Journal of Psychoeducational Assessment
Volume 25 Number 1
March 2007 3-16

10.1177/0734282906291768
http://jpa.sagepub.com
hosted at
http://online.sagepub.com

Is the Woodcock-Johnson III a Test for All Seasons?

Ceiling and Item Gradient Considerations in Its Use With Older Students

Nancy Krasa
Columbus, Ohio

This study assesses the adequacy of item gradients and ceilings for the subtests of the Woodcock-Johnson III (WJ III) Cognitive and Achievement batteries, including the Diagnostic Supplement, in their use with participants ages 16 to 25 and Grades 10 to 18. Of the 52 subtests, 18 contain adequate item gradients and ceilings for the entire age and grade range. The remaining 34 subtests have inadequate ceilings and/or inadequate item gradients in at least the interval between the first and second standard deviations above the mean, as predicted by the growth curves of the factor clusters, which peak around age 25. The report discusses developmental and practical implications and suggests improvements for the next revision of the WJ.

Keywords: *Woodcock-Johnson; ceiling; item gradient; adolescent; college*

The Woodcock-Johnson III Cognitive (COG) and Achievement (ACH) batteries (WJ III; Woodcock, McGrew, & Mather, 2001), including the newly added Diagnostic Supplement (SUPP; Woodcock, McGrew, Mather, & Schrank, 2003), constitute a broad, comprehensive instrument for the evaluation of both cognitive processing and academic achievement. Normed on samples ranging from early childhood to old age, and from kindergarten through graduate school, it is equally expansive in its developmental scope. Because of the wide range of ages and educational levels it encompasses, it is serviceable for evaluating a broad range of competence levels. The ability to measure strengths as well as weaknesses is important not only for any comprehensive intellectual and academic evaluation but specifically for the assessment of deficits and disabilities, which are most usefully defined relative to an individual's strengths.

At the high school and college levels, the WJ III batteries are often used to document learning disabilities for the purpose of educational evaluation and planning. For example, the Educational Testing Service (1998, 1999) lists the WJ III among the tests it regards as acceptable for documentation of cognitive ability, information processing, and academic achievement in the diagnosis of both learning disabilities and attention deficit hyperactivity disorder. Results from WJ III assessments contribute to the selection of appropriate remediation and

Author's Note: Special thanks to Bruce A. Bracken, for his valuable comments, and to Terri Yoder, for her help with data management and manuscript preparation. Correspondence concerning this article should be sent to Nancy Krasa, 268 North Parkview Avenue, Columbus, OH 43209; e-mail: nkrasa@columbus.rr.com.

treatment for these disorders, as well as to decisions regarding reasonable accommodations in high school and college and on high-stakes college, graduate school, and professional school admissions tests. Among older students and adults, the WJ III can also provide information that is valuable for career decisions. For psychologists who are called upon to evaluate comprehensively a diverse population, the WJ III is an economical, all-purpose instrument. Thus, in many ways, it is a test for all seasons.

However, a closer look at the life span developmental profile of the WJ III suggests that caution may be in order for its use with older students. The growth curves of the cognitive and achievement factor clusters, illustrated in the WJ III *Technical Manual* (McGrew & Woodcock, 2001, pp. 55-56), demonstrate that performance rises rapidly through the early years and peaks around age 25. After age 25 or 30, performance in most areas goes into gradual decline. This means that for older students—those in mid-adolescence through their early twenties—it is important that test instruments be able to tap this peak performance even among the strongest individuals. This feature would be equally important for middle-aged individuals, prior to any significant age-related decline in function.

The goal of this study is to evaluate the ability of the WJ III to measure adequately the full range of performance levels in older students. Specifically, it investigates the item gradients and ceilings of all 52 subtests of the WJ III for ages 16-0 to 25-0 and for Grades 10.0 to 18.0 (the "target range").

Method

Data

Because the WJ III *Technical Manual* (McGrew & Woodcock, 2001) does not offer tables of raw scores and standard scores, manual scoring was not an option. Thus, the WJ III Compuscore and Profiles Program (Schrank & Woodcock, 2003) was the source for all the data.

