Journal of Psychoeducational Assessment
2004, *22*, 261-274

TEST REVIEWS

Woodcock, R. W., McGrew, K., & Mather, N. (2001) *The Woodcock-Johnson Tests of Achievement: Third Edition.* Itasca, IL: Riverside.

GENERAL DESCRIPTION

The Woodcock-Johnson Tests of Achievement: Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2001) is designed for children, adolescents, and adults ranging from 2 through 90 plus years. The two forms of the test (A & B) each consist of two batteries: a Standard Battery (tests 1-11, supplemental test 12, and two writing scales) and an Extended Battery (tests 13-19 and supplemental tests 20-22). The purpose of the Extended Battery is to further assess academic areas to determine students' strengths and areas of difficulty.

Table 1
Organization of WJ-III Tests of Achievement

Academic Area	Standard Battery	Extended Battery
Reading	Letter-Word Identification	Word Attack
	Reading Fluency	Reading Vocabulary
	Passage Comprehension	
Oral Language	Story Recall	Picture Vocabulary
	Understanding Directions	Oral Comprehension
Mathematics	Calculation	Quantitative Concepts
	Math Fluency	
	Applied Problems	
Written Language	Spelling	Editing
	Writing Fluency	
	Writing Samples	
Academic Knowledge		Academic Knowledge
Supplemental	Story Recall-Delayed	Spelling of Sounds
	Handwriting Legibility	Sound Awareness
	Writing Evaluation	Punctuation & Capitalization

Materials for the test consist of an Examiner's Manual that describes administration and scoring procedures and a Technical Manual that describes test development and technical adequacy. A Test Record Form is used by examiners to record students' responses, and a Subject Response Book is used by students for their written responses. The Compuscore and Profile Program is necessary for scoring, and a cassette tape is needed to present items for five subtests. A Standard Test Book and an Expanded Test Book, both with an easel format, are used to present test items.

Either age- or grade-based norms can be used to interpret results, which can be advantageous. For example, if results from the achievement test are to

be compared with results from intelligence tests, age-based norms might be selected because norms for intelligence tests are based on age. However, if a student's results are to be compared with those of other students in the same grade, regardless of his or her chronological age (e.g., if a student has been held back in school), then grade-based norms might be selected for this purpose.

Numerous derived scores are possible, including age and grade equivalents, percentiles, W difference scores (difference between student's score and the average score for the reference group in the norm sample), Relative Proficiency Index (e.g., an RPI of 60/80 indicates that the norm group demonstrated an 80% proficiency, whereas the student tested displayed 60% proficiency), an Instructional Zone indicating whether skills would be easy or difficult for the student, Cognitive-Academic Language Proficiency (which reflects how difficult the student will find English Language demands at different levels), and standard scores ($M = 100$, $SD = 15$), z scores, T scores, stanines, and normal curve equivalents (NCEs). Only age or grade equivalents can be obtained without the use of the Compuscore and Profile Program; consequently, norms are not available for inspection. However, because of psychometric limitations, the American Psychological Association (APA, 1999) and other organizations advise against the use of age or grade equivalents. Anastasi and Urbina (1997) as well as Salvia and Ysseldyke (2004) provide detailed discussions of the concerns associated with use of these derived scores. The type of derived score usually considered the most psychometrically satisfactory is a standard score (Anastasi & Urbina, 1997). Nonetheless, most test authors still offer these age- and grade-based equivalent scores because of their intuitive appeal to nonpractitioners.

According to the manual, the Standard Battery takes about 60 to 70 minutes to administer. The Writing Samples subtest requires approximately 15 to 20 minutes and the remaining subtests each take 5 to 10 minutes. The administration time varies depending upon the number of subtests used. The only timed subtests are those that assess fluency in the areas of reading, math, and writing.

DESCRIPTION OF SUBTESTS

Standard Battery

Letter-Word Identification. There are 76 items on this subtest. Reading words is not tested until item 15. The first 14 items assess the ability to name letters rather than give their sounds, and it is knowledge of letter sounds that is the skill useful in decoding words. If only the first 14 items or fewer are administered, test results would reflect letter naming and not a student's ability to identify letter sounds or words. The majority of the remaining items include phonetically regular words or irregular words that appear frequently in print. However, several items at the upper end of the subtest are phonetically irregular and cannot be decoded using phonics. Also, these are not words that should be recognized as sight words because they appear infrequently in print. Consequently, these items are less helpful for assessing reading skills.

