

## Georgia State University ScholarWorks @ Georgia State University

Psychology Faculty Publications

Department of Psychology

2005

# The Woodcock Reading Mastery Test: Impact of Normative Changes

Hye Kyeong Pae

*University of Cincinnati - Main Campus*, [hye.pae@uc.edu](mailto:hye.pae@uc.edu)

Justin Coy Wise

*Georgia State University*, [psycjcw@langate.gsu.edu](mailto:psycjcw@langate.gsu.edu)

Paul Cirino

*University of Houston*, [pcirino@uh.edu](mailto:pcirino@uh.edu)

Rose Sevcik

*Georgia State University*, [rsevcik@gsu.edu](mailto:rsevcik@gsu.edu)

Maureen Lovett

*University of Toronto*, [mwl@sickkids.ca](mailto:mwl@sickkids.ca)

*See next page for additional authors*

Follow this and additional works at: [https://scholarworks.gsu.edu/psych\\_facpub](https://scholarworks.gsu.edu/psych_facpub)

 Part of the [Psychology Commons](#)

### Recommended Citation

Pae, H. P., Wise, J.C., Cirino, P.T., Sevcik, R.A., Lovett, M.W., Wolf, M. & Morris, R. (2005). The Woodcock Reading Master Test: Impact of Normative Changes. *Assessment*, 12(3), 347-357.

This Article is brought to you for free and open access by the Department of Psychology at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Psychology Faculty Publications by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact [scholarworks@gsu.edu](mailto:scholarworks@gsu.edu).

---

**Authors**

Hye Kyeong Pae, Justin Coy Wise, Paul Cirino, Rose Sevcik, Maureen Lovett, Robin Morris, and Maryanne Wolf

# The Woodcock Reading Mastery Test

## Impact of Normative Changes

**Hye Kyeong Pae**

**Justin C. Wise**

*Georgia State University*

**Paul T. Cirino**

*University of Houston*

**Rose A. Sevcik**

*Georgia State University*

**Maureen W. Lovett**

*University of Toronto*

**Maryanne Wolf**

*Tufts University*

**Robin D. Morris**

*Georgia State University*

*This study examined the magnitude of differences in standard scores, convergent validity, and concurrent validity when an individual's performance was gauged using the revised and the normative update (Woodcock, 1998) editions of the Woodcock Reading Mastery Test in which the actual test items remained identical but norms have been updated. From three metropolitan areas, 899 first to third grade students referred by their teachers for a reading intervention program participated. Results showed the inverse Flynn effect, indicating systematic inflation averaging 5 to 9 standard score points, regardless of gender, IQ, city site, or ethnicity, when calculated using the updated norms. Inflation was greater at lower raw score levels. Implications for using the updated norms for identifying children with reading disabilities and changing norms during an ongoing study are discussed.*

**Keywords:** Woodcock Reading Mastery Test (WRMT); reading measures; reading disabilities; normative changes; systematic inflation

---

Support for this research was provided by the National Institute of Child Health and Human Development Grant No. HD30970 to Georgia State University, Tufts University, and the Hospital for Sick Children/University of Toronto and by the Research Program Enhancement Fund of Georgia State University. We thank the families of our participants for their cooperation and support. Correspondence concerning this article should be addressed to Robin Morris at the Department of Psychology, Georgia State University, Atlanta, GA 30303; e-mail: psyrdm@langate.gsu.edu.

As measures of academic achievement and cognitive processes are updated on a frequent basis, questions about the comparability and validity between old and new versions of the tests have arisen (Graf & Hinton, 1994; Lyon, 1995; Schultz, 1997; Scott, Bengston, & Gao, 1998; Stockman, 2000; Ukrainetz & Duncan, 2000; Washington & Craig, 1992, 1999). Such updates are usually in response to changes across a spectrum of curricula, educational practices, cultural milieu, and population demographics and subgroups. A recent example is the re-release of the Woodcock Reading Mastery Test–Revised/ Normative Update (WRMT-R/NU; Woodcock, 1998), 11 years after the appearance of the Woodcock Reading Mastery Test–Revised (WRMT-R; Woodcock, 1987). The essential difference between those two editions is that the norms of the WRMT-R/NU have been updated with a new, normative sample of children without any modification of the actual test items.

When a test is re-normed, typically there is a reduction in the resulting standard scores with the new version (see Flynn, 1984, 1987, 1996; Graf & Hinton, 1994; Lyon, 1995; Schultz, 1997). With regard to measures of intelligence, for example, the “Flynn effect” (Flynn, 1984, 1987, 1996) is a well-known phenomenon that suggests that the newer norming cohorts’ intelligence scores rise; in other words, raw scores on intelligence tests have risen steadily and dramatically over time, resulting in a lower standard score when the same individual’s performance is compared using the new test’s normative base with the results of the older previous version.

The decrease in standard scores on newer versions of intelligence tests is highly consistent across studies. The release of Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1991) prompted a number of studies (Graf & Hinton, 1994; Lyon, 1995; Schultz, 1997) comparing it with the older Wechsler Intelligence Scale for Children–Revised (WISC-R; Wechsler, 1974). Although the extent and scope of the differences of standard scores between the old and new editions of the WISC have varied across studies, lower standard scores on the new edition have been consistently found. Lyon (1995), for example, reported that the WISC-III Full Scale, Verbal, and Performance intelligence quotients (IQs) were significantly lower than WISC-R IQs by approximately one third to one half of a standard deviation at the time of reevaluation.