Item gradients. An *item gradient* refers to "the steepness with which standard scores change from one raw score unit to another" (Rathvon, 2004, p. 49). When the item gradient of a test or subtest is too steep, it does not sufficiently absorb "noise"—that is, errors irrelevant to the construct being tested (such as carelessness or distractibility) can lead to an abrupt change in standard score that does not reflect a true difference in the ability being tested. The test or subtest, thus, would only very roughly discriminate between levels of competence, limiting the power of test interpretation (Bracken, 2000). Moreover, a too-steep item gradient in only part of the ability range may reflect a highly skewed score distribution, which would invalidate the use of percentile scores and other standard scores, and of intertest comparisons (Anastasi & Urbina, 1997).

To assess item gradients for each power subtest (i.e., the untimed measures of ability across difficulty levels, such as spelling or calculations), raw scores at the mean (M), at $M \pm 1$, and at one ($\pm$ 1 SD) and two ($\pm$ 2 SD) standard deviations were identified. Because no tables were available, making it difficult to assess item gradients at every age and grade in the target range, subtests and cluster scores were assessed at ages 16-0 and 25-0 and at

Table 1
Spatial Relations Scoring, Age 16-0

	−2 *SD*	−1 *SD*	*M* − 1	*M*	*M* + 1	1 *SD*	2 *SD*
Raw score	46	59	69	70	71	77	80
Z score	−2.01	−1.03	−0.09	0.02	0.12	1.00	2.06

Grades 10.0 and 18.0, the outer boundaries of the target range. Relevant raw scores were identified according to the following rules.

The mean (*M*) on a given subtest at a given age or grade was defined as the raw score associated with the *z* score closest to 0.00. If two raw scores had symmetric *z* scores (e.g., –0.05 and 0.05), the one that minimized the discrepancy between the *z* scores associated with *M* + 1 and *M* − 1 was defined as *M* under the assumption that standard scores tend to cluster most tightly around the mean, at least in normal distributions (Anastasi & Urbina, 1997).

The raw scores associated with −1 *SD* and −2 *SD* were defined as the highest raw score for which $z \leq -1.00$ and $z \leq -2.00$, respectively (on tests for which errors rather than correct responses serve as raw scores, the lowest raw score satisfying these criteria was used). Similarly, raw scores associated with 1 *SD* and 2 *SD* were defined as the lowest raw score for which $z \geq 1.00$ and $z \geq 2.00$, respectively (or the highest error raw scores satisfying these criteria). The *z* scores associated with the raw scores so defined were considered the target *z* scores. Finally, to evaluate the adequacy of the item gradient around the mean, the *z* scores associated with $M \pm 1$ were also noted. Table 1 illustrates the data for COG subtest Spatial Relations at age 16-0.

For some multisection subtests (Verbal Comprehension, General Information, Reading Vocabulary, Quantitative Concepts, and Academic Knowledge), for which the ratio of section scores affects the total raw score for each subtest, care was taken to input balanced section scores. For other multisection subtests (e.g., Sound Awareness), the *z* score depended only on the total raw score. For multi-level subtests based on developmental considerations (e.g., Understanding Directions), scoring was evaluated for the age- or grade-appropriate sections. For multilevel subtests based on performance cutoff criteria (e.g., Visual-Auditory Learning), calculations were based on perfect performance up to failure at the raw score.

The data for the speed subtests (i.e., timed tests consisting of low-difficulty items, measuring skill fluency rather than range of ability, such as Reading Fluency) were developed in the same manner. (On Reading Fluency, the error score was held at zero, and the number correct was used as the raw score.) However, for some of those subtests at some age and grade levels, the maximum allowed time limit did not fully test the upper limits of ability. When time bonuses were needed in order to reach the target *z* scores, the relevant times were converted to raw score equivalents (based on the maximum possible raw score) for data analysis. The conversion was accomplished by finding the *z* scores associated with each 1-second decrease in time from the allowed limit. On these subtests, changes of anywhere from 1 to 6 seconds were associated with a single *z*-score change. Therefore, each decrement in the *z* score was counted as one raw-score decrement, and raw-score equivalents could thus be found for each target *z* score.