Reading Fluency. On this subtest, students read as many of the 98 sentences as they can within 3 minutes and then circle yes or no to indicate whether each statement about the material read is correct or not. Reading speed and comprehension are assessed. This subtest requires reading isolated sentences. This format is not the same as that encountered in most reading tasks that require reading passages for meaning and that allow use of passage context cues to aid word recognition as well as comprehension. We could find no information in the manual to indicate how results from this subtest correlate with results obtained from reading passages.

Passage Comprehension. This 47-item subtest assesses students' understanding of what they read. Rather than assessing understanding of words read, the first four items require matching a picture with a corresponding line drawing. Because neither words nor passages are used, these items seem inappropriate for this subtest; presumably they assess some prerequisite skill. For most of the remaining items a cloze format is used; i.e., examinees supply missing words in a passage. In our experience, picture cues may confound results on seven items.

Story Recall. Students listen to an audiotape for this subtest, which requires that they listen to as many as 10 stories that increase in complexity. Then students are asked to recall as many story details as possible. Both memory and listening comprehension are required.

Understanding Directions. Students listen to directions presented on an audiotape and then are asked to point to items in a picture that were mentioned on the tape. The number of items that students are required to point to gradually increases. Although understanding directions is an important classroom skill, the directions on this subtest seem atypical of (i.e., not like) those used in classrooms; consequently, generalization to classroom behavior is problematic. This subtest appears to assess memory more than understanding of meaningful directions.

Calculation. The 45 items on this subtest range from writing single numbers to logarithmic and calculus operations. Considering the wide age range of the WJ-III (to 90+ years), the 45 items on this subtest provide only a limited sample of skills in this area. For example, for whole numbers only one addition item on Form A and two on Form B require regrouping, one subtraction item on Form A and two on Form B require regrouping, and the largest multiplication problem on both forms is two digits by one digit.

Math Fluency. Knowledge of 160 addition, subtraction, and multiplication facts through 10 is assessed within a 3-minute time limit. Fluency with math facts is an important skill to assess and, unfortunately, is one that is often overlooked on other tests.

Applied Problems. The 63 items on this subtest tap students' ability to apply their math skills to solve orally presented problems. Eleven of the first 12 problems require counting pictured objects; the remaining 51 range from simple story problems to items that require knowledge of probability and algebra.

Spelling. The 59 items on this subtest assess students' ability to draw marks and letters and spell dictated words. Written spelling necessarily involves writing words; however, writing words is not required on this subtest until item 15 on Form A and item 14 on Form B because the preceding items require

prewriting skills such as making marks or writing individual letters. Again, considering the age range of the test and the fact that the first 14 items on Form A (13 on Form B) assess prespelling skills, the remaining 45 items on A (46 on B) provide only a limited sample of spelling skills.

Writing Fluency. On this 40-item subtest, students are shown pictures one at a time and asked to write a sentence within 7 minutes about each picture using three given words. Spelling, punctuation, or capitalization mistakes are not counted as errors because the focus is on speed. To receive credit, students must write complete sentences related to the pictures and incorporate the given words. Thus, speed of writing sentences is assessed.

Writing Samples. The 30 items on this subtest each require students to write a sentence; some items have picture prompts and others use words.

Story Recall-Delayed (Supplemental). The same items are administered for this subtest as for Story Recall, except that this subtest is given in a delayed manner from 30 minutes to 8 days after administration of the Story Recall subtest. The range in time delay for administering this subtest seems curious; we would anticipate that students would remember the stories better if this subtest were administered 30 minutes following the Story Recall subtest than they would 8 days later. We could find no data in the manual to support the range of delay of 30 minutes to 8 days or to suggest that a delay of 30 minutes and a delay of 8 days yield similar results.

Handwriting Legibility (Supplemental). Samples of students' handwriting are compared to writing samples in the manual. Scoring handwriting has always been subjective and is on this subtest as well, although helpful guidelines are provided. A useful Handwriting Elements Checklist on the record book also may be used to rate aspects of handwriting such as slant and spacing.

Writing Evaluation (Supplemental). Samples of students' writing are scored analytically by using an evaluation rating scale in the manual to note student strengths and problems such as spelling, sequencing of ideas, use of complete sentences, and inclusion of major details. Several samples of a student's written stories or essays can be used to complete the rating.