Despite the well-known score decreases at the time of restandardization for intelligence tests, there is a question as to whether the Flynn effect is consistently found in other measures, such as those that examine academic achievement. Scott et al. (1998) compared multiple versions of the Wide Range Achievement Test (WRAT) and found a decrease in standard scores only on the Arithmetic subtest of the WRAT-3 (Wilkinson, 1993). The converse, however, emerged with the Reading and Spelling subtests, on which higher scores were obtained on the newer edition of the WRAT, with standard score differences of 2.3 and 1.5 points, respectively.

This increase in standard scores has also been observed with newer editions of other measures, such as the Peabody Picture Vocabulary Test–III (PPVT-III; Dunn & Dunn, 1997). Although strong correlations were found between the PPVT-R and PPVT-III, indicating that the two editions measured single-word receptive vocabulary skills in a similar way, the modification of content and new normative sample resulted in an increase in standard scores. This increase was largely observed in the performance of children from 4 to 10 years of age, who scored 10 standard score points higher, on average, on the PPVT-III than on the PPVT-R (Ukrainetz & Duncan, 2000). Such changes may be related to several causes: Test items and materials may change over time, subtests may be added or eliminated, and different rules for discontinuing administration or scoring may arise (Dunn & Dunn, 1997; Wechsler, 1991).

In addition, there may be shifts in the demographic composition of the normative base, such as the inclusion of more or fewer individuals at a given age or differences in individuals’ socioeconomic, ethnic, or instructional status, for example, served within special education or general education programs. The normative sample of the PPVT-III (Dunn & Dunn, 1997), for example, included more individuals who were receiving special education services (K. T. Williams & Wang, 1997) compared to the sample recruited for the norming of the PPVT-R (Dunn & Dunn, 1981).

The differential impact of test revisions versus a new normative sample may be examined more clearly in the case of tests that make no revisions, that is, when the actual test items do not change but the instruments are subjected to re-norming. This is the case for the WRMT-R/NU (Woodcock, 1998), along with the Peabody Individual Achievement Test – Revised/Normative Update (Markwardt, 1998), in which the same item set was evaluated with a new standardization sample of children.

In the WRMT-R (Woodcock, 1987), the normative sample consisted of 6,089 individuals aged 5 to 80 from 60 geographically diverse U.S. regions and was closely matched to the 1980 U.S. census data. The normative sample was stratified by geographical distribution, community size, race, sex, and socioeconomic status. The WRMT-R contained neither information about students identified as receiving special educational services nor about gifted students in the manual. The total sample for the WRMT-R/NU (Woodcock, 1998), in contrast, was much smaller than that of the WRMT-R (Woodcock, 1987) and was composed of 3,429 individuals (3,184 K-through-12 students and 245 individuals aged 18 to 22) from 129 U.S. communities. The normative sample for the WRMT-R/NU was also selected to match the demographic composition of the 1990 U.S. census reports, stratified by age, grade, gender, socioeconomic status (examinee or parental education), race/ethnicity (African American, Hispanic, White, and other), and geographic distributions (Northeast, North Central, South, and West).

There appeared to be several additional differences between the two samples in terms of their relevant demographic characteristics that may have influenced scores on the WRMT-R/NU. Although the overall WRMT-R/NU sample was smaller than the WRMT-R, the percentage of the total sample representing the kindergarten-through-12th-grade range was much higher (93% vs. 69%). Whereas the proportion of African American individuals (14.7% to 13.4%) did not change substantially, that of Caucasian (75.6% to 69%) participants declined by 6%, and other ethnic populations changed substantially. On the WRMT-R, for example, the proportions of Native American and Asian American populations were 0.8% and 2.2%, respectively, and Hispanics accounted for 7.3% of the sample represented. In contrast, in the WRMT-R/NU (Woodcock, 1998) normative sample, Native American and Asian participants composed the category listed as “other” (4.9%), whereas the country’s largest growing minority (Edmonson, 1997; Reimers, 1997), Hispanics, accounted for 12.1% of the sample in the K-to-12 pool. The inclusion of more ethnic minority students may have resulted in higher standard scores from the WRMT-R/NU norms than those obtained with the WRMT-R because children from different ethnic backgrounds have been found to perform differently, and sometimes lower, on vocabulary and reading tests (Washington & Craig, 1992) partly due to English as a second language.

The WRMT-R/NU (Woodcock, 1998) sample also included students who were enrolled in gifted (2.3%) and special education (10.8%) programs, including learning disabled (5.9%), speech/language delayed (2.4%), emotionally/behaviorally disturbed (0.8%), mentally retarded (0.8%), physically impaired (0.1%), and other (0.8%). However, the WRMT-R manual did not specify the educational placement of gifted or disabled students. Therefore, it was not clear, on the basis of the WRMT manual, whether the change in these demographic make-ups resulted in the differences of standard scores between the two editions.

Because the W scores, which are the result of a mathematical transformation of raw scores into Rasch-based ability scores, and reference (R) scores are used as intermediate scores in the process of getting standard scores, we also looked at the change that took place in the W scores (a mathematical transformation of raw scores into Rasch-based ability scores) and R scores from the WRMT-R to WRMT-R/NU. The psychometric model of the WRMT development was based on the Rasch model, which provided a mechanism that gave a child’s degree of mastery as well as his or her predicted degrees of mastery on easier and harder material. From a microscopic view on scoring procedure, the WRMT/NU W score underwent a systematic change within and across its subtests. More specifically, 3 points were subtracted from the W score of the Revised version to construct that of the Normative Update edition in the Word Identification subtest, whereas 5 points were added to the Word Attack subtest and 4 points to the Passage Comprehension subtest. Furthermore, the

justification for the systematic W score changes was not explained fully in the manual.