The three delayed-recall subtests (Visual-Auditory Learning Delayed, Memory for Names Delayed, and Story Recall Delayed) are also time dependent. All the data from these subtests

were based on the minimum delay (30 minutes), as that was the least challenging condition. On only one subtest, Memory for Names Delayed, did that minimum time delay not always provide an adequate ceiling, precluding the assessment of some interval gradients.

Ceilings. *Z* scores for the maximum raw score (or for the minimum error score, or for the minimum response time at the maximum raw score) attainable were obtained for each subtest at ages 16-0 and 25-0 and at Grades 10.0 and 18.0.

Data Analysis

Item gradients. Item gradient data were analyzed according to the adequacy criteria proposed by Bracken (2000), who suggested that a change in one raw score point should not result in a change of more than one third of a standard deviation. Although this is the only published guideline available, it is generally accepted and has been used in several critiques of cognitive and academic tests (see Bradley-Johnson & Durmusoglu, 2005; Bradley-Johnson, Morgan, & Nutkins, 2004; Flanagan & Alfonso, 1995; Rathvon, 2004). To this end, the adequacy of the item gradient of each ability interval (−2 *SD* to −1 *SD*, −1 *SD* to *M*, *M* to 1 *SD*, and 1 *SD* to 2 *SD*) for each subtest at each of the target range boundaries was assessed by dividing the discrepancy between the target *z* scores by the discrepancy between the associated raw scores. This transformation resulted in an average *z* score change per 1-point raw score change for that ability interval on that subtest at that age or grade. An item gradient was deemed adequate for an ability interval if the average *z* score for that interval measured ≤ 0.33. The item gradient around the mean was calculated as the discrepancy between the *z* scores associated with $M + 1$ and $M - 1$, divided by 2, and likewise judged according to the Bracken (2000) guidelines.

Subtests with adequate item gradients for an ability interval at ages 16-0 and 25-0 and at grades 10.0 and 18.0 (the outer boundaries of the target range) were deemed adequate for that ability interval for the entire target range. For ability intervals in which a subtest did not meet the item gradient adequacy criterion for both boundary ages and both boundary grades, item gradients were judged to be inadequate. When the item gradient criterion was not met for an ability interval at either the upper or lower age or grade boundary, item gradients were calculated for each age or grade (at 1-year intervals) in the target range for that ability interval to determine at which age(s) and/or grade(s) in that ability interval the item gradient was inadequate.

Ceilings. Ceiling adequacy was judged according to Bracken's (2000) criterion that every test and subtest should provide accurate scores at least two standard deviations above the mean ($z \geq 2.00$). Again, this guideline was the only one available but has been used in other related work (see Bradley-Johnson & Durmusoglu, 2005; Flanagan & Alfonso, 1995; Rathvon, 2004). All subtests with adequate ceilings at age 25-0 and Grade 18.0 were deemed to have adequate ceilings, under the assumption that maximum performance could be expected at the upper bounds of the target range. For subtests on which the maximum raw score did not reach a *z* score ≥ 2.00 at age 25-0 or Grade 18.0, *z* scores were obtained for the maximum raw score at 1-year age or grade intervals for the entire target range, and

Table 2
WJ III Subtests With Adequate Item Gradients and Ceilings for Ages 16 to 25 and Grades 10 to 18

Cognitive	Supplement	Achievement
1 Verbal Comprehension	22 Visual Closure	2 Reading Fluency
6 Visual Matching	26 Cross Out	3 Story Recall
7 Numbers Reversed	27 Memory for Sentences	5 Calculations
11 General Information		8 Writing Fluency[a]
12 Retrieval Fluency[b]		10 Applied Problems
16 Decision Speed		12 Story Recall Delayed
20 Pair Cancellation		17 Reading Vocabulary
		18 Quantitative Concepts

Note: WJ III = Woodcock-Johnson III Test.
a. The 7-min time limit provided an adequate ceiling for the whole target range at all ability intervals.
b. The maximum scoreable number of responses provided an adequate ceiling for the whole target range at all ability intervals.

ceiling inadequacies were noted. Generally, the absence of an adequate ceiling (usually at 2 *SD*, occasionally at 1 *SD*) meant an inadequate item gradient within the associated ability interval. In one instance, on Visual-Auditory Learning, an examination of item gradients in the top ability level revealed unexpected ceiling inadequacies at Grades 15 to 17, despite an adequate ceiling at Grade 18.0.