Extended Battery

According to the manual, the purpose of the Extended Battery is to determine students' strengths and areas of difficulty for the purpose of planning instruction, although none of the subtests provide truly comprehensive assessment. Typically, each skill is assessed with one item, which does not provide a sufficient sample of performance for instructional planning. Consequently, results from the Extended Battery subtests yield general information regarding areas that may require instruction (e.g., vocabulary or word attack), but supplemental measures will be required to obtain the information on students' specific strengths and difficulties needed for planning comprehensive instruction.

Word Attack. Except for the first 3 items, the 32 items on this subtest assess students' ability to decode nonsense words. This is a useful format for assessing word attack skills because students would not encounter words they have learned to recognize by sight. However, the first 3 items involve letters rather than nonsense words and may not be helpful for assessing this skill.

Reading Vocabulary. Three sections make up this subtest: Synonyms (26 items), Antonyms (26 items), and Analogies (21 items). According to the manual, "Low performance on Reading Vocabulary may be a function of limited basic reading skills, limited word comprehension or both" (p. 82). However, on Synonyms, for example, a student may read and understand the meaning of a word but not be able to supply another word that means the same thing as the word presented. In this case, the student would not receive credit for the item even though he or she was successful in reading and comprehending the vocabulary word. Thus, cognitive functions in addition to reading and understanding words (i.e., supplying similar words, words that mean the opposite, and words that form an analogy) are required for success on this subtest.

Picture Vocabulary. This 44-item subtest covers expressive vocabulary. Students are asked to name pictures. Only knowledge of nouns is tested, and we could find no rationale as to why these particular words were selected.

Oral Comprehension. This 34-item listening subtest requires that passages be presented on an audiotape. Students listen to the tape and are asked to supply a missing word.

Quantitative Concepts. Two sections comprise this subtest: Concepts and Number Series. At the beginning of the 34-item Concepts section, counting, number identification, shapes, and sequences are covered; later items cover mathematical terms and formulas. The 23 items on the Number Series section require students to recognize a pattern in a number series and supply missing numbers in the series. Because two areas are assessed on this subtest (concepts and reasoning), it is difficult or impossible to determine if a low score reflects problems with one or both areas.

Editing. This subtest consists of 34 items and requires that students correct errors in spelling, capitalization, punctuation, or word usage in a written passage. Most rules are assessed only once. Although it is useful to assess these skills, planning for comprehensive instruction will require additional assessment. On Form A, 8 items assess capitalization rules, 7 assess punctuation rules, 13 assess word usage, and 6 require spelling skills. For students from 1st through 12th grade, the 6 spelling items provide a limited sample of skills. On Form B, 5 items require capitalization, 6 assess punctuation, 13 cover word usage, and 10 assess spelling.

Academic Knowledge. Three sections make up this subtest: Science, Social Studies, and Humanities. These are not areas typically assessed on individually administered achievement tests, and we commend the authors for including them. On initial items, students point to indicate a response; on later items, they must give oral responses. Science consists of 28 items, Social Studies 28, and Humanities 22. Whether the skills tested are representative of commonly used curricula is not addressed, and the number of items per section is limited considering the age range of the test.

Spelling of Sounds (Supplemental). The 28 items for this subtest are presented on audiotape. The first 5 items present single letters. Consequently, spelling items begin with item 6 and involve nonsense or low-frequency words that use regular patterns in English spelling. Use of nonwords and low-frequency words circumvents the problem of students memorizing the spelling of words rather than using phonics.

Sound Awareness (Supplemental). Items on this subtest for phonological awareness are presented on an audiotape. The sections of this subtest are: Rhyming (17 items), Deletion (10 items), Substitution (9 items), and Reversal (9 items).

Punctuation & Capitalization (Supplemental). Items on this 36-item subtest require students to apply capitalization and punctuation rules. However, relatively few of the rules are included. For example, on Form A, of the 24 capitalization rules commonly taught (Brigance, 1999), 9 are tested on 12 WJ-III items. Of the 26 punctuation rules commonly taught (Brigance, 1999), only 13 are tested by the 23 WJ-III items. Most rules are assessed only once. Because both punctuation and capitalization are tested on one subtest, it is difficult to determine if a low score reflects a problem in one or both areas. Again, for comprehensive instruction purposes additional assessment will be necessary.

