A Review of the *Woodcock Reading Mastery Test—Revised* (WRMT-R)

Marybeth De Rose

The *Woodcock Reading Mastery Test—Revised* (1997) is the latest edition of the *Woodcock Reading Mastery Test*, which was originally published in 1973. The revised version, first published in 1987, is now in its 12th year of service and remains a popular choice among available reading tests. The objective of this review is to take a renewed critical look at the test in general, and in particular to inquire whether test results might be useful in making inferences about ESL learner reading ability.

The WRMT-R was designed to assess the reading levels of test-takers within an age range of 4 to 75 years of age and over. The manual suggests that the WRMT-R may be used with ESL (English as a Second Language) students, although provisions must be made for "ensuring that the subject understands the instructions, and providing additional practice opportunities, or reviewing the instructions" in order to "provide a valid measure of the English reading skills" (p. 18).

The WRMT-R, like its predecessor the WRMT, consists of two forms that facilitate retesting at short intervals. The original test consisted of five subtests: Letter Identification, Word Identification, Word Attack, Word Comprehension, and Passage Comprehension across two parallel forms (A and B). The revised WRMT has retained the two forms, but they are no longer parallel. Form H has been reduced to four subtests by eliminating Letter Identification from the original, and Form G has been expanded to six subtests to improve readiness assessment. Form G consists of the five subtests from the WRMT, plus an additional one taken from the Woodcock-Johnson Psycho-Educational Battery (Woodcock & Johnson, 1977). The six subtests of Form G, four of which comprise Form H, are as follows.

First, the new subtest Visual-Auditory Learning assesses the ability to learn 26 symbols and their names, as well as two symbols for word endings (-ing, -s), then read them in sentence form across seven brief test stories.

The Letter Identification subtest involves reading alphabet characters in several fonts, including cursive writing. Supplementary lists have been added to the revised WRMT, in upper and lower case, which also contain several diphthongs and digraphs for the purpose of assessing sound-symbol association. These tests are not normed and are therefore of diagnostic value only.

The Word Identification subtest entails reading and pronouncing words in isolation from lists of increasing difficulty, "even though [the test taker may have] had no previous personal experience with the word." Essentially the listed words must be decoded and pronounced in a manner consistent with the articulation guidelines in order to be counted as correct.

The Word Attack subtest assesses phonetic decoding skills through the reading aloud of increasingly complex nonsense words. Nonsense words have been used in order to eliminate the possibility of test-takers knowing how to pronounce them in advance, as might be the case with the occasional real stimulus word, no matter how obscure.

The Word Comprehension subtest consists of three distinct sections. The first two sections entail reading stimulus words and responding with either a synonym or an antonym, and the third section requires the test-taker to read the stimulus words, ascertain the relationship between them, and then continue the analogy to complete another stimulus pair of words. The vocabulary items have been categorized as either general reading, science-mathematics, social studies, or humanities. Separate category scores can be calculated for diagnostic purposes, but are not normed.

The last subtest, Passage Comprehension, resembles a cloze exercise, which requires test-takers to fill in a series of blanks according to the meaning of the surrounding sentences or phrases. Picture clues have been included for approximately one third of the easiest items.

The WRMT-R provides norms for subjects from kindergarten to college seniors in terms of grade, and for ages up to and beyond 75 years.

The original normative data were gathered from 6,089 subjects in 60 geographically diverse United States communities. The kindergarten to grade 12 sample comprised 4,201 subjects, and the college/university sample contained 1,023 subjects. The adult sample of subjects aged 20-80 but not enrolled in college comprised 865 subjects. The norming sample was structured to resemble the distribution of variables included in the 1980 US Census. Individual subject weighting was applied during data analysis to obtain "exact" proportion matching with the census data.

The WRMT-R was renormed in 1995-1996. In Canada these norms have just been made available commercially in recent months.

The current norms (WRMT-R NU) are based on a somewhat smaller sample population of approximately 3,700, varying somewhat for each subtest/cluster, whose ratios are reflective of the 1994 American census. The sample controlled for age, gender, race, geographic region, SES, parent education, and community size (adult norms only). A representative proportion of special education students were also included in the norm sample. The new technical chapter of the manual stated that individuals not proficient in English were not included in the norm sample; however, an operative definition of proficiency was not included.

The WRMT-R renorming process was part of a program aimed at renorming the PIAT-R, KeyMath-R, and the K-TEA in addition to the WRMT-R. According to the updated WRMT-R technical chapter, the four achievement batteries included in the normative updating project measure a number of achievement domains in common: word reading, reading comprehension, mathematics computation, mathematics applications, and spelling. The Visual-Auditory Learning, Word Attack, and Word Comprehension tests of the WRMT-R are not measures of any of the five shared domains; therefore, these tests were normed individually with sample sizes controlled for through the plan's testing assignments.