Results

Of the 52 subtests in the WJ III, 18 have adequate ceilings throughout the target age and grade range, and have adequate item gradients within all ability levels at all ages and grades in the target range (see Table 2).

The remaining 34 subtests were found to have inadequate ceilings for all or part of the target range and/or inadequate item gradients for all or part of at least one ability level. Table 3 itemizes the various adequacies and inadequacies of those subtests.

Ten subtests, representing both the COG and ACH batteries, have inadequate ceilings for at least a portion of the target age or grade range. Notably, three of those subtests (Letter-Word Identification, Word Attack, and Sound Awareness) tap early building blocks of reading—phonological awareness and decoding. The only factor cluster score that does not have an adequate ceiling for the target range was Basic Reading, which topped out at age 17 and Grade 11.

The item gradient inadequacies of the 34 subtests were found to lie chiefly in the uppermost ability interval, between the first and second standard deviations above the mean. Among those with partial item gradient inadequacies, most fall short in the uppermost ages and grades of the target range. This is even the case with Writing Sample, despite its having developmentally determined item blocks. Twelve of the 34 subtests have inadequate item gradients above the mean. One subtest, Memory for Words, has inadequate item gradients in all four ability intervals, although it has an adequate ceiling, and another, Sound Patterns–Music, has an adequate item gradient for only the lowest ability interval.

(text continues on p. 13)

Table 3
Item Gradient and Ceiling Adequacy of WJ III Subtests for Ages 16 to 25 and Grades 10 to 18

		Item Gradient				
Battery	Subtest	–2 *SD* to –1 *SD*	–1 *SD* to *M*	*M* to 1 *SD*	1 *SD* to 2 *SD*	Ceiling
COG	2 Visual-Auditory Learning	Adequate	Adequate	Adequate	Inadequate	Adequate for age norms; inadequate for Grades 15-17
	3 Spatial Relations	Adequate	Adequate	Adequate	Inadequate	Adequate
	4 Sound Blending	Adequate	Adequate	Adequate	Inadequate for age norms; inadequate above Grade 11	Adequate
	5 Concept Formation	Adequate	Adequate	Adequate	Inadequate	Inadequate
	8 Incomplete Words	Adequate	Adequate	Adequate	Adequate for age norms; inadequate for Grades 11-15, 17-18	Adequate
	9 Auditory Working Memory	Adequate	Adequate	Adequate	Adequate for age norms; inadequate for Grade 18 only	Adequate
	10 Visual-Auditory Learning Delayed	Adequate	Adequate	Adequate	Inadequate	Adequate
	13 Picture Recognition	Adequate	Adequate	Adequate	Inadequate	Adequate
	14 Auditory Attention	Adequate	Adequate	Adequate for age norms; inadequate above Grade 12	Inadequate	Adequate
	15 Analysis/ Synthesis	Adequate	Adequate	Adequate for age norms; inadequate for Grades 12-14, 18	Inadequate	Adequate

