Reading & Writing Quarterly, 17: 279–289, 2001
Copyright © 2001 Taylor & Francis
1057-3569/01 $12.00 + .00

A COMPARISON OF WOODCOCK–JOHNSON PSYCHO-EDUCATIONAL BATTERY-REVISED AND QUALITATIVE READING INVENTORY-II INSTRUCTIONAL READING LEVELS

Patrick P. McCabe, Ph.D.
St. John's University, Jamaica, NY, USA

Howard Margolis, Ed.D.
Queens College, Flushing, NY, USA

Edna Barenbaum, Ph.D.
Cabrini College—Education Division, Radnor, PA, USA

The Qualitative Reading Inventory-II (QRI-II) and the reading subtests of the Woodcock–Johnson Psycho-Educational Battery-Revised (WJ-R) were administered to 34 fourth-grade males reading at or below the 25th percentile on the Iowa Test of Basic Skills. Spearman rank-order correlation coefficients indicated that WJ-R reading scores and QRI-II oral instructional reading levels were moderately and significantly related. Fifty percent of children obtained identical instructional levels on the WJ-R and QRI-II, while 50% demonstrated differences of half a year or more. For the portion of students who obtained different WJ-R and QRI-II instructional levels, WJ-R levels exceeded QRI-II levels 92.1% of the time. Implications for placing poor readers in instructional level reading materials are discussed.

One of the more daunting and important challenges that teachers, reading specialists, and school psychologists face is correctly supplying below average readers with curriculum materials at their proper reading instructional level. To do this, many rely on scores from commercially published norm-referenced reading tests (NRT) or criterion-referenced reading tests (CRT).

Address correspondence to Patrick P. McCabe, St. John's University, 8000 Utopia Parkway, Jamaica, NY 11439. E-mail: mccabe@stjohns.edu

We wish to thank Dr. Donna Wilson, formerly of Valdosta State University, for helping to administer the tests in this study.

As with all articles published in *Reading & Writing Quarterly,* this article went through the anonymous review process.

Although they are often used to make instructional placement decisions, NRTs and CRTs differ in purpose and usually differ in design. NRTs are designed to "compare an individual's performance to the performance of his or her peers" (Salvia & Ysseldyke, 1995, p. 745), whereas CRTs are designed to "measure a person's skills in terms of absolute levels of mastery" (Salvia & Ysseldyke, 1995, p. 742). NRTs have students engage in many different tasks, whereas CRTs emphasize very few tasks.

Using NRT scores to place students in reading and other curriculum materials has raised many concerns. Three have particular relevance here. The first concern is that NRTs emphasize the measurement of lower-order skills using artificial formats (Falk, 1998). The second is that NRT scores often reflect the specific curriculum emphasized in school. Selecting an NRT that does not directly measure what instruction has emphasized can result in low test scores and invalid interpretations of scores (Salvia & Ysseldyke, 1995; Webster & Braswell, 1991). Third, scores on different NRTs that claim to measure the same constructs can differ dramatically, even when the tests themselves are significantly related (Slate, 1996).

To avoid the problems and limitations of NRTs, many educators use informal reading inventories (IRIs) to help place below average readers in reading materials. Because IRIs seek to directly measure a student's absolute mastery level in a specific area (e.g., Can Carol comprehend a 3rd grade reading passage with 100% accuracy?), they are considered CRTs.

Commercial IRIs typically consist of a series of graded word lists and reading passages that yield information about the student's independent, instructional, and frustration reading levels. At their independent level, students can read material without difficulty; materials at this level are appropriate for recreational reading and homework. At their instructional level, students can successfully read material without frustrating or losing motivation, but with teacher assistance; at the instructional level, teachers often preview vocabulary with students, help them activate prior knowledge and set purposes for reading, break selections into manageable units, ask questions, and provide assistance whenever help is needed. In contrast, students cannot successfully handle material at the frustration level, even with teacher assistance. At this level, they make too many mistakes and often get quite anxious. As Newcomer (1986) has noted, "Continuing to expect a child to read material at his or her frustration level can create serious achievement and emotional problems" (p. 26).

DO NRTS AND IRIS PRODUCE COMPARABLE RESULTS?

Evaluators generally choose which test(s) to administer to place students in reading materials or to determine a student's degree of reading difficulty.

In choosing one test over another, practitioners appear to rely on the verity of at least one assumption—that they would obtain essentially the same reading level whatever reading test they administered (Slate, 1996). That is, an NRT would yield essentially the same instructional reading level as an IRI, or, the very least, the test scores would not differ significantly.