In the Examiner's Manual, seven Standard Battery subtests and five Extended Battery subtests are said to be appropriate for preschoolers. However, as noted later in this review, limited floors and item gradients suggest that there are too few items for comprehensive assessment of preschoolers on some of these subtests.

TECHNICAL ADEQUACY

The following information pertains to use of the WJ-III with preschoolers and school-aged students.

Standardization. The authors obtained a large norm sample of 1,143 participants for ages 2 through 5 years. The manual states that demographic characteristics of the sample "approximated the U.S. population distribution" but notes that weighting was required to bring the data more in line with the census data. Weighting is used in an attempt to correct when the sample does not sufficiently match the population, but it may introduce its own sources of error. Prior to weighting, demographic characteristics of the WJ-III sample were similar to the census data in terms of community size, sex, race, and ethnicity (defined as Hispanic or non-Hispanic). However, the data were somewhat discrepant for geographic distribution and parents' education. The South was overrepresented (46% compared with 35.5% in the U.S. population) and the Midwest and West somewhat underrepresented (15.7% rather than 23.1%, and 17% rather than 22.4%, respectively). Parents with a high school education were underrepresented (50% compared with 62.4% in the U.S. population), whereas those with more than a high school education were somewhat overrepresented (34% compared with 25.9% in the U.S. population).

The school-aged norms for kindergarten through grade 12 are based on 4,783 participants. Prior to weighting, the sample was very similar to the census data in terms of geographic distribution, sex, community size, race, ethnicity, and parents' education. Percentages for demographic characteristics are not presented in the manual in a straightforward manner and have to be calculated from other data in the manual. Having the percentages readily available would be helpful for examiners.

Students with disabilities were included if they were selected as part of the random sampling procedure used and if they were in regular classes at least part time. Unfortunately, the percentage of students with disabilities included

in the standardization is not given, nor are demographic characteristics for these students described.

Reliability. Correlations for internal consistency were determined using the Rasch method for timed subtests and the split-half method for untimed subtests. Bracken (1987) and Bracken and McCallum (1998) suggested .80 or higher as a reasonable criterion for acceptable subtest reliability for most purposes. Correlations for the Standard Battery subtests were .80 or higher for nearly all ages for 8 of the 12 subtests. However, the uniformity of test items is problematic for Story Recall for ages 2 through 5, Understanding Directions for ages 11 through 18, Passage Comprehension for ages 15 through 18, and Story Recall-Delayed for ages 3 through 9. For the Extended Battery, correlation coefficients were acceptable for nearly all ages for 6 of the 10 subtests. Those subtests with correlations less than .80 at a number of ages were Picture Vocabulary, Oral Comprehension, Sound Awareness, and Punctuation & Capitalization.

A test-retest study for the three speeded subtests (Reading, Math, and Writing Fluency) used only a 1-day retest interval. The correlation for Writing Fluency for ages 7 to 11 was .76, but other correlations ranged from .80 to .95. However, retest intervals of only 1 day do not provide the stability data needed to generalize to practice. Rarely are individual educational planning meetings held the same day that students' assessments are conducted; typically, planning meetings are held several weeks later. Consequently, whether test results remain relatively stable over a time period of about 2 to 4 weeks, rather than 1 day, is important.

A second retest study examined three Standard Battery subtests (Letter-Word Identification, Passage Comprehension, and Applied Problems) and one Extended Battery subtest (Academic Knowledge). Spelling subtest results were combined with Punctuation & Capitalization, making it impossible to evaluate the stability of each subtest individually. Further, combining the two subtests increases the number of items used to determine the correlation and results in a higher correlation than would be obtained for either subtest alone. The Synonyms and Antonyms sections of the Reading Vocabulary subtest were evaluated also, but the Analogies section of the subtest was not. We did not find a rationale to explain the absence of correlation coefficients for four nonspeeded Standard Battery subtests and seven Extended Battery subtests, nor is a rationale given for combining two subtests and examining only two of the three sections for Reading Vocabulary. Retest intervals varied from less than 1 year (exact interval not given), 1 to 2 years, and 3 to 10 years. Thus, it may be that prior editions of the test were used. Most subtest correlations were .80 or higher. Data in this study are presented for very broad age groups (2-7 years, 8-18 years), rather than by age level. Data by 1-year age levels would enable examiners to determine, for example, how reliable the test results are for a 10-year-old.