	17 Memory for Words[a]	Inadequate	Inadequate	Inadequate	Inadequate	Adequate
	18 Rapid Picture Naming	Adequate	Adequate	Adequate	Inadequate above age 23; inadequate for Grades 13, 16-18	Inadequate for ages 24-25; inadequate for Grades 13, 16-18[b]
	19 Planning	Adequate	Adequate	Adequate	Adequate for age norms; inadequate above Grade 14	Adequate
SUPP	21 Memory for Names	Adequate	Adequate	Adequate	Inadequate above age 16; inadequate for Grades 11, 13, 15-18	Adequate
	23 Sound Patterns–Voice[c]	Adequate	Adequate	Inadequate	Inadequate	Adequate
	24 Number Series	Adequate	Adequate	Adequate	Adequate for age norms; inadequate for Grades 13, 15-18	Adequate
	25 Number Matrices	Inadequate for age 16 only; inadequate for Grade 10 only[d]	Adequate	Adequate	Adequate for age norms; inadequate for Grades 17-18	Adequate
	28 Block Rotation	Adequate	Adequate	Adequate	Inadequate above age 20; inadequate for Grades 14, 16-18	Adequate
	29 Sound Patterns–Music[a]	Adequate	Adequate for age norms; inadequate above Grade 10	Inadequate	Inadequate	Adequate

(continued)

Table 3 (continued)

Battery	Subtest	Item Gradient				Ceiling
		–2 *SD* to –1 *SD*	–1 *SD* to *M*	*M* to 1 *SD*	1 *SD* to 2 *SD*	
	30 Memory for Names Delayed	Adequate	Adequate	Adequate	Inadequate for age norms; inadequate for grade norms at 30-minute delay[e]	Adequate for age norms; inadequate above Grade 13 at 30-minute delay[e]
ACH	1 Letter-Word Identification[f]	Adequate	Adequate	Adequate	Inadequate above age 16; inadequate above Grade 10	Inadequate above age 17; inadequate above Grade 13
	4 Understanding Directions	Adequate	Adequate	Inadequate for age norms; inadequate above Grade 11	Inadequate	Inadequate above age 16; inadequate above Grade 13
	6 Math Fluency	Adequate	Adequate	Adequate[g]	Inadequate above age 20; inadequate above Grade 14[h]	Inadequate above age 20; inadequate above Grade 14[i]
	7 Spelling	Adequate	Adequate	Adequate	Inadequate above age 19; inadequate above Grade 12	Inadequate above age 21; adequate for grade norms
	9 Passage Comprehension	Adequate	Adequate	Adequate	Inadequate above age 17; inadequate above Grade 11	Adequate

11 Writing Sample	Adequate[j]	Adequate	Inadequate above age 17 on Block D[k] and for ages 22-23 on Block E; adequate for grade norms[l]	Inadequate for age norms on Block D[m] and on Block E; inadequate for grade norms on Block D[m] and above Grade 10 on Block E	Adequate
13 Word Attack[f]	Adequate	Adequate	Inadequate	Inadequate	Inadequate
14 Picture Vocabulary	Adequate	Adequate	Adequate	Inadequate above age 17; inadequate for Grades 13, 15-18	Adequate
15 Oral Comprehension	Adequate	Adequate	Adequate for age norms; inadequate for Grades 13, 14, 16-18	Inadequate	Adequate
16 Editing	Adequate	Adequate	Adequate	Inadequate above age 16; inadequate above Grade 13	Adequate
19 Academic Knowledge	Adequate	Adequate	Adequate	Adequate for age norms; inadequate for Grade 18 only	Adequate
20 Spelling of Sounds	Adequate	Adequate	Inadequate above age 17; inadequate above Grade 12	Inadequate	Adequate

(continued)

Table 3 (continued)

Battery	Subtest	Item Gradient				Ceiling
		–2 *SD* to –1 *SD*	–1 *SD* to *M*	*M* to 1 *SD*	1 *SD* to 2 *SD*	
	21 Sound Awareness	Adequate	Adequate	Inadequate	Inadequate	Inadequate
	22 Punctuation/ Capitalization	Adequate	Adequate	Inadequate above age 17; inadequate above Grade 12	Inadequate	Adequate

Note: *Adequate* and *inadequate*, unelaborated, signify that on that subtest, the item gradient within that ability level or the ceiling of the subtest was either adequate or inadequate for both ages (16-0 and 25-0) and both grades (10.0 and 18.0) examined, and thus, presumably, for the entire target age and grade range. *M* = raw score mean; *SD* = *standard deviation*; WJ III = Woodcock-Johnson III Test; COG = Cognitive; SUPP = Diagnostic Supplement; ACH = Achievement.