The advantage of the domain-norming procedure is primarily the increased comparability of test batteries in terms of their common cross-battery domains. In addition, smaller sample sizes are acceptable.

Rasch scaling was used to calibrate item difficulty of the subtests included in the common domains. However, norms for the Visual-Auditory Learning, Word Attack, and Word Comprehension tests were constructed from raw scores with no Rasch analysis.

Criticisms of the WRMT-R

From its debut in 1973 the WRMT has met with controversy. Critics were either impressed by its purported precision and ease of administration (Bannatyne, 1974; Allington, 1976), or disappointed by its theoretical framework and questionable psychometric properties (Houck & Harris, 1976; Dwyer, 1978).

Current critical review continues to raise many of the same issues left outstanding after the revision process. Both editions were criticized for the Eurocentric and gender-stereotype-linked artwork (Dwyer, 1978; Jaeger, 1989).

Of continuing concern is the assessment of fragmented skills rather than the reading act as a whole. The test author apparently views reading as a collection of subskills acting in concert. In reaction, Cooter (1989) argues that "instruments like the WRMT-R are useless and represent a bygone age that viewed the reading act in a fractured or splintered way" (p. 910).

Jaeger (1989) and Eaves (1990) debated whether readiness skills such as Letter Identification should be included in a reading test per se. The recent renorming project pointed out that Letter Identification tended to correlate positively with Word Recognition scores in all age groups but the youngest. The lack of correlation in the early grades was attributed to a curriculum effect, in that young children have a greater portion of their program dedicated to the development of letter recognition skills as opposed to word recognition. As well, both authors had difficulty accepting whether the Auditory-Visual Learning subtest (part of the readiness cluster) actually generalizes to the auditory-visual skills required by written English. For the

purposes of the renorming project, Auditory-Visual learning was not considered part of either the word reading or reading comprehension domains.

Although the author of the test asserts that the Word Identification items represent the culmination of a variety of high frequency word lists, Jaeger (1989) is of the opinion that "the words contained in the test range from those found in early elementary school readers to polysyllabic tongue-twisters likely to be familiar only to persons whose leisure hours are spent perusing the Oxford English Dictionary [sic]" (p. 914). In addition, he states that reading lists of words in isolation has been faulted as a useful measure of sight vocabulary, because examinees cannot use semantic context cues found in authentic reading experiences. Tuinman (1978) adds that some subjects will manage to read a number of the words correctly as sight words, whereas others will rely on phonetic word attack skills. Consequently, this subtest measures different skills with different test-takers. One might anticipate that the person with superior decoding skills would have an advantage on this subtest.

Both editions of the WRMT have drawn criticism for the Word Comprehension subtest. Tuinman (1978) decided that this was more a test of reasoning than word knowledge because of the necessity to ascertain the relationship between the stimulus words in order to complete the analogy. As well, Tuinman felt that poor readers were penalized twice because they may miss additional items due to poor decoding skills, not necessarily because they lack the word knowledge necessary to complete the item. Jaeger (1989) was unsure of the diagnostic and prescriptive value of a low score on this subtest because of the number of possible constructs being measured. This subtest was not considered part of the word reading or reading comprehension domains in the renorming project.

Houck and Harris (1976), Dwyer (1978), and Tuinman (1978) were particularly critical of the Passage Comprehension subtest because of the modified cloze procedure. Houck and Harris (1976) revealed that the test items do not appear to meet Bormuth's (1969) criteria for determining the location of deletions as stated in the manual. In their opinion, "the items resemble more closely a completion test where deletions are made using subjective concepts such as key words" (Houch & Harris, 1976, p. 77).

Bachman (1985) concluded that not all deletions in a given cloze passage measure exactly the same abilities. Among L2 cloze test-takers, Turner (1989) found that other factors in addition to knowledge in a second language do contribute to successful performance on the cloze procedure. Therefore, changing the type of deletions in any given cloze test affects the construct validity of the test. Straying from Bormuth's (1969) criteria may have significantly altered what the Passage Comprehension subtest purports to be measuring.

Although the WRMT-R has been renormed, it is important to remember that many schools and institutions may be slow in updating their test packages as a result of current funding limitations. It is, therefore, prudent to consider the criticisms of both norm editions.

Eaves (1990) and Jaeger (1989) have expressed concerns regarding the sample used for normative purposes in the original WRMT-R. In particular they had difficulty with the author's statement in the manual that selective weighting was used in order to have the sample reflect the variables outlined in the 1980 US Census, but did not specify the procedure or offer any details.

Eaves (1990) had particular difficulty with how Woodcock "indicated that communities were selected according to socio-economic characteristics in a manner that would obtain a sample matching the [1980] US Census data, yet no evidence was provided to verify the outcome of his selection procedure for communities or individuals" (p. 288). This evidence was similarly not forthcoming in the background information accompanying the new norms.