The R score refers to “the median W score minus 100” (Woodcock, 1998, p. 38) for any given grade or age, according to the examiner’s manual. With respect to an R- score change, there were unsystematic variations across subtests and ages in the two editions. Concerning the difference of R scores for 5 years of age between the two editions, for example, the Normative Update version’s R score for Word Identification involved a 6-point addition to that of the Revised edition and 9 points and 5 points for the Word Attack and Passage Comprehension, respectively. For 6 years of age, -7, -8, and -4 points for the Word Identification, Word Attack, and Passage Comprehension, respectively, were observed as changes from the Revised edition to the Normative Update edition, whereas for 7 years of age, 25, -12, and -12 points were shown for the given three subtests from the Revised to the Normative Update editions. However, some age ranges demonstrated no change of the R scores from the Revised to Normative Update norms: R scores for 8-year-olds did not change on either Word Attack or Passage Comprehension, and those for 9-year-olds had no changes on the Word Attack subtest.

Given these changes in the re-norming sample of the WRMT but not in the item content or their ordering, the purpose of this study is to investigate the magnitude of any changes in standard score performance that arose when an individual’s performance was compared on the WRMT-R and WRMT-R/NU norms. Because the WRMT is used frequently by school systems to qualify students for special educational services (Ross, 1995; Schultz, 1997; T. O. Williams & Eaves, 2001) as well as to evaluate intervention effects in research efforts, understanding the impact of such normative changes is essential. Because the identification of reading disabilities (RD) is often based on an assessment of discrepancies between ability and achievement scores (Schultz, 1997; Share & Silva, 2003), a problem may arise when a change in one of the instruments’ norms is made. The resulting new standard scores may potentially qualify or disqualify a student for services. The convergent validity between the two normative versions, along with their concurrent validity with other cognitive and reading measures, was also examined.

## **METHOD**

### **Participants**

Participants were 899 first to third-grade students from public elementary schools who were referred by their teachers concerned about their difficulties in reading and screened for entry into a longitudinal reading intervention project. The sample group was predominantly monolinguals, because the project limited the participants who spoke English as a first language. Participants were from three large metropolitan areas (Atlanta,  $n = 284$ ; Boston,  $n = 283$ ; and Toronto,  $n = 332$ ); 404 participants (44.9%) were African Americans, and 495 (55.1%) were Caucasians; 41.8% ( $n = 376$ ) were female students, and 58.2% ( $n = 523$ ) were male. Their mean age in months was 89.5 ( $SD = 7.1$ ), and ranged from 76 to 109.

### **Measures and Their Characteristics**

Woodcock Reading Mastery Test–Revised (WRMT-R; Woodcock, 1987). Three subtests of the WRMT-R were used: Word Identification, Word Attack, and Passage Comprehension. The Basic Skills cluster (a decoding composite score using Word Identification and Word Attack subtests) and a Total Reading cluster (Short Scale— composed of Word Identification and Passage Comprehension) representing a broader measure of reading ability (Woodcock, 1998) were also derived. Internal consistency reliability coefficients ( $r_{11}$ ) of the WRMT-R obtained by split-half reliability for the first through the third grades ranged from .91 to .98 ( $M = .94$ ; Woodcock, 1987). The description of reliability was not changed in the Normative Update edition and was based on the 1987 norms (Woodcock, 1998).

Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). The K-BIT is an IQ screening

measure of verbal (vocabulary) and nonverbal (matrices) abilities for children and adults aged 4 through 90. Split-half reliability coefficients for the Vocabulary subtest were high, ranging from .89 to .98 ( $M = .92$ ), demonstrating high reliability for children and adolescents ages 4 through 19 ( $M = .91$ ; Kaufman & Kaufman, 1990). Matrices' split-half coefficients ranged from .74 to .95. Lower reliability coefficients occurred for the youngest children in the sample, ages 4 to 6 years ( $M = .78$ ), whereas the values were excellent for ages 20 through 90 ( $M = .94$ ). For the age group of 5 through 12, test-retest reliability coefficients for Vocabulary was .86, for Matrices was .83, and for the K-BIT Composite was .92 (Kaufman & Kaufman, 1990).

WRAT-3 (Wilkinson, 1993). The WRAT-3 includes three subtests that measure reading, spelling, and arithmetic skills. Internal consistency using coefficient alpha for Reading and Spelling subtest ranges from .90 to .95. Test-retest reliability ranges from .91 to .98 were obtained from a sample of 142 individuals from the normative group between the ages of 6 and 16 (Wilkinson, 1993).

## **Procedure**

Testing was conducted by psychologists or doctoral trainees who were trained extensively in test administration. Participants were administered all measures in their schools and were in the first, second, or third grade in the years spanning 1996 to 2000. In the process of RD designation, there were two determinations to identify RD: the low achievement (LA) definition and the regression-based discrepancy definition. The LA criterion was based on the various formulas that consisted of four reading subtests: Word Identification, Word Attack, Passage Comprehension of the WRMT, and Reading subtest of the WRAT. Individuals with a K-BIT IQ Composite score greater than 70 and whose reading skills were equal to or less than a standard score of 85 on the WRMT-R and the WRMT-NU were identified as meeting the LA criteria for RD. The Ability Achievement Regression Corrected Discrepancy (AA-D) definition included participants whose actual reading performance was at least 1 standard error of the estimate below their expected achievement standard score (EASS), calculated based on an average correlation of .60 between measures of reading performance and intellectual ability. To receive an RD designation, students met either LA and/or D criteria.

