must wait for such information to emerge from the extant literature. In the meantime if following the methods of interpretation articulated in the Technical and Interpretive Manual, they risk overinterpretation and misinterpretation of WPPSI-IV scores in clinical application. Empirical evidence for the WPPSI-IV clinical interpretation schemes (ipsative strengths and weaknesses, pairwise comparisons, PSW) was also lacking. Perpetuation of these subtest and profile analyses continues clinical test interpretation tradition that appears more a shared professional myth of subtest and profile utility and belief without empirical evidence. Subtest and profile interpretation methods are not consistent with Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) and should not be used in clinical decision-making until psychometric support for them is provided. In the words of Weiner (1989), the ethical psychologist will "(a) know what their tests can do and (b) act accordingly" (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.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. Journal of Personality, 80, 796–846.

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. (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

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

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. Multivariate Behavioral Research, 30, 429–452.

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 Press.

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 Press.

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

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.

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.

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

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.

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

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.

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.

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

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

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.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. Psychometrika, 2, 41–54.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185.

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 Press.

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

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, 5, 1–55.

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.

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

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.

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

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.

***Pearson***

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

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47, 667–696.

Sattler, J. M. (2008). Assessment of children: Cognitive foundations (5th ed.). San Diego, CA: Author

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

Spearman, C. (1904). “General intelligence”: Objectively determined and measured. American Journal of Psychology, 15, 201–293.

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

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

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

Velicer, W. F. (1976). Determining the number of components form the matrix of partial correlations. Psychometrika, 41, 321–327.

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. (2010). Structure of the Wechsler Intelligence Scale for Children–Fourth Edition among a national sample of referred students. Psychological Assessment, 22, 782–787.

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.

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

Wechsler, D. (1967). Wechsler Preschool and Primary Scale of Intelligence. New York, NY: The Psychological Corporation.

Wechsler, D. (2002). Wechsler Preschool and Primary Scale of Intelligence-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, Inc.

Wechsler, D. (2008b). Wechsler Adult Intelligence Scale—Fourth Edition technical and interpretive manual. San Antonio, TX: NCS Pearson, Inc.

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

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.

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 . Applied Psychological Measurement, 30, 121–144.

## Cite this review

Canivez, G. L. (in press). [Test review of 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*. Retrieved from http://marketplace.unl.edu/buros/

REVIEW 2 OF 2

# *Review of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition by TRACY THORNDIKE, Associate Professor of Special Education and Education Leadership, Western Washington University, Bellingham, WA:*

## DESCRIPTION

The Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV) is an individually administered intelligence test designed for use with children ages 2 years, 6 months (2:6) to 7 years, 7 months (7:7). The WPPSI-IV is a substantially revised version of the Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; T8:2911). As with the WPPSI-III, the WPPSI-IV is divided into two age bands (2:6 to 3:11 and 4:0 to 7:7) with the second age band extended upward to overlap more fully with the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; 16:262). The scale includes 15 subtests with different batteries of subtests used for each age band.

Scores can be interpreted at three levels: full scale, index scale, and individual subtest. Scores on multiple

core subtests are combined to yield up to five primary, factor-based composite index scores reflective of different aspects of intellectual functioning and a second-order factor-based composite score representing general cognitive ability (i.e., g or Full Scale IQ [FSIQ]). Available supplemental and optional subtests allow for a broader sampling of specific intellectual skills and make possible the computation of up to four additional theoretically based ancillary index scores. The breadth and depth of information provided by the primary and ancillary index scores, in the hands of a skilled user, could increase the clinical utility of the WPPSI-IV as a component of a comprehensive neuropsychological assessment of cognitive and/or adaptive functioning in young children.

## DEVELOPMENT

The Wechsler intelligence scales—the Wechsler Adult Intelligence Scale (WAIS), Wechsler Intelligence Scale for Children (WISC), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI)—are among the instruments most widely used to measure intellectual functioning in English-speaking populations and have been revised frequently over the past seven decades. The WPPSI-IV is newest revision of the scale designed to measure intelligence in young children.

Goals of this revision were to: (a) update the scale to reflect more accurately current structural models of intelligence, theories of working memory, and findings from neurodevelopmental and neurocognitive research; (b) increase the developmental appropriateness and user-friendliness of the materials and administration and scoring procedures; (c) improve the psychometric properties of the scale; and (d) enhance the clinical utility of the scale for use in various types of psychological evaluation. The scope of these goals necessitated substantial modifications to the scale including removal of four subtests, addition of five new subtests, and revision of item content and administration and/or scoring procedures on all subtests retained from the WPPSI-III.

The major alterations to subtests for the WPPSI-IV involved those measuring working memory (two new subtests) and processing speed (three new subtests). Both working memory subtests now employ proactive interference rather than sequencing as the means of increasing cognitive processing demands. The new processing speed subtests were designed to be more developmentally appropriate than the subtests they replaced. Specifically the tasks are more play-like, realistic stimuli are favored over abstract stimuli, and level of fine motor skill needed to generate a response quickly has been decreased through use of an ink dauber instead of a pencil.

Multiple rounds of revision, piloting, and refinement, all clearly described in the test manual, were conducted to arrive at the WPPSI-IV. Major revisions to items, administration procedures, and scoring were first tested in pilot studies before moving to a national tryout phase of development.