a. Inadequate gradient about the raw score mean *M* (*M* − 1 to *M* + 1) for age norms and grade norms.

b. Compuscore does not compute *z* scores for *t* < 1:09.

c. Inadequate gradient about the raw score mean *M* (*M* − 1 to *M* + 1) for Grade 18.

d. Inadequate floor for age 16 only; inadequate floor for Grade 10 only.

e. Longer delays are necessary to provide adequate ceilings above Grade 13; gradients are difficult to assess.

f. The Basic Reading Cluster has an inadequate ceiling above age 17 and Grade 11. It is the only cluster with an inadequate ceiling in the target age and grade range.

g. Time bonus required in order to reach 1 *SD* above age 20 and above Grade 14.

h. Time bonus required in order to reach 2 *SD* for all ages and grades.

i. For *t* < 2:28, Compuscore *z* scores remain flat and < 2.00.

j. Subjects with scores in the lowest portion of this interval on Block E are administered, and scored on, Block D.

k. Subjects with scores in the top portion of this interval on Block D are administered, and scored on, Block E.

l. Grades 10-12 on Block D; Grades 13-18 on Block E.

m. Subjects with any score in this interval on Block D are administered, and scored on, Block E.

Discussion

The results of this study on the use of the WJ III with older students complement those of previous studies on its use with the youngest students, which raised concerns about inadequate item gradients and subtest floors (Bradley-Johnson & Durmusoglu, 2005; Bradley-Johnson et al., 2004). Specifically, in those studies, the Letter-Word Identification, Passage Comprehension, Calculation, and Applied Problems subtests of the Achievement battery were found to have inadequate floors for the lowest ages in the WJ III norming sample and to have excessively steep item gradients in the age range of 4 years to 8 years, compromising test interpretation and usefulness.

As predicted, the majority of WJ III subtests (34 of 52), plus one cluster, were found to have inadequacies in either the ceiling or item gradient of the top ability interval for all or part of the age and/or grade range under consideration. Several of those subtests also have item gradient inadequacies in lower ability levels. Adequately measuring competence at the top (or bottom) of the ability scale, where the range of responses is restricted, is a common problem for test design. As the mean approaches the upper (or lower) limit of the test, the score distribution is bound to become skewed and/or truncated.

Developmental Implications

Of the 10 WJ III subtests that do not reach two standard deviations for all or part of the age and grade range tested, 7 appear to adequately test their respective domains, thus implying that those skills are essentially mastered prior to age 25 or Grade 18. Two of those subtests, Concept Formation and Understanding Directions, tap executive functions, such as attention and concentration. In addition, Concept Formation requires deductive reasoning, and the Understanding Directions subtest demands visual and verbal working memory. Further exploration of the data revealed that Concept Formation does not reach the second standard deviation above age 15 or above Grade 7; likewise, the subtest Understanding Directions tops out at age 16 and Grade 13. Therefore, it is reasonable to speculate that, among normally developing individuals, the executive and related cognitive functions necessary to perform these tasks are in place by midadolescence.

The five remaining subtests that do not reach two standard deviations for all or part of the target range are tests of academic skill. Spelling lacks an adequate ceiling for ages 22 to 25 and also has an inadequate item gradient above age 19 and above Grade 12 for the uppermost ability interval. Although it would certainly be possible to increase the number of difficult spelling words on the subtest, it is unlikely that common words that are more difficult to spell than the ones currently on the subtest can be found. It is thus reasonable to suggest that, on average, by college and young adulthood, the art of spelling is more or less mastered.

Similarly, Math Fluency lacks adequate item gradients or ceilings above age 20 or Grade 14. Although the software allows for the scoring of faster response times, the z scores flatten out short of 2.00. It may be that, like spelling, there is a natural point in early adulthood at which math fluency must be considered mastered and higher ability distinctions dropped.