A third study used data from 457 participants who ranged in age from 4 to 17 with a retest interval of 1 year. Data are presented for four age groups (4-7, 8-10, 11-13, and 14-17). Unfortunately only 17 "selected" subtests were included. We could find no explanation as to why these 17 subtests were selected or why no data were presented for 5 subtests. Of the 68 correlations, more than

half (35) were lower than .80 (range .53 to .92). For 10 of the 17 selected subtests, correlations for at least three of the four age groups were lower than .80 (Reading Fluency, Word Attack, Reading Vocabulary, Calculation, Writing Fluency, Writing Samples, Editing, Handwriting, Story Recall, and Oral Comprehension).

When making eligibility decisions, cluster rather than subtest scores for reading, writing, math, and oral language should be used because they provide a better sample of students' performance than do subtest results alone. Results used for this purpose should meet a criterion of at least .90 (Salvia & Ysseldyke, 2004). Test-retest correlations are provided for the WJ-III achievement clusters over a 1-year period for the 457 students used to evaluate the subtests. For the three reading clusters (Broad Reading, Basic Reading Skills, and Reading Comprehension), 5 of 12 correlations are lower than .90. Reading Comprehension appears strongly stable only for ages 4 to 7. For the three math clusters, data are provided only for Broad Math and Math Calculation Skills. For the Broad Math composite, correlations exceed .90 for all four age groups. For Math Calculation, however, correlations are lower than .90 for all four age levels. The rationale for excluding Math Reasoning data is not mentioned. For the three written expression clusters (Broad Written Language, Basic Writing Skills, and Written Expression), 9 of 12 correlations are lower than .90. Three of the four correlations for Broad Written Language reached .90 or higher, but none of the correlations for Basic Writing or for the Written Expression cluster reach the .90 criterion for any age group. No data are provided for the four composites for Oral Language, nor is a rationale provided for excluding these composites in the evaluation of test-retest reliability. Hence, the stability of these clusters is unknown.

Although the test-retest reliability appears acceptable for the cluster scores for Broad Math and for most age levels for Broad Reading, Basic Reading, and Broad Written Language, concerns exist. Many clusters do not meet the .90 criterion and, therefore, they are not strongly stable over the 1-year test-retest interval. Further, no test-retest data are provided for the Math Reasoning cluster, all four clusters for oral language, and several subtests. Examiners cannot be confident that the following clusters meet the criterion of .90 recommended for making eligibility decisions: Math Reasoning, Math Calculation, Oral Language, reading clusters for half the age groups, Basic Writing, and Written Expression. Why the authors used a 1-year retest interval, which is longer than the interval used in most studies, is unclear. This lengthy interval may have resulted in lower correlations for some clusters.

Interscorer reliability was evaluated using the earlier version of the test, the Woodcock Johnson-Revised (Woodcock & Johnson, 1989). According to the manual, only three subtests require subjective judgments, and only these three were evaluated: Writing Samples for grades 2, 3, 9, and a few high school students; Writing Fluency for grades 3 and 7; and Handwriting for grades 3 and 7. Correlations were .81 or higher except for Handwriting at grade 3, which was .75. Interscorer reliability at other grade levels and for other subtests is unknown.

Regarding alternate form reliability, the authors state that a spiraling omnibus item-banking method was used to assign items to the two forms of the

test. According to the manual, because of the large number of tests, only two examples (Calculation and Passage Comprehension) were provided as additional evidence of alternate-form reliability. The rationale for selecting these two subtests is not mentioned. For Calculation, plotted item difficulties and the raw score/W-ability ogive for the two forms suggest their similarity. For Passage Comprehension, Rasch W scores were correlated for the two forms for selected age groups. Why the particular age groups were chosen is not indicated. For school-aged students, correlations range from .80 to .96. Again, we could find no rationale for why such correlations were not presented for all subtests. Further data on alternate form reliability would enable examiners to determine whether Forms A and B yield similar results and can be used interchangeably.