## **RESULTS**

### **Consistency of WRMT-R and WRMT-R/NU Standard Scores**

We use age-referenced standard scores instead of grade-normed standard scores throughout this article. To measure convergent validity, Pearson correlations were calculated as a measure of the relationship between the standardized scores derived from each set of norms. All correlations were near unitary, as expected (Word Identification,  $r = .97$ ,  $p < .001$ ; Word Attack,  $r = .93$ ,  $p < .001$ ; Passage Comprehension,  $r = .97$ ,  $p < .001$ ; Basic Skills Cluster,  $r = .97$ ,  $p < .001$ ; and Total Reading Cluster,  $r = .98$ ,  $p < .001$ ). As can be seen in Figure 1, the relationship between standardized scores derived from each set of norms is not perfectly linear. Furthermore, this departure from linearity is most exaggerated at the lowest end of the distribution.

To gauge concurrent validity, Pearson correlations were also calculated among the two WRMT standard scores, the WRAT-3, and the K-BIT. Overall, high correlations were found between the WRMT and the Spelling and Reading subtests of the WRAT-3 across the two norms, whereas moderate correlations were observed with the others (i.e., the Arithmetic subtest of the WRAT-3 and the Vocabulary and Matrices subtests of the K-BIT). As seen in Table 1, the correlation coefficients were highly similar for the Revised and Normative Update scores.

**FIGURE 1**  
**Scatterplots of Woodcock Reading Mastery Test (WRMT-R) and Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU) Standard Scores for the Word Attack Subtest, the Word Identification Subtest, the Passage Comprehension Subtest, and the Basic Skills Cluster**

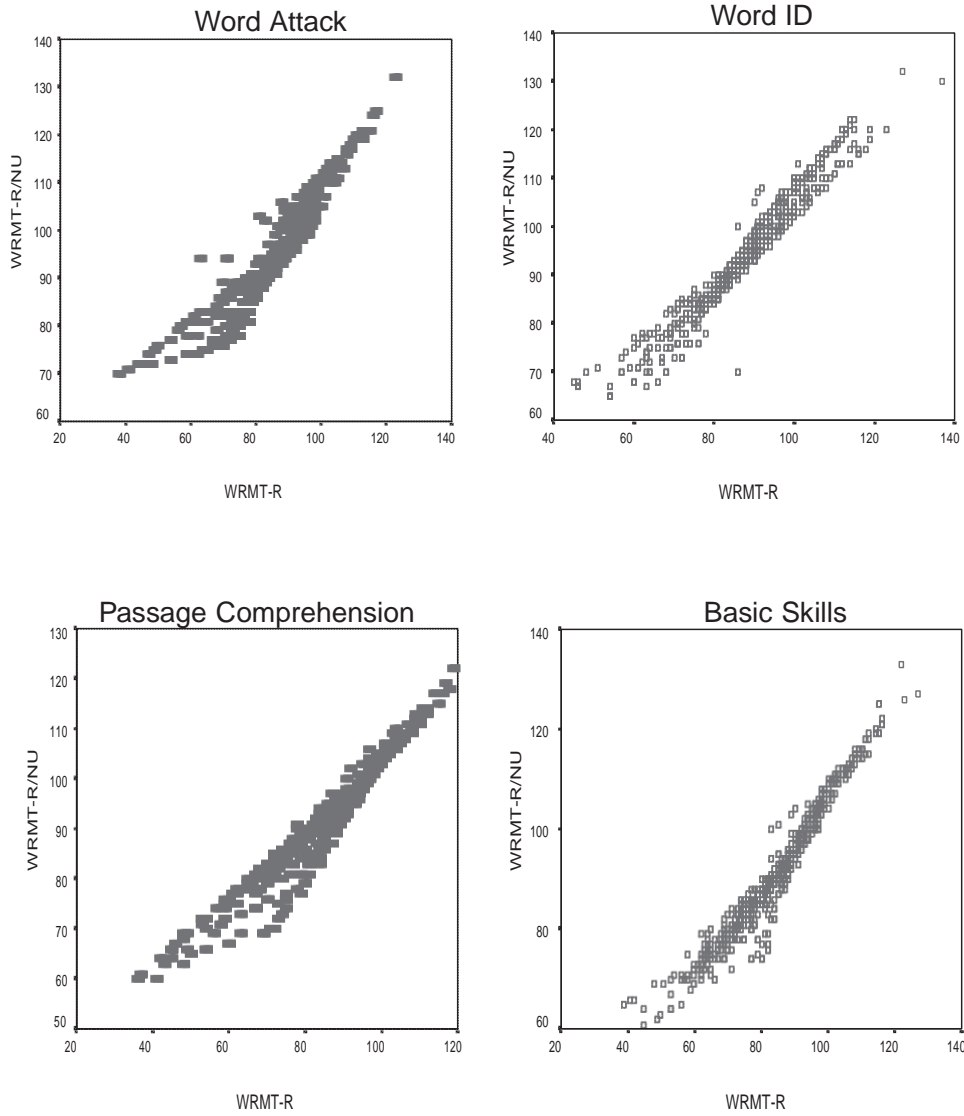


Table 2 presents the mean and standard deviation of each subtest and correlation coefficients between the two norms. Means associated with standardized scores calculated using WRMT-R/NU norms were significantly higher (range across subtests = 4.9 to 8.6 points) than norms calculated using WRMT-R norms.



**TABLE 1**  
**Correlations Between Woodcock Reading Mastery Test (WRMT) Scores and Other Measures (N = 899)**

Criterion Measure	Word Identification		Word Attack		Passage Comprehension	
	Revised	NU	Revised	NU	Revised	NU
WRAT Spelling	.81	.82	.65	.66	.74	.72
WRAT Reading	.88	.87	.72	.71	.81	.76
WRAT Arithmetic	.56	.57	.47	.47	.56	.55
K-BIT Vocabulary	.32	.31	.37	.36	.35	.36
K-BIT Matrices	.26	.36	.25	.27	.32	.33
K-BIT Composite	.37	.35	.39	.39	.41	.42

NOTE: Revised = Revised version of the WRMT; NU = Normative Update version of the WRMT; WRAT = Wide Range Achievement Test; K-BIT = Kaufman Brief Intelligence Test.  
 All correlations:  $p < .001$ .