However, this assumption may not be valid. Kress (1988) noted, "The results of an IRI will lead to placement in reading materials that are significantly less difficult than those particular standardized tests would recommend" (p. 2). In other words, IRIs may routinely produce reading level scores that are lower than NRT achievement scores. Analogous to Kress, McCormick (1999) asserted that IRI scores are often closer to students' actual classroom performance than standardized test (i.e., NRT) scores.

To examine whether or not NRTs and IRIs yielded comparable, interchangeable results, this study hypothesized that fourth grade boys of below average reading ability would obtain functionally discrepant reading scores on the reading portion of the Woodcock–Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock & Johnson, 1990), an NRT, and the Qualitative Reading Inventory-II (QRI-II; Leslie & Caldwell, 1995), an IRI.

METHOD

The Tests

The WJ-R is a widely used, individually administered, multi-skilled battery that assesses educational achievement (Witt, Elliott, Daly, Gresham, & Kramer, 1998). It produces a grade level instructional band that specifies a range of tasks the student "would perceive as quite easy... to a level that would be perceived as quite difficult" (Woodcock & Mather, 1990, p. 57). The easy instructional level is one standard error of measurement below the student's obtained score; the difficult instructional level is one standard error of measurement above the student's obtained score. Many practitioners use the grade equivalents that students obtain on the WJR's reading achievement subtests to identify the students' instructional reading level.

The WJ-R has been well reviewed. Salvia and Ysseldyke (1995) concluded that "the WJ-R provides a comprehensive assessment of cognitive and academic ability" (p. 651). Witt, Elliott, Kramer, and Gresham (1994) concluded that it serves as an excellent model for the reporting of test validity data (p. 176).

The QRI-II is a widely used IRI designed to "help find the level at which the student can read independently, with instructional guidance, and with frustration" (Leslie & Caldwell, 1995, p. 16). Depending on the student's anticipated level of achievement and the purposes of assessment, students are asked to read orally and/or silently, narrative and/or expository, familiar and/or unfamiliar text. Lower level QRI-II passages were written to resemble those found in basal readers; higher level passages were taken directly from social studies and science textbooks and slightly modified. Readability formulae and empirical analysis of student reading performances confirmed that increases in text difficulty were commensurate with grade level designations (Leslie & Caldwell, 1995). Related studies supported the QRI-II's oral reading acceptability criteria for determining instructional level. These studies found that "as the total number of miscues increases, the number of miscues that change the authors' meaning and are not corrected also increases" (Leslie & Caldwell, 1995, p. 309).

Test reviews for the QRI-II's predecessor, the QRI, were exceptionally positive. Lipson and Wixson (1997) concluded that "the QRI provides a set of procedures that is both conceptually and psychometrically sound and that represents a major step forward in the development of interactive assessment devices" (p. 341). Taylor, Harris, Peason, and Garcia (1995) concluded that the QRI "represents a standard to which other reading tests formal and informal might aspire" (p. 374).

Participants

Thirty-four fourth grade boys who attended a public school in rural, south central Georgia were randomly selected from a pool of 99 fourth grade boys who scored at or below the 25th percentile in reading on the Iowa Test of Basic Skills (ITBS) and obtained less than a third grade reading level on the ITBS. In this school, students scoring at or below the 25th percentile on the ITBS national norms were deemed remedial. On the basis of classwork, teachers had also identified each student as having reading problems.

These students remained in their regular classes and did not receive additional reading support. The sample had a mean age of 10.2 years with a standard deviation of seven months. Twenty-five participants were African American and nine were Caucasian. Families in this area had a mean annual income of approximately $35,000, which is considered lower middle class.

Procedures

All students were administered the QRI-II and the WJ-R Letter-Word Identification, Word Attack, Passage Comprehension, and Reading Vocabulary subtests. The senior author and a doctoral level psychologist from Valdosta State University administered the tests. Both examiners were highly proficient in the administration of these tests.

The tests were administered in counterbalanced order over two sessions of approximately 30 minutes each. Half the students first took the WJ-R and half first took the QRI-II. The tests were administered approximately one week apart, at the same time of day, in a quiet, private room in the students' school. The same room was used for both testing sessions.

One examiner administered all QRI-II tests and the other administered all WJ-R tests. After initial introductions, including questions and comments designed to build rapport and put the students at ease, the WJ-R reading achievement subtests and the QRI-II were administered in accord with each manual's directions.

The QRI-II manual suggests that "examiners use an oral reading format with... students suspected of reading below third grade level" (Leslie & Caldwell, 1995, p. 47). As each student had an ITBS reading score below third grade, they were asked to orally read the QRI-II's graded paragraphs.