Validity. To establish content validity, items were developed with input from experts and the items were selected using the Rasch model to measure particular traits. In the manual, subtests and clusters are said to correspond to major curricular areas noted in federal law. Because content validity is extremely important for achievement tests, we believe that additional attention to this aspect of test development is warranted. For example, for Letter-Word Identification the basis on which words were selected is not mentioned. Are the words for the Spelling subtest words that are commonly taught and, therefore, important words for students to be able to spell? Are items from Academic Knowledge (Science, Social Studies, and Humanities) addressing material typically taught? A considerable amount of information is available to address such questions, such as scope and sequence charts, word frequency lists, and research articles, yet we could find no such references in the WJ-III manual. As noted in the Description of Subtests section of this review, a number of subtests appear to assess educationally important skills such as Math Fluency and Word Attack but others appear to be less effective measures of educationally relevant content. For example, the measurement of understanding directions as used on the Understanding Directions subtest is of questionable educational usefulness. Some of these subtests (e.g., Calculation) have too few items for the test's age range, resulting in limited information for instructional planning.

We found relatively little data that addressed possible item bias. Nine experts reviewed items for bias. How the experts were selected is not explained. Differential item functioning was used for the following groups: male/female, White/non-White, and Hispanic/non-Hispanic. For this analysis, the section "most likely to be biased because of language and achievement influences" (p. 97) was used. The authors selected the Academic Knowledge from the Achievement Scale and the Comprehension-Knowledge section from the Cognitive Scale. Results suggested five items with possible bias, but only one with practical and statistical significance. This item was eliminated. Results of a factor-structure invariance study for the same three subgroups suggested a lack of bias. To fully evaluate possible bias, examiners will need additional data (e.g., mean scores, reliability, and validity data for these subgroups). Item bias data for the remaining 21 subtests are unknown, although apparently from the authors' perspective, bias is unlikely.

No norm tables appear in the manual; the test must be scored using the Compuscore and Profiles Program. Without norm tables, the only way to eval-

uate possible floor effects and item gradients is to enter various raw scores into the Compuscore Program and examine the results. If a test has inadequate floors, steep item gradients, or both, results will not discriminate well between students with low-average or average performance and those with extremely low levels of performance. Tests with floor problems can yield spuriously high scores. Consequently, Bracken and Walker (1997) recommended that, to avoid obtaining misleading test results, floors for subtests and total scores should extend at least two standard deviations below the mean. We used Bracken and Walker's criterion and the Compuscore Program for the WJ-III and entered raw scores beginning with 1 for the subtests for the Standard Battery and continued until we found an adequate floor for each subtest. All subtests for ages 2 to 3-6 failed to meet Bracken and Walker's criterion. For 2-year-olds, the lowest standard score possible on any subtest is 80. Inadequate floors were also found for the Calculations subtest through age 6, Passage Comprehension through age 5, and for both Handwriting and three of the Writing Samples through age 6. Thus, these subtests at these ages may yield inflated scores because they lack a sufficient number of easy items.

Another factor that may contribute to inflated results for younger children is the limited content validity of some subtests. For example, because the first 14 items on Spelling assess copying marks such as a scribble or writing letters, a 7-year-old could receive a standard score of 89 and not be required to spell a single word. Six-year-olds could receive a standard score of 98 for Letter-Word Identification and a standard score of 84 on Passage Comprehension and not have to read one word; they could receive a standard score of 107 on Spelling and not have to spell a word. Thus, before interpreting results from such subtests for young examinees, examiners should review the items used with a student to examine the content covered.

Bracken (1987) suggested that, if a test has adequate item gradients, each raw score point will be equal to or less than one third of a standard deviation. We used Bracken's criterion and the Compuscore Program to sample the adequacy of item gradients for the Standard Battery by entering various raw scores for the youngest children. We found a number of item gradients that do not meet Bracken's criterion. For example, item gradient violations ranging from 6 to 28 standard score points were found with a 1-point change in the raw score through age 10 for Letter-Word Identification, Calculation, and Passage Comprehension. We were unable to enter all raw scores, for all ages, for all subtests, to obtain the tables needed to evaluate all item gradients for this test. Item gradient violations closest to the mean are the most troublesome. Examiners need norm tables in the manual so that item gradients can be determined easily. Because these tables are not available, we recommend that examiners check their results by entering raw scores 1 point above and 1 point below an obtained raw score for each subtest to determine whether standard scores would change by more than one third of a standard deviation for the particular student tested.