**TABLE 2**  
**Means, Standard Deviations, Correlation Coefficients, and Paired t Tests of Standard Scores on Both WRMT Norms (N = 899)**

Measures	Edition	M	SD	Range	M Difference	r	t	Effect Size
Word Identification	Revised	87.75	12.57	45-137	6.48	.97*	-63.25*	.82
	NU	94.23	11.46	65-132				
Word Attack	Revised	82.20	12.13	38-123	8.56	.93*	-57.56*	.79
	NU	90.76	11.41	70-132				
Passage Comprehension	Revised	85.83	12.88	36-119	6.07	.96*	-44.79*	.69
	NU	91.90	10.27	60-122				
Basic Skills	Revised	85.31	12.60	39-127	6.88	.97*	-66.55*	.83
	NU	92.19	11.70	61-133				
Total Reading	Revised	87.56	12.34	42-130	4.87	.98*	-55.95*	.78
	NU	92.43	11.51	54-128				

\* $p < .001$ .

Dividing the standard scores of the two norms into their respective percentile ranks yielded a differential distribution between the two WRMT standard scores. For the WRMT-R, 54.7% of the sample scored below the 25th percentile. In contrast, 39.3% of the sample scored below the 25th percentile when standard scores were computed using the WRMT-NU norms. The 26th-to-75th percentile range contained 42.5% of the WRMT-R scores and 54.7% of the WRMT-NU scores. Finally, 2.8% of the WRMT-R scores and 6.0% of the WRMT-NU scores were above the 75th percentile.

## Demographic Variables

To evaluate whether the test norms differed with respect to demographic variables, results were examined using the variables of age, IQ, gender, collection sites, and ethnicity. For age, the sample was divided into children who were 7.2 years old and younger ( $n = 356$ ) and those who were 7.3 years old and older ( $n = 543$ ). The decision to split the sample at 7.2 years of age was made for two reasons: (a) our use of age-referenced standard scores and (b) demarcation of reading level. The demarcation of age at 7.2 was based on the fact that learning to read is developmental in nature and that students older than this should be developing skills beyond the decoding level. Most children younger than this age, however, are just beginning to learn to read and evidence reading skills that are primarily confined to word identification. It was speculated that children in the older age group would demonstrate more automatized and fluent reading abilities, and thus, this discrepancy in ability would provide more power to detect whether differential relations across normative bases existed. A second reason for using age rather than grade to divide the sample was because a grade division would result in a number of children older than 7.2 being subsumed into the first-grade group and children younger

than 7.2 being subsumed into the second-grade group. Because learning to read is developmental in nature, we decided age was a more appropriate variable to make group classifications. For the IQ scores, the data set was divided into below average (below the 25th percentile;  $n = 370$ ), average (25th percentile through 74th percentile;  $n = 483$ ), and above average (75th percentile or higher;  $n = 46$ ) IQ groups.

**TABLE 3**  
**Standard Scores on Word Attack of the Woodcock Reading Mastery Test (WRMT)**  
**and WRMT–Revised/Normative Update (WRMT-R/NU) Based on Demographic Variables**

Variable	Group	n	R			NU			M Difference	r
			M	SD	Range	M	SD	Range		
Ethnicity	AA	404	80.31	11.74	38-116	88.74	10.73	70-124	8.43	.92*
	Caucasian	495	83.75	12.24	38-123	92.42	11.53	70-132	8.67	.94*
Age	≤ 7.2	356	84.79	10.36	63-115	92.99	12.75	78-121	8.20	.96*
	> 7.2	543	80.50	12.90	38-123	89.30	10.19	70-132	8.80	.94*
Gender	Female	376	82.72	12.33	38-123	91.45	11.76	70-132	8.73	.94*
	Male	523	81.83	11.99	38-117	90.27	11.13	70-125	8.44	.92*
Sites	Atlanta	284	82.45	13.32	38-116	90.79	12.53	70-124	8.34	.94*
	Boston	283	85.76	11.35	38-123	94.40	10.46	70-132	8.64	.93*
	Toronto	332	78.96	10.80	45-110	87.64	10.22	72-119	8.68	.90*
IQ	Below average IQ	370	77.25	11.65	38-117	86.09	9.91	70-125	8.84	.88*
	Average IQ	483	85.12	10.84	45-116	93.47	10.86	72-124	8.35	.94*
	High IQ	46	91.43	13.69	57-123	99.91	13.55	77-132	8.48	.96*

\* $p < .001$ .

NOTE: NU = Normative Update version of the WRMT; AA = African American. Analysis was conducted based on the students who had Word Attack scores.

As shown in Table 3, there was very little difference between the correlations computed on the Word Attack subtest across the demographic groups. The Word Attack subtest was analyzed because it evidenced the widest range of scores between the old and new norms, although similar results were found for each of the other subtests. There were systematic differences between the WRMT-R and WRMT-R/NU within the demographic groups that paralleled the differences in the sample as a whole.