## TECHNICAL

Normative information for the WPPSI-IV is based on a stratified sample of 1,700 children ages 2:6 to 7:7 years. The sample is representative of the U.S. population of English-speaking children in this age range in terms of race/ethnicity, parent education level, and geographic region and includes an equal number of males and females. The sample was divided into nine age groups with 200 children each for the 2:6–2:11 through 6:0-6:11 groups and 100 children for the 7:0–7:7 group.

Scores on individual subtests are scaled to a metric with a mean of 10 and a standard deviation of 3. Composite scores are scaled with a mean of 100 and a standard deviation of 15. Age equivalent scores are provided for 1-month intervals across the full 2:6 to 7:7 age range. Percentile ranks, standard errors of measurement, and confidence intervals also are available to aid interpretation of scores.

The reliability of scores on the individual subscales and composites was examined separately. Internal consistency for all scores except those measuring processing speed was estimated using the split-half method later adjusted with the Spearman-Brown formula to estimate reliability of the full-length scales. Reliability coefficients ranged from .85 to .93 for the subscales and from .86 to .94 for the composite scores. Reliability estimates for the processing speed subtests were based on test-retest data and ranged from .75 to .83. The values were somewhat lower than the other subtests, but still in an acceptable range given that they were based on test-retest data.

A subset of the standardization sample (N = 172) took the WPPSI-IV on two occasions to estimate score stability over time. Test-retest reliability of subtest and composite scores was calculated for three age ranges (2:6–3:11, 4:0–5:5, and 5:6–7:7) over an interval of an average of 23 days. Reliability coefficients were in the .70s and .80s for the subtests, .80s for the primary index scores, and over .90 for the FSIQ score indicating acceptable to excellent levels of stability over time for all age ranges. Evidence of interscorer agreement is also presented for the five subtests that require judgment in scoring. A sample of 60 cases was randomly selected from the norming sample and scored independently by nine scorers with no prior experience with the WPPSI-IV scoring rules. Interrater reliability coefficients ranged from .96 to .99, indicating that these subtests can be reliably scored.

Extensive validation evidence is provided to support proposed interpretations and uses of scores from the WPPSI-IV. Expert and advisory panel reviews were employed at each stage of development to help ensure that test content adequately sampled all relevant domains of intellectual functioning and that items and administration procedures were developmentally appropriate. Evidence describing the internal structure of the instrument is available for each of the two age bands. Confirmatory factor analysis was used to generate and test a variety of structural models. A model with three first-order factors (Verbal Comprehension, Visual Spatial, and Working Memory) and one second-order factor (g) best fit the data

for the 2:6 to 3:11 age band. For the 4:0 to 7:7 age band, a model with five first-order factors (Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed) and one second-order factor (g) best reproduced the observed correlations among subtest scores. Both the WISC-IV and WAIS-IV have a four-factor structure, but the test manual notes that evidence collected since the publication of these scales is more compatible with the five-factor model observed on the WPPSI-IV for the older age band.

Strong evidence of convergent concurrent validity was demonstrated by the strength of relationship between scores on the WPPSI-IV and scores on multiple other measures of cognitive ability (e.g., WPPSI–III, Bayley Scales of Infant and Toddler Development–Third Edition [Bayley–III; 17:17], and Differential Ability Scales–Second Edition [DAS–II; 18:45]). Additional validation evidence in the form of correlations with scores on tests of achievement, executive functioning, and emotional and behavioral functioning commonly used clinical application in conjunction with the WPPSI-IV is also presented for 13 "special groups" of children. Although based on relatively small, nonrandom samples, the "special group" evidence generally supports the clinical utility of the WPPSI-IV as part of a comprehensive diagnostic evaluation of cognitive functioning in young children.

## COMMENTARY

The WPPSI-IV is a psychometrically strong instrument that yields scores that estimate both overall cognitive ability and more specific intellectual skills. New norms are based on a large, representative sample of the U.S. population of English-speaking children ages 2 years 6 months to 7 years 7 months. Extensive evidence is provided to support both the reliability of scores and the validity of a wide range of score-based inferences. The test manual includes detailed information about administration, scoring, and score interpretation presented so clearly that even inexperienced test users should find it easy to understand. Of particular note is the chapter in the manual on interpretative considerations with step-by-step instructions for interpreting and reporting scores and important cautions that, if heeded, would decrease dramatically the likelihood of score misinterpretation and misuse. The major goals of this revision–updating the scale to reflect current structural models of intelligence, increasing developmental appropriateness and user-friendliness, improving psychometric properties, and expanding the evidence base supporting clinical application–have largely been achieved.

## SUMMARY

The WPPSI-IV is a substantially revised version of the Wechsler Preschool and Primary Scale of Intelligence, an individually administered test battery used to measure cognitive ability in young children. The structure of the test has been updated to reflect accurately current theoretical conceptualizations of intellectual functioning. Changes to items, subtests, and administration and scoring procedures have

resulted in a more developmentally appropriate instrument that is psychometrically stronger than its predecessors. Evidence supporting the use of the WPPSI-IV with a wide variety of groups enhances the clinical utility of the scale for diagnostic and intervention planning purposes.

## Cite this review

Thorndike, T. (in press). [Test review of 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*. Retrieved from http://marketplace.unl.edu/buros/