As evidence for construct validity, Achievement subtests were intercorrelated and results suggest that, except for three, the subtests are interrelated. The 63 correlations for the Handwriting Legibility Scale range from .01 to .38; for

Story Recall, 32 of 78 correlations range from .08 to .38; and for Story Recall-Delayed, 43 of 71 correlations range from .08 to .39. As expected, correlations between clusters that assess the same area of achievement (e.g., Basic Reading and Reading Comprehension) were higher than correlations between clusters less related (e.g., Basic Reading and Math Calculation). Confirmatory factor analysis indicates that many Achievement subtests load on a single factor for the broad Cattell-Horn-Carroll theory on which the test is based. However, several load on more than one factor, suggesting that they measure more than one characteristic.

Cluster growth curves based on median performance across age are also provided to address construct validity. As anticipated, the curves show that skills in math, reading, writing, and academic knowledge increase during the school years and curves for oral language peak at later ages. Growth curves for Reading Decoding, Reading Comprehension, Basic Writing, Writing Ability, Quantitative Reasoning, and Math Achievement describe the uniqueness of each of these factors.

The relationship of Achievement clusters to results for several intelligence tests was examined as well. Correlations with the Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991) are moderate for the Verbal and Full Scale, but the clusters do not appear to have a strong relationship to the Performance Scale; correlations range from .22 to .43. Correlations between Achievement clusters and WJ-III Cognitive clusters are generally moderate to high, except for Visual-Spatial Thinking, Auditory Processing, and Phonemic Awareness, which have correlations ranging from .08 to .27 across the Achievement clusters. Five subtests for preschoolers and the cluster results for Oral Expression, Listening Comprehension, and Oral Language (Standard and Extended sections) were compared with results from the WJ-III Cognitive Section, the Differential Ability Scale (Elliott, 1990), and the Wechsler Individual Preschool and Primary Scale of Intelligence-Revised (Wechsler, 1989). Correlations range from .41 to .86 for Oral Expression, Listening Comprehension, and Oral Language (Standard and Extended sections) clusters. For the Letter-Word Identification subtest, most correlations were also within this range. However, the majority of correlations for the subtests Passage Comprehension, Word Attack, Applied Problems, and Spelling range from .05 to .39, indicating that results from these four subtests are not strongly related to results from intelligence tests. No data for Academic Knowledge or Sound Awareness subtests are presented.

To address concurrent validity, the Achievement clusters were compared with subtests and composites from the Kaufman Test of Educational Achievement (KTEA; Kaufman & Kaufman, 1985) and the Wechsler Individual Achievement Test (WIAT; Wechsler, 1992). Correlations for WJ-III Achievement clusters and related KTEA and WIAT subtests and composites range from .44 to .82 for reading, .29 to .70 for math, .31 to .77 for written language, and .43 to .56 for oral language.

SUMMARY: WHAT EXAMINERS SHOULD KNOW

The norm sample for the WJ-III for school-aged students appears to consist of a large, nationally representative sample even before results were weighted. For children aged 2 through 5, the sample seems representative before weighting except for geographic distribution and parents' education. The South was somewhat overrepresented, as were parents with more than a high school education. Careful norming is preferable to weighting because weighting places too much emphasis on smaller numbers of participants. Including the number of participants with disabilities and a description of their demographic characteristics would have been helpful because this is an important variable.

Internal consistency data suggest homogeneity of the test items for 14 of the 22 subtests. However, the remaining 8 subtests have weak correlations at many age levels.

Unfortunately the test-retest reliability data for the WJ-III Achievement are problematic. No test-retest data are presented for the four oral language clusters or the Math Reasoning cluster, and no reason for this omission is given. Hence, whether results for these areas are stable estimates of performance is unknown. Particularly troublesome is that no rationale is provided as to why stability data are presented for only certain subtests selected by the authors, why data for some subtests are combined, and why only several sections of Reading Vocabulary were evaluated for stability. Also, many clusters do not meet the minimum criterion of .90 for acceptable reliability for results that are used in making important educational decisions. Rather than the 2- to 4-week retest interval typically used to evaluate stability, data are presented for either a 1-day retest interval or over periods of several years, possibly using prior editions of the test. Consequently, these data are less than ideal in terms of their application to situations that examiners face in their practice.