To examine these systematic differences in greater depth, a 2 (gender) x 2 (ethnicity) x 2 (age) x 3 (IQ) x 2 (normative edition) repeated-measures ANOVA was conducted using standard scores on the Word Attack subtest as the dependent variable. Results revealed a significant main effect for gender,  $F(7, 876) = 6.35$ ,  $p < .05$ ,  $\eta^2 = .01$ . Female students ( $M = 87.09$ ,  $SD = 11.86$ ) scored significantly higher on the Word Attack subtest than did the male students ( $M = 86.05$ ,  $SD = 11.34$ ).

In addition, a significant main effect for IQ split was found,  $F(2, 876) = 56.74$ ,  $p < .001$ ,  $\eta^2 = .12$ . Tukey HSD post hoc analyses indicated that all three IQ groups (below average, average, and above average) were significantly different from each other ( $p < .05$ ). The above average group had the highest standard scores ( $M = 95.67$ ,  $SD = 13.47$ ) followed by the average ( $M = 89.30$ ,  $SD = 10.70$ ) and below average ( $M = 81.67$ ,  $SD = 10.47$ ) groups, respectively.

Finally, a significant main effect was found for normative base,  $F(1, 876) = 701.44$ ,  $p < .001$ ,  $\eta^2 = .45$ . Standard scores computed using the WRMT-R norms produced significantly lower standard scores ( $M = 82.20$ ,  $SD = 12.13$ ) than standard scores computed using the WRMT-R/NU norms ( $M = 90.76$ ,  $SD = 11.41$ ).

No other significant main effects were found. Analyses did, however, evidence three three-way interactions. A significant three-way interaction was found between normative base, sex, and age,  $F(1, 876) = 4.38$ ,  $p < .05$ ,  $\eta^2 = .01$ . For those students older than 7.2 years of age, the magnitude of

difference between standard scores for males and females was essentially zero under both WRMT-R and WRMT-R/NU norms; however, for those children 7.2 years of age or younger, females evidenced higher standard scores than males under both WRMT-R and WRMT-R/NU norms.

In addition, a significant Normative Base x Age x IQ three-way interaction was found,  $F(2, 876) = 16.14, p < .05, \eta^2 = .04$ . For scores standardized with the WRMT-R norms, in the above average IQ group, there was no difference between the younger and older students. However, in the average IQ group, younger children scored higher than did older children, and in the below average IQ group, this difference was even larger. For the scores standardized with the WRMT-R/NU norms, there was a small difference in the above average IQ group with the younger students scoring slightly higher than the older students. The magnitude of this difference was larger in the below average IQ group and the largest difference between age classifications was seen in the average IQ group.

Finally, a significant Ethnicity x Age x IQ three-way interaction was found,  $F(2, 876) = 4.94, p < .05, \eta^2 = .01$ . In the older children, there was essentially no difference in standard scores between African Americans and Caucasians in the below and average IQ groups; however, in the above average group, African Americans had higher scores than did Caucasians. In the younger students, there was a small difference in scores between the two ethnic groups in the below average IQ group with African Americans scoring higher than Caucasians. Conversely, Caucasians scored higher than African Americans in the average and above average IQ groups, with the greatest difference between the ethnic groups existing in the above average IQ group.

**TABLE 4**  
**Change in Reading Disability Categorization From WRMT-R to WRMT-R/NU**

	WRMT-R (n = 899)		WRMT-R/NU (n = 899)	
LA only	138	(15.3)	131	(14.6)
AA-D only	50	(5.6)	46	(5.1)
LA+AA-D	346	(38.5)	180	(20.0)
Total RD	534	(59.4)	357	(39.7)
Individual change from the WRMT-R to the WRMT-R/NU				
LA+AA-D → LA			80	(8.9)
LA+AA-D → AA-D			34	(3.8)
LA+AA-D → No D/LA			53	(5.9)
LA → LA + AA-D			1	(0.1)
LA → No D/LA			86	(9.6)
AA-D → No D/LA			38	(4.2)
Total change			292	(32.5)
No change			607	(67.5)

NOTE: Percentages in parentheses. WRMT = Woodcock Reading Mastery Test; WRMT-R/NU = Woodcock Reading Mastery Test-Revised/Normative Update; LA = meets low-achievement criteria; AA-D = meets discrepancy criteria; LA+D = meets both low-achievement and discrepancy criteria.

## RD Classification

Table 4 indicates the changes for those students who qualified for an RD diagnosis under either LA or AA-D criteria, utilizing either the WRMT-R or the WRMT-R/NU norms. Using the WRMT-R norms, 534 (59.4%) of the 899 children qualified as having a reading disability. In contrast, using the NU norms, the number of children classified as having a reading disability was reduced to 357 (39.7%). Importantly, no new classifications of RD were made.

The most noteworthy finding when moving from WRMT-R to WRMT-R/NU norms, however, is related to the loss of an AA-D classification. Failing to meet AA-D criteria has important implications for children receiving special education services and will be discussed in more detail in the discussion section. The largest decline of AA-D classifications was in the group who met both LA

and AA-D criteria. The number of children who moved from a classification of LA and AA-D to a classification of LA only was 80 (8.9%) and to a classification of no RD was 53 (5.9%). The number of children who met AA-D criteria only ( $n = 50$  to  $n = 46$ ) did not change substantially. As can be seen in Table 4, however, whereas a number of children moved from a classification of AA-D only to a classification of no RD ( $n = 38$ ; 4.2%), this change is masked by a similar proportion of children moving from a classification of LA and AA-D to a classification of AA-D only ( $n = 34$ ; 3.8%). Thus, a total of 19% ( $n = 171$ ) of the children in this sample lost a classification of AA-D when moving from the WRMT-R to the WRMT-R/NU norms. As for the subgroup that loses eligibility for RD with the use of the WRMT-R/NU edition, the descriptive information is as follows: 79 (46%) African Americans, 92 (54%) Caucasians; 74 (43%) female, 97 (57%) male; 37 (22%) 7.2 years old and younger, 134 (78%) 7.3 years old and older; and 74 (43%) below average IQ, 87 (51%) average IQ, 10 (6%) above average IQ. The conspicuous change was observed in the age split, whereas other variables did not make notable differences. Of the 543 students who were older than 7.2 years of age, 134 (25%) lost an RD designation using the WRMT-R/NU version, whereas only 37 of 356 younger children (10%) lost RD eligibility.

## DISCUSSION

The results of this study demonstrate that changing normative samples can result in systematic changes in a measure's standard scores and, in turn, can have implications for the classification of children with RD. In this study, standard scores calculated for the WRMT/NU increased nearly systematically an average of 7 points when compared to standard scores calculated for the WRMT/R. This increase occurred regardless of the child's gender, IQ, site, or ethnicity. A closer examination of the descriptive data between the two standardizing procedures reveals a systematic inflation averaging 5-to-9 points in standard scores when they are calculated with the Normative Update norms. The mean score differences between the two norms were as follows: Word Identification, 6.5; Word Attack, 8.6; Passage Comprehension, 6.0; Basic Skills, 6.9; and Total Reading, 4.9.

At the same time, correlations between the standard scores did not change from one normative sample to the next, and correlations with other measures (i.e., IQ, reading, and spelling) did not differ importantly between the two normative editions. Because correlations are not affected by mean differences, however, these results are not surprising.

These results are consistent with previous research showing the inverse Flynn effect on academic achievement tests. Higher standard scores have been obtained on the newer editions of achievement tests, such as the WRAT-3 and PPVT-III, whereas lower scores are acquired on the newer versions of the intelligent tests (Flynn effect).

The systematic increase in standard scores was present across a number of demographic variables (i.e., gender, IQ, site, and ethnicity) and was found to be statistically significant between the Revised and Revised/Normative Update normative bases. Because some groups included in the re-norming sample, however, were not represented in the current study (e.g., Hispanic children), we still do not know how the normative change would affect the reading achievement scores of bilingual or non-English dominant immigrant children.

The only exception to the systematic increase in standard scores when moving from the Revised normative base to the Revised/Normative Update normative base occurred for the variable of age. When examining mean differences across subtests, a difference was found between ages in the Passage Comprehension subtest; younger children produced a smaller discrepancy (a 3-point difference) than older children (an 8-point difference). This finding may be the result of a floor effect as reading comprehension is a skill that relies on later developing, more complex reading abilities. The presence of a floor effect may also explain younger children scoring higher on the Word Attack subtest when compared to the older group because there is less variability among the younger children in their decoding skills and low raw scores may result in relatively high standard scores.

The finding that females produced significantly higher scores than males on the Word Attack

subtest is not surprising. It has been well documented that females outperform males early in language development and reading achievement (Berk, 2003). Furthermore, the finding that the above average IQ group significantly outperformed the average group, which subsequently significantly outperformed the below average IQ group also was not surprising. Although recent evidence suggests that IQ is unrelated to the phonological-core deficit documented in children classified with RD (Fletcher et al., 2002), higher IQ provides for compensatory skills that can result in elevated achievement scores.

Additional evidence of differences between age groups comes from the three significant three-way interactions found in the repeated measures ANOVA. In all three interaction effects, the younger age group evidenced the greatest amount of variability in concordance with the other interacting variables when compared to the older age group. As this is an age that is speculated to represent a developmental period in which children are beginning to acquire basic reading skills, this group may evidence a large degree of variability in performance. The effect sizes of these interaction effects, however, were small (range = .01 to .04) and any interpretations based on these results should be made cautiously.

There was some evidence that children whose scores fell at the lower end of the distribution experienced an exaggerated inflation of their scores when compared to children who performed at average or above average levels (see Figure 1). Fifteen percent more children scored below the 25th percentile using the WRMT-R norms than when the WRMT-NU norms were used. Furthermore, the standardized scores of the WRMT-R/NU evidenced a greater restriction in range when compared to the standardized scores of the WRMT-R, resulting at times in a floor effect. For example, the average discrepancy across subtests at the lowest standard score was 22 points.<sup>1</sup> The average discrepancy across subtests at the highest standard score, however, was 5 points.<sup>2</sup>

Under the WRMT-R norms, 534 (60%) children in this sample qualified as RD using either LA or AA-D criteria. In contrast, using the revised norms, the number of children classified as RD was reduced to 357 (40%). As a result, a number of children (20% of the children examined in this study) who had been previously classified as having RD using the WRMT-R norms would no longer qualify under the WRMT-R/NU norms. The greatest change of classification rates, however, occurred in the group meeting criteria for both LA and AA-D scores. The majority (47.9%) of this change resulted in a modification of the LA and AA-D classification to an LA-only classification. This specific type of change poses a concern because federal and state regulations typically mandate that children who have low reading scores but do not meet the discrepancy criteria are not eligible for special education services (Stuebing et al., 2002). Therefore, in this sample, a large portion of children (19%) who met criteria for an IQ discrepant classification using the WRMT-R norms would no longer be eligible for special education services if the WRMT-R/NU norms were used when computing their standard scores.

Because reading measures such as the WRMT frequently are used to assess educational interventions, these results stress the importance of remaining cognizant of the effects of changing norms during an ongoing study when newer versions of assessment measures may be used. As illustrated, actual improvement, that is, improvement over time or intervention effects, can be confounded with spurious improvement due to differences in the versions of the normative tests employed. The decision to continue using an older version of a test is usually related to the addition or removal of test items in the updated version, which in turn may affect participants' scores in repeated measures designs. As seen in this study, however, even when no item changes are made, significant difference in outcomes may occur. In such situations, which scores (e.g., old vs. new norms) are considered more valid indices of a child's reading level are difficult to determine.

Some concerns may exist as to why no attempts were made to test for differences between the two normative bases in terms of *W* scores. The basis for this decision originates from the fact that the systematic changes in *W* scores made from the old to the new normative base were not fully explained in the manual. Therefore, any analyses based on this information would not have produced results that would have allowed for meaningful interpretations.

One cautionary note concerning the sample used in the data analysis involves a restriction of the age range. Children in this sample ranged from 6.3 to 9.1 years of age. The systematic inflation of standard scores seen in younger children when moving from the WRMT-R norms to the WRMT-NU norms may not exist in samples composed of older participants. Although it is possible this inflation of standard scores is restricted to younger children, the size of the sample and the intellectual, ethnic, and regional variability that were present within it make it unlikely that any of the reported findings are an artifact of this restriction of age range. In addition, considering this is the age range during which children are most likely to be identified as RD or as needing special services, the results of these analyses are especially important.

## NOTES

1. Woodcock Reading Mastery Test–Revised/Normative Update: Word Identification = 45/65; Word Attack = 38/70; Passage Comprehension = 36/60; Basic Skills = 39/61; Total Reading = 42/54; and range across subsets = 12 to 32.

2. Woodcock Reading Mastery Test–Revised/Normative Update: Word Identification = 137/132; Word Attack = 123/132; Passage Comprehension = 119/122; Basic Skills = 127/133; Total Reading = 130/128; and range across subsets = 2 to 9.

## REFERENCES

- Berk, L. E. (2003). *Child development* (6th ed.). Boston: Allyn & Bacon.
- Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test—Revised*. Circle Pines, MN: American Guidance Service.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test—III*. Circle Pines, MN: American Guidance Service.
- Edmonson, B. (1997). Hispanic Americans in 2001. *American Demographics*, 19(1), 16-17.
- Fletcher, J. M., Lyon, G. R., Barnes, M., Steubing, K., Francis, D. J., Olson, R. K., et al. (2002). Classification of learning disabilities: An evidence-based evaluation. In R. Bradley, L. Danielson, & D. P. Hallahan, (Eds.), *Identification of learning disabilities: Research to practice* (pp. 185-229). Hillsdale, NJ: Lawrence Erlbaum.
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29-51.
- Flynn, J. R. (1987). Massive IQ gain in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171-191.
- Flynn, J. R. (1996). What environmental factors affect intelligence: The relevance of IQ gains over time. In D. K. Detterman (Ed.), *Current topics in human intelligence, Vol. 5: The environment* (pp. 17-29). Norwood, NJ: Ablex.
- Graf, M. H., & Hinton, R. N. (1994). A 3-year comparison study of WISC-R and WISC-III IQ scores for a sample of special education students. *Educational and Psychological Measurement*, 54(1), 128-133.
- Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman brief intelligence test*. Circle Pines, MN: American Guidance Service.
- Lyon, M. A. (1995). A comparison between WISC-III and WISC-R scores for learning disabilities reevaluations. *Journal of Learning Disabilities*, 28(4), 253-255.
- Markwardt, F. C., Jr. (1998). *Peabody individual achievement test—Revised/Normative update (PIAT-R/NU) manual*. Circle Pines, MN: American Guidance Service.
- Reimers, C. W. (1997). The progress of Mexican and white non-Hispanic immigrants in California and Texas, 1980-1990. *Quarterly Review of Economic and Finance*, 37, 315-343.
- Ross, R. P. (1995). Impact on psychologists of static guidelines for evaluating underachievement. *Learning Disabilities Quarterly*, 18, 43-56.
- Schultz, M. K. (1997). WISC-III and WJ-R tests of achievement: Concurrent validity and learning disability identification. *Journal of Special Education*, 31(3), 377-386.
- Scott, R., Bengston, H., & Gao P. (1998). The 'Flynn effect': Does it apply to academic achievement? *Mankind Quarterly*, 39(1), 109-118.
- Share, D. L., & Silva, P. (2003). Gender bias in IQ-discrepancy and post-discrepancy definitions of reading disability. *Journal of Learning Disabilities*, 36(1), 4-14.
- Stockman, I. J. (2000). The new Peabody Picture Vocabulary Test-III: An illusion of unbiased assessment. *Language, Speech, and Hearing Services in Schools*, 31, 340-353.
- Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classification of reading disabilities: A meta-analysis. *American Educational Research Journal*, 39(2), 469-518.
- Ukrainetz, T. A., & Duncan, D. S. (2000). From old to new: Examining score increases on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools*, 31, 336-339.
- Washington, J., & Craig, H. (1992). Performances of low-income, African American preschool and kindergarten children on the Peabody Picture Vocabulary Test—Revised. *Language, Speech, and Hearing Services in Schools*, 23, 329-333.
- Washington, J., & Craig, H. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test—III. *Language, Speech, and Hearing Services in Schools*, 30, 75-82.
- Wechsler, D. (1974). *Wechsler intelligence scale for children—Revised*. New York: Harcourt, Brace.

- Wechsler, D. (1991). *Wechsler intelligence scale for children-Third edition*. New York: Harcourt Brace.
- Wilkinson, G. S. (1993). *Manual for wide range achievement test* (3rd ed.). Wilmington, DE: Wide Range.
- Williams, K. T., & Wang, J. J. (1997). *Technical references to the Peabody Picture Vocabulary Test-Third Edition (PPVT-III)*. Circle Pines, MN: American Guidance Service.
- Williams, T. O., & Eaves