As the reader may recall, the US Census in 1980 restricted itself to distinguishing between "white" and "non-white" ethnic groups. Only with the 1988 US Census did further details regarding ethnic background become available. Therefore, the author used only age, socioeconomic status, and white-non-white variables in his original sample. The current norm sample is more inclusive with regard to race, gender, and community size (adult norms only). However, the reader must bear in mind that the 1994 US Census was still limited to the categories White, Black, Hispanic, and other. It remains questionable whether the norm sample is representative of the Canadian population with whom it is widely compared.

More recent work by Klesmer (1994) has found strong evidence to suggest that the academic/linguistic development of ESL students follows a distinct pattern. At least six years are required for ESL students to approach native English speakers' norms in a variety of areas, and it appears that even after six years, full comparability may not be achieved. With this in mind, it would seem prudent to develop separate achievement criteria for these students based on their age and length of residence. Klesmer's (1994) attempts to renorm the Peabody Picture Vocabulary Test illustrated the difficulty and complexity involved in establishing norms appropriate for individual ESL populations.

Given Klesmer's (1994) findings, it would appear that the crude unidimensional norms presented in the WRMT-R are hardly adequate to account for the potential variability in performance among ESL test-takers. Indeed, it is questionable whether the ESL population is represented in any meaningful way that would allow Woodcock to advocate the use of the test with only minor provisions.

Also of current concern is the validity evidence presented in the manual. As with the former edition, there is lack of evidence of content validity. No

information is offered regarding the possible match between the test and various curricula, or the "items and the psychological processes involved in the reading act for each grade/age level" (Cooter, 1989, p. 912). The manual suggests that the WRMT-R items were developed with the assistance of "outside experts" and experienced teachers, but does not elaborate beyond this overt statement regarding the specific input of these sources. To several critics this information is far from sufficient for claiming that the WRMT-R has any kind of content validity (Cooter, 1989; Jaeger, 1989; Eaves 1990).

In terms of concurrent validity (how this test compares with others in the field for measuring various reading skills), the authors compared the WRMT-R only with the Woodcock-Johnson Reading Tests in the original manual. "It seems implausible that this exercise constitutes a comparison between independent criterion measures" (Cooter, 1989, p. 912). Jaeger (1989) and Eaves (1990) were similarly unimpressed with evidence supplied in the manual.

The original WRMT-R manual did present a table comparing the earlier edition with various other tests, arguing that it remained pertinent because the "psychometric characteristics of the original WRMT (1973) and the WRMT-R are so similar that many generalizations from one to the other can be made" (p. 100). This statement begged the question: Why bother with a new edition then? The author took great pains in the first chapter to illustrate the changes effected from the original test, yet he suggested at that point that the two were so similar that they shared psychometric properties.

By using the domain-norming approach in the renorming project, WRMT-R concurrent validity is established with the PIAT-R and the K-TEA insofar as the subtests in the common domain are concerned.

In response to the criticisms regarding validity levelled at the WRMT-R, Eaves (1990) reported that the test's publisher, American Guidance Service, made available intercorrelations between the WRMT-R Form G and the K-ABC (Kaufman Assessment Battery for Children) and WISC-R (Wechsler Intelligence Scale for Children—Revised) intelligence scales. In both cases correlations were fairly high (range .75 to .96). Eaves (1990) also reported that the studies were without date. One might surmise that in making these particular intercorrelations available, the publisher was making a brilliant marketing move. Both the WISC and the K-ABC are widely used in education and psychology. By aligning the WRMT-R with these popular tools, the publisher implies that they measure the same constructs. This makes the WRMT-R especially attractive for use in diagnosing a variety of learning difficulties that entail the use of an achievement measure. However, without a date on the study, one has to question the credibility of the statistics.

Curiously, it must be noted that several recent studies used to validate the WISC-III (Wechsler Intelligence Scale for Children—Third Edition) involved the WRMT-R, to establish credibility (Hishinuma & Yamakawa, 1993;

Newby, Recht, Caldwell, & Schaefer, 1993; Schwean, Saklofske, Yakulic, & Quinn, 1993). It would appear that the validation arrangement is mutually beneficial.

The K-ABC and the Wechsler tests are not reading tests per se. They purport to measure intelligence; therefore, the only conclusion that can be drawn from the above information is that the K-ABC, WISC-R, and WISC-III may measure the same undefined constructs as does the WRMT-R Form G.

The construct validity of the WRMT-R is not discussed in the original manual. The renorming project, in defining the domains and selecting subtests appropriate for those domains, effectively establishes the constructs being measured. However, three of the WRMT-R subtests did not fall within the confines of the definitions of word reading and reading comprehension and therefore are measuring other, undefined constructs. The paucity of validity evidence should again raise questions regarding whether the test should be used as the basis for decision-making of any kind.