Like most IRIs, the QRI-II assesses word recognition and comprehension accuracy to determine independent, instructional, and frustration oral reading levels. When analyzing word recognition in context, the QRI-II requires that examiners use either "total acceptability" (counting semantically acceptable miscues as correct) or "total accuracy" (counting deviations from the printed text as miscues). Total acceptability was used to assess word recognition miscues because the test authors found it to be the "best predictor of instructional level comprehension" (Leslie & Caldwell, 1995, p. 52).

Data Analysis

In accord with WJ-R procedures, test statistics were calculated for the battery's three reading clusters: Reading Comprehension, Broad Reading, and Basic Reading Skills. Each cluster was composed of two subtests. Reading Comprehension combined Passage Comprehension and Reading Vocabulary; Broad Reading combined Letter-Word Identification and Passage Comprehension; and Basic Reading Skills combined Letter-Word Identification and Word Attack. Spearman rank-order correlation coefficients were calculated to measure the relationship between the obtained instructional levels on the WJ-R clusters and the QRI-II instructional reading level.

WJ-R clusters were used instead of individual subtests as clusters tend to be more reliable than individual subtests and better reflect the integration of reading skills. Each cluster incorporates the information found in the constituent subtests and represents the arithmetic mean of the included subtests (Woodcock & Mather, 1990).

The WJ-R uses numerical year and month grade equivalent (GE) designations throughout its scale to designate different reading levels (e.g., 2.3, 2.5, 2.6, 2.9). In contrast, the QRI-II uses only one ordinal grade level designation for each grade (e.g., third grade), except for first grade. First grade is divided into three ordinal levels to represent historical and current reading practice: pre-primer, primer, and first grade. To compare WJ-R instructional reading levels to QRI-II pre-primer, primer, and first grade instructional reading levels, the following numerical designations were assigned:

- QRI-II pre-primer level: WJ-R GE of 1.0 to 1.2
- QRI-II primer level: WJ-R GE of 1.3 to 1.4
- QRI-II first grade level: WJ-R GE of 1.5 to 1.9

These correspondences retain the order in which pre-primer, primer, and first grade basal reading materials are introduced in first grade and approximate the academic month in which average achieving first graders read these or similar materials.

RESULTS

As the correlation coefficients in Table 1 indicate, moderate, significant relationships ($r = .68$ to $.73$) were found between WJ-R and QRI-II instructional reading levels.

TABLE 1 Relationship Among WJ-R and QRI-II Instructional Levels

Reading Tests ($N = 34$)	Spearman Rank-Order Correlation Coefficient Level of Significance	Coefficient of Determination/Shared Variance
QRI-II/WJ-R Reading Comprehension (Passage Comprehension and Reading Vocabulary subtests)	$r = .7377$ $p < .0001$ (2-tail)	$R^2 = 54.4\%$
QRI-II/WJ-R Broad Reading (Letter-Word Identification and Passage Comprehension subtests)	$r = .6826$ $p < .0001$ (2-tail)	$R^2 = 46.6\%$
QRI-II/WJ-R Basic Reading Skills (Letter-Word Identification and Word Attack subtests)	$r = .7043$ $p < .0001$ (2-tail)	$R^2 = 49.6\%$

TABLE 2 Comparison of WJ-R and QRI-II Instructional Levels

WJ-R Clusters and Totals	Difference Between WJ-R and QRI-II Instructional Levels*	Percent of Differences in Which WJ-R Instructional Levels Exceeded QRI-II Levels
Reading Comprehension (Passage Comprehension and Reading Vocabulary)	16/34 = 47%	15/16 = 93.7%
Broad Reading (Letter-Word Identification and Passage Comprehension)	21/34 = 61.7%	21/21 = 100%
Basic Reading Skills (Letter-Word Identification and Word Attack)	14/34 = 41.1%	11/14 = 78.5%
Totals (Reading Comprehension, Broad Reading, Basic Reading Skills)	51/102 = 50%	47/51 = 92.1%

*Difference is greater than or equal to (a) 1 level at first grade, or (b) .5 year on the WJ-R for QRI levels at grade 2 and above.

Because correlations can mask meaningful instructional level differences, each student's WJ-R and QRI-II scores were directly compared. As seen in Table 2, scores often differed. In such instances, WJ-R scores were almost always higher than QRI-II scores. Fifty percent of the time, student WJ-R GEs and QRI-II scores differed by one more QRI-II levels, or .5 of a year. In 92.1% of such instances, WJ-R instructional reading levels were higher than QRI-II levels.