The three remaining academic subtests, which all have substantial inadequacies in both item gradients and ceilings in the target range, are Letter-Word Identification, Word Attack,

and Sound Awareness. In fact, Sound Awareness and Word Attack have inadequate item gradients above the mean and on further exploration were found to have adequate ceilings only to age 14 (Grade 8) and age 12 (Grade 7), respectively. As noted, these subtests measure the basic building blocks of early reading, and the deficits of Letter-Word Identification and Word Attack are reflected in the inadequate ceiling of the Basic Reading Cluster, which tops out at age 17 and Grade 11. In the case of all three of these subtests, it is reasonable to conclude that they measure fundamental skills that are normally mastered by some point in adolescence.

These findings are inconsistent with the developmental curves in the WJ III *Technical Manual* (McGrew & Woodcock, 2001). Those curves, which reportedly are drawn from the *W* scores in the appendices, indicate (counterintuitively) that all skills measured by the COG and ACH batteries, including decoding, peak around age 25, as noted previously.

Implications for Practitioners

The findings of this study suggest that practitioners can feel secure in using and interpreting the 18 WJ III subtests listed in Table 2 with older students as directed in the test manual. Those subtests, representing several components of the Achievement, Cognitive, and Supplemental batteries, have adequate ceilings and item gradients for this age and grade range. Two additional subtests, Auditory Working Memory and Academic Knowledge, are adequate as well, except for use with the brightest Grade 18 graduate students, and thus can also generally be employed with confidence.

However, the findings of this study also strongly suggest that caution must be exercised in interpreting the results of the 32 remaining WJ III subtests and of the Basic Reading Cluster above the age and grade limits noted in Table 3. As noted, deviation from a normal score distribution, as reflected in steep item gradients and lack of an adequate ceiling, calls into question the use of such standard scores as percentiles at any point in the ability distribution.

Regarding specific subtests, Memory for Words should not be used at all in the target age and grade range because the items are too steeply graded throughout all ability levels. The subtest Sound Patterns–Music should be avoided as well, for similar reasons. The Number Matrices subtest should not be used at (or below) age 16 or Grade 10, where it lacks a floor. Rapid Picture Naming is useful and valid for most of the target range; however, for graduate students or those in their mid-20s who name all of the objects correctly in less than about 1½ minutes, scores should be interpreted only as "above average."

On the seven subtests with developmental limitations (Concept Formation, Understanding Directions, Spelling, Math Fluency, Letter-Word Identification, Word Attack, and Sound Awareness), discussed in the previous section, interpretation should be limited to "adequate development" (for at least average-range scores) or "inadequate development" (for below-average scores) for individuals whose ages or grade placements exceed the cutoffs noted for the specific subtests.

Of the remaining 21 subtests, 14 have excessively steep item gradients in the 1 *SD* to 2 *SD* range only, in all or part of the target age or grade range, although all but one of those subtests (Visual-Auditory Learning) has an adequate ceiling. When testing individuals who score above the mean in the indicated problematic age and/or grade range, it will be necessary to interpret their scores cautiously ("above average") and to avoid using percentiles.

Finally, seven of the subtests (Auditory Attention, Analysis/Synthesis, Sound Patterns–Voice, Writing Sample, Oral Comprehension, Spelling of Sounds, and Punctuation/Capitalization) have excessively steep item gradients above the mean for all or part of the target age or grade range, again despite adequate ceilings. In these cases, the integrity of the distribution is compromised, and caution must be applied in interpreting scores at all ability levels. Writing Sample presents additional complications because of its block administration.

Recommendations for Test Improvement

The only way to improve the item gradient of a sparsely tested ability interval is to add nonredundant items of the same difficulty as the other items in that interval. Where there is already an adequate ceiling, this should not be difficult to do. The risk would be in making the subtest too long. However, of the 24 subtests that have adequate ceilings but inadequate item gradients, 5 (Incomplete Words, Planning, Block Rotation, Sound Patterns–Music, and Writing Samples E) have maximum raw scores that produce z scores > 4.00 at age 25 and Grade 18, suggesting that items could be taken off the top to make room for more items in the inadequate ability intervals on these subtests. For the age and grade range examined in this study, the goal of increasing the item gradients in the ability intervals above the mean would be to ensure adequate discrimination of competent from very competent from truly outstanding individuals. For any given test subject, a more normally distributed test would also ensure that an individual's strengths could be distinguished with as much accuracy as his or her weaknesses. Specifically, it would guarantee that reported statistics, such as percentiles, would have meaning, which they do not have for nonnormal distributions.