With regard to ESL populations, it should be apparent that reading skills might not be effectively assessed using this tool. The original manual does not define "English reading skills"; therefore, any number of constructs may be measured by this instrument.

Letter Identification might be of some use with persons whose L1 is encoded with characters other than those of English; however, this skill should not be interpreted as reading. The Word Identification subtest may not be a valid measure of acquired English sight words if the individual has developed sufficient phonetic decoding skills. Word Comprehension may at best define some sense of the test-taker's knowledge of synonyms and antonyms. However, because what the analogies subtest is measuring (decoding? verbal reasoning? word knowledge?) is unclear, it may be dangerous to ascribe the construct *word comprehension* to this subtest. The vocabulary chosen for the test as a whole may not realistically reflect everyday language at a variety of age levels or lengths of residence. Last, the Passage Comprehension subtest may not be measuring true comprehension as much as the ability to fill in a blank adequately. It would appear that the US sample used to establish norms for the WRMT-R may not adequately represent the Canadian and/or North American newcomer; therefore, an ESL test-taker may not be adequately represented in the norm population. Essentially, under these conditions, the ESL test-taker's performance would be compared with a majority white American norm. This is hardly fair testing procedure.

To conclude, the *Woodcock Reading Master Test—Revised* is "unsafe at any speed," especially for ESL test-takers.

The Author

Marybeth De Rose is an 18-year veteran of Ontario classrooms; for the past decade she has been a special education teacher in a predominantly ESL community. She is currently completing the final year of her PhD program at the Ontario Institute for Studies in Education.

References

Allington, R.L. (1976). Woodcock reading mastery tests review. *Journal of Reading, 20,* 162-163.

Bachman, L.F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19,* 535-556.

Bannatyne, A. (1974). Woodcock reading mastery tests review. *Journal of Learning Disabilities, 7,* 398-9.

Bormuth, J. (1969). Factor validity for cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4,* 358-365.

Cooter, R. (1989). *Review of the Woodcock Reading Mastery Tests—Revised.* In J.C. Conoley & J.J. Kramer (Eds.), *The 10th mental measurements yearbook* (pp. 910-912). Highland Park, NJ: Gryphon.

Dwyer, C.A. (1978). Woodcock Reading Mastery tests review. In O.K. Buros (Ed.), *The 8th mental measurements yearbook* (pp. 1303-1305). Highland Park, NJ: Gryphon.

Eaves, R.C. (1990). Woodcock Reading Mastery Tests—revised (WRMTR). *Diagnostique, 15,* 277-297.

Hishinuma, E.S., & Yamakawa, R. (1993). Construct and criterion related validity of the WISC-III for exceptional students at risk. *Journal of Psychoeducational Assessment: Advances in Psychoeducational Assessment Monograph Series,* 94-104.

Houck, C., & Harris, L. (1976). Woodcock Reading Mastery Tests review. *Journal of School Psychology, 14*(1), 77-79.

Jaeger, R.M. (1989). Review of the Woodcock Reading Mastery Tests—revised. In J.C. Conoley & J.J. Kramer (Eds.), *The 10th mental measurements yearbook* (pp. 913-916). Highland Park, NJ: Gryphon.

Klesmer, H. (1994). Assessment and teacher perceptions of ESL student achievement. *English Quarterly, 26*(3), 8-11.

Newby, R.F., Recht, D.R., Caldwell, J., & Schaefer, J. (1993). Comparison of WISC-III and WISC-R IQ changes over a two-year time span in a sample of children with dyslexia. *Journal of Psychoeducational Assessment: Advances in Psychoeducational Assessment Monograph Series,* 87-93.

Schwean, V.L., Saklofske, D.H., Yakulic, R.A., Quinn, D. (1993). WISC-III performance of ADHD children. *Journal of Psychoeducational Assessment: Advances in Psychoeducational Assessment Monograph Series,* 56-60.

Tuinman, J.J. (1978). Woodcock reading mastery tests review. In O.K. Buros (Ed.), *The 8th mental measurements yearbook* (pp. 1306-1308). Highland Park, NJ: Gryphon.

Turner, C.E. (1989). The underlying factor structure of L2 cloze test performance in Francophone, university-level students: Causal modeling as an approach to construct validation. *Language Testing, 6,* 172-197.

Woodcock, R.N. (1973). *Woodcock Reading Mastery Tests.* Circle Pines, MN: American Guidance Service.

Woodcock, R.N. (1987). *Woodcock Reading Mastery Tests—revised examiner's manual.* Circle Pines, MN: American Guidance Service.

Woodcock, R.N. (1997). Woodcock Reading Mastery Tests—revised/normative update. Circle Pines, MN: American Guidance Service.

Woodcock, R., & Johnson, M. (1977). *Woodcock-Johnson Psychological Battery.* Allen, TX: DLM Teaching Resources.