DISCUSSION

Instructional Decision Making

At first glance, the moderate correlations between WJ-R and QRI-II instructional reading levels (ranging from $r = .68$ to .73) suggest that examiners might use these tests almost interchangeably, thus rejecting the hypothesis that fourth grade boys reading below average would obtain functionally discrepant reading scores on the WJ-R reading clusters and the QRI-II. However, direct comparisons of actual student scores indicate that students often obtained functionally discrepant reading scores, supporting the study's hypothesis. WJ-R scores differed from QRI-II scores 50% of the time. In these instances, WJ-R scores exceeded QRI-II levels by .5 of a year or more 92.1% of the time. This supports Kress's

(1988) contention that IRIs yield lower scores than standardized tests. Given that both tests are widely used to help identify the instructional reading level of students who function below average in reading and that there is currently no fully validated way to determine instructional reading levels, it is critical to ascertain which measure provides the most practical information for instructional placement in reading and which measure best avoids placing students in reading materials they are likely to find frustrating.

Daub and Colarusso (1996) offer a perspective that can help solve the dilemma of test selection for helping to determine a student's instructional reading level. They ask which measure most directly represents the reading tasks in which the student engages. This requires examiners to compare the typical reading demands made of the student with the tasks on the test. For example, to determine a student's proficiency to read and answer questions about relatively long selections of connected text, the QRI-II appears more valuable than the WJ-R. This is because the QRI-II has the student read relatively long selections of connected text whereas the WJ-R does not. Three of the four WJ-R reading subtests have the student read isolated words; only one subtest, Passage Comprehension, has the student read connected text. On this subtest, the passages are quite short, with the last eight items (presumably the subtest's most difficult items) averaging only 33.75 words ($SD = 5.70$ words). The WJ-R manual itself suggests that generalization to tasks unlike those measured by WJ-R is likely to be limited (Woodcock & Mather, 1990), making suspect the validity of any generalization of WJ-R reading levels to relatively longer selections of connected text.

Similarly, if instructional level for automaticity is needed, the QRI-II will likely provide a more appropriate instructional level because it directly measures words-per-minute on relatively long selections of connected text, whereas the WJ-R does not. In both instances, the QRI-II dramatically reduces the inference involved in decision making, which reduces the probability of error. In contrast, the WJ-R might prove more helpful than the QRI-II for selecting initial materials for short cloze activities. This is because the WJ-R uses a cloze procedure with short, disconnected paragraphs, whereas the QRI-II does not employ a cloze procedure.

When instruction aims to develop poor readers' word analysis abilities, neither the WJ-R nor the QRI-II offers adequately reliable or adequately precise information for determining which sound–symbol correspondences and decoding principles to teach. The QRI-II depends on the analysis of oral reading miscues to determine a student's word recognition and word analysis needs. For poor readers, the QRI-II often generates small samples of word recognition errors on specific type materials, limiting reliability and generalizability. Moreover, the mental processes involved in word recognition are invisible (Manzo & Manzo, 1995), limiting the understanding of error causation to hypothesizing. In essence, the limited oral reading task and small error sample make it impossible to draw more than tentative conclusions. Similarly, the WJ-R offers only a small number of items that assess sound–symbol relationships and decoding principles. Students reading at or below a third grade level are unlikely to have an opportunity to attempt all the WJ-R Letter-Word Identification and Word Attack subtest items; this results in a small error sample, which precludes a highly reliable analysis of errors. Both tests, however, can provide a starting point and guidance for an additional, more precise functional assessment of decoding skills, which might include running records of oral reading (see Tierney, Readence, & Dishner, 1995), curriculum-based evaluation (see Howell, Fox, & Morehead, 1993), informal word analysis tests (see Shanker & Ekwall, 2000), and planned diagnostic lessons aimed at identifying how the student responds to specific instructional procedures and the degree of teacher-support needed (see Lipson & Wixson, 1997).

When faced with conflicting information about instructional reading levels, evaluators should consider the Criterion of the Least Dangerous Assumption (CLDA). The CLDA states that "[i]n the absence of conclusive data, education decisions should be based on assumptions which if incorrect, will have the least dangerous effect on the student" (Donnellan, 1984, p. 142). For example, if WJ-R and QRI-II instructional reading levels differ, the lower level should be used as it is less likely to frustrate the student and thus potentially be less harmful. Starting a student at a lower level does not mean that the student should be kept at this level, as all any test can do is suggest a starting point for instruction. By carefully monitoring and quickly responding to a student's functioning, teachers can move the student up to higher level materials if the initial test results appear to underestimate the student's abilities.