Similarly, interscorer and alternate-form reliability data are not complete. That is, interscorer reliability data are provided for only 3 of the 22 subtests and alternate form reliability data are provided for only the Calculation and Passage Comprehension subtests. Whether Forms A and B yield similar results for the remaining 20 subtests cannot be ascertained from the information in the Technical Manual.

Content validity should be the basis on which a test is developed; in our opinion, this construct deserves considerably more attention than it was given in the manual. A number of subtests on the WJ-III assess important skills using appropriate formats. However, some items on some subtests do not appear appropriate and may contribute to inflated scores for younger students. Justification for the educational relevance of some test items and formats is lacking. Additional efforts to evaluate item bias are needed (e.g., reliability and validity data for various subgroups of students).

Growth curves and confirmatory factor analysis generally support the test's construct validity. Except for the Handwriting Legibility scale, Story Recall, and Story Recall-Delayed subtests, the Achievement subtests are interrelated. The Achievement clusters seem to have an appreciable relationship with most results from intelligence tests. However, except for Letter-Word Identification, subtest results for preschoolers either are not strongly related to intelligence test performance or no data are provided.

For concurrent validity, Achievement clusters typically have moderate to high correlations with KTEA and WIAT subtests and composites. However, no data are presented to describe the relationship of WJ-III Achievement subtests with these measures.

By systematically entering data for various subtests at different ages into the Compuscore Program, we found a number of subtests with limited or inadequate floors and item gradients. When such problems occur results will not discriminate well between examinees with low average performance and those with very low performance levels. Because norm tables are not presented in the manual, considerable time and effort is required for examiners to obtain information on floors and item gradients.

According to the manual, the Extended Battery provides information for determining students' strengths and weaknesses for planning instruction. Yet the test is based primarily on the principle of providing a broad sampling of achievement rather than an in-depth assessment in a relatively narrow area (p. 52, Examiner's Manual). The sample of skills is too limited for comprehensive instructional planning, although information on general areas of strength and difficulty may be determined.

The Technical Manual for the WJ-III is written in such a way that many hours are required to extract the basic information needed to evaluate the test's technical adequacy. For example, to compare the norm sample with the percentages given for the census, one has to calculate the percentages for the norm sample using the reported number of participants and the total number in the sample. Though not a difficult task, examiners expect to have this information available. For test-retest and alternate-form and interscorer reliability, no data are presented for many subtests and clusters. To determine which subtests lack data, one has to compare the subtests for which correlations are given with a list of all of the subtests for the test. The difficulty required to extract information makes it difficult for examiners to obtain critical information from the manual.

If WJ-III results are used to aid in making eligibility decisions, caution should be exercised for various age levels because sections of the test do not meet the minimum criterion for reliability of .90 suggested by Salvia and Ysseldyke (2004) for tests used to make important educational decisions about students. "Reliability sets the limits for validity" (Bracken & Walker, 1997, p. 488); thus, when reliability coefficients are lower than experts recommend as a minimum, validity will suffer as well.

This well-known and widely used measure seems to have considerable potential for assessing achievement levels. The WJ-III addresses some important areas of achievement, and a number of subtests provide useful information. Unfortunately, we found the technical adequacy to be considerably more variable than expected for the third edition of the test. We hope future editions will address the specific concerns raised.

Sharon Bradley-Johnson
Sandra Kaouse Morgan
Christie Nutkins
Central Michigan University

REFERENCES

American Psychological Association (1999). *Standards for educational and psychological tests.* Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 4,* 313-326.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test examiners manual.* Itasca, IL: Riverside.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for 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: Guilford.

Brigance, A. (1999). *Comprehensive Inventory of Basic Skills-Revised.* North Billerica, MA: Curriculum Associates.

Elliott, C. D. (1990). *Differential Ability Scales.* San Antonio, TX: The Psychological Corporation.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement.* Circle Pines, MN: American Guidance Service.

Salvia, J., & Ysseldyke, J. (2004). *Assessment* (9th ed.). Boston: Houghton-Mifflin.

Wechsler, D. (1989). *Wechsler Individual Preschool and Primary Scale of Intelligence-Revised.* San Antonio, TX: The Psychological Corporation.

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

Wechsler, D. (1992). *Wechsler Individual Achievement Test.* San Antonio, TX: The Psychological Corporation.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery-Revised.* Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K., & Mather, N. (1989). *Woodcock-Johnson Tests of Achievement: Third Edition.* Itasca, IL: Riverside.