To create a ceiling where one does not exist, items of greater difficulty than those already in the subtest need to be added. Whether or not a subtest with an inadequate ceiling can admit new items of greater difficulty, however, depends on the extent to which the subtest already tests the full upper range of its domain. As noted, seven subtests appear to assess cognitive and academic skills that are essentially mastered by normally developing individuals before or during the target age and grade range of this study. For Spelling, Letter-Word Identification, Word Attack, and Sound Awareness, scoring software and the examiner's manual should eliminate standard scores above the age and grade for which the subtest is no longer structurally sound and provide broader developmentally appropriate interpretive guidelines (e.g., "mastered," "not yet mastered"). Regarding Math Fluency, it is possible that the written format of the responses may impose a graphomotor limitation on response time and that another format (e.g., computer keyboard) may elicit faster response times and more refined distinctions between levels of mental mathematical fluency among older individuals. Finally, as to Concept Formation and Understanding Directions, different subtests may need to be developed that adequately test more advanced levels of executive functioning in older individuals.

Three remaining WJ III subtests do not reach two standard deviations for all or part of the age and grade range tested for reasons that do not seem to relate to development. Visual-Auditory Learning does not adequately tap the learning prowess of the most talented students in Grades 15 to 17. It is reasonable to speculate that a few more difficult items could be added to this subtest without making it unduly long, thus providing an adequate ceiling for the entire target range. The final two subtests are time dependent. Rapid Picture Naming

comes close to a ceiling for the uppermost ages and grades, but Compuscore does not accommodate response times under 1:09. An adjustment of the software should solve the ceiling problem for that subtest. The other time-dependent subtest that lacks a reliable ceiling is Memory for Names Delayed, which does not register z scores ≥ 2.00 above Grade 13 at the shortest (30-minute) delay. However, a ceiling could be reached with the delay extended anywhere from 1 hour to 3 days, depending on the grade level. In this case, an adjustment of administration rules to require longer delay intervals for college and graduate students would solve the ceiling problem for this subtest.

Finally, score tables would be a valuable addition to the next edition of the WJ, as would revised developmental curves. Researchers and practitioners alike would benefit.

It is hoped that the authors of the WJ III will consider the findings of this study and incorporate the recommendations that stem from them in their next revision of the batteries. If that happens, the WJ IV may indeed be a test for all seasons.

References

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

Bracken, B. A. (2000). Maximizing construct relevant assessment: The optimal preschool testing situation. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 33-44). Needham Heights, MA: Allyn & Bacon.

Bradley-Johnson, S., & Durmusoglu, G. (2005). Evaluation of floors and item gradients for reading and math tests for young children. *Journal of Psychoeducational Assessment, 23,* 262-278.

Bradley-Johnson, S., Morgan, S. K., & Nutkins, C. (2004). A critical review of the Woodcock-Johnson III. *Journal of Psychoeducational Assessment, 22,* 261-274.

Educational Testing Service. (1998, January). *Policy statement for documentation of a learning disability in adolescents and adults.* Retrieved February 7, 2006, from http://www.ets.org/distest/ldpolicy.html

Educational Testing Service. (1999, June). *Policy statement for documentation of attention deficit/hyperactivity disorder in adolescents and adults.* Retrieved February 7, 2006, from http://www.ets.org/distest/adhd plcy.html

Flanagan, D. P., & Alfonso, A. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13,* 66-90.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock-Johnson III*. Itasca, IL: Riverside.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford.

Schrank, F. A., & Woodcock, R. W. (2003). WJ III Compuscore and Profiles Program (Version 2.0) [Computer software]. Woodcock-Johnson III. Itasca, IL: Riverside.

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

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. A. (2003). *Diagnostic Supplement to the WJ III Tests of Cognitive Abilities.* Itasca, IL: Riverside.