Because students cannot wait for research to provide adequate validation of specific methods for determining instructional reading levels, educators are urged to carefully consider the recommendations in this discussion. In essence, the recommendations ask that educators analyze the more important and frequent classroom reading tasks that students engage in, select the test(s) that most closely matches the tasks, use test results and other relevant data to tentatively determine the student's instructional reading level, frequently monitor the student's progress, and quickly respond to observed needs. Although not a fool-proof formula, following such recommendations can compensate for the weaknesses inherent in all tests and help many below average readers achieve high levels of success.

Limitations of the Study

The relatively small sample size and the students' rural background limit generalization of the study's findings. Replication with similar (e.g., rural) and dissimilar (e.g., urban) below average readers is needed to validate the results. Thus, the article's conclusions about the specific tests used must be viewed as tentative. More important, however, than the specific tests used in this study is its validation of the often observed clinical phenomenon that students' reading levels differ markedly on different reading tests. If verified by future research, this has considerable implications for selecting tests to help ascertain a student's reading instructional level and legal eligibility as learning disabled.

REFERENCES

Daub, D., & Colarusso, R. (1996). The validity of the WJ-R and DAB-2 reading subtests with students with learning disabilities. *Learning Disabilities Research and Practice*, *11* (2), 90–95.

Donnellan, A. M. (1984). The criterion of least dangerous assumption. *Behavior Disorders*, *9*, 141–150.

Falk, B. (1998). Testing the way children learn: Principles for valid literacy assessments. *Language Arts*, *76* (1), 57–66.

Howell, K. W., Fox, S. L., & Morehead, M. K. (1993). *Curriculum-based evaluation* (Second ed.). Pacific Grove, CA: Brooks/Cole.

Kress, R. (1988). Some caveats when applying two trends in diagnosis: Remedial reading. ERIC Digest Number 6 (ERIC Document Reproduction Service No. ED 297 303; http://ericae.net/edo/ED297303.htm).

Leslie, L., & Caldwell, J. (1995). *Qualitative Reading Inventory-II*. New York: Harper Collins.

Lipson, M. Y., & Wixson, K. K. (1997). *Assessment and instruction of reading disability: An interactive approach* (Second ed.). NY: Longman.

Manzo, A. V., & Manzo, U. C. (1995). *Teaching children to be literate*. Fort Worth, TX: Harcourt Brace Jovanovich.

McCormick, S. (1999). *Instructing students who have literacy problems* (Third ed.). Englewood Cliffs, NJ: Merrill.

Newcomer, P. L. (1986). *Standardized reading inventory* (manual). Austin, TX: Pro-Ed.

Salvia, J., & Ysseldyke, J. E. (1995). *Assessment* (Sixth ed.). Boston: Houghton Mifflin.

Shanker, J. L., & Ekwall, E. E. (2000). *Ekwall/Shanker Reading Inventory* (Fourth ed.). Boston: Allyn & Bacon.

Slate, J. R. (1996). Interrelations of frequently administered achievement measures in the determination of specific learning disabilities. *Learning Disabilities Research & Practice*, *11* (2), 86–89.

Taylor, B., Harris, L. A., Pearson, P. D., & Garcia, G. (1995). *Reading difficulties: Instruction and assessment* (Second ed.). NY: McGraw-Hill.

Tierney, R. J., Readence, J. E., & Dishner, E. K. (1995). *Reading strategies and practices: A compendium*. Boston: Allyn & Bacon.

Webster, R., & Braswell, L. (1991). Curriculum bias and reading achievement test performance. *Psychology in the Schools*, *28* (3), 193–198.

Witt, J. C., Elliott, S. N., Daly, E. D., III, Gresham, F. M., & Kramer, J. J. (1998). *Assessment of at-risk and special needs children* (Second ed.). Boston, MA: McGraw-Hill.

Witt, J. C., Elliott, S. N., Kramer, J. J., & Gresham, F. M. (1994). *Assessment of children: Fundamental methods and practices*. Madison, WI: WCB Brown & Benchmark.

Woodcock, R. W., & Johnson, M. B. (1990). *Woodcock–Johnson Psycho-Educational Battery-Revised, Examiner's manual*. Chicago: Riverside.

Woodcock, R. W., & Mather, N. (1990). *Woodcock–Johnson Psycho-Educational Battery-Revised, Tests of Achievement, examiner's manual*. Chicago: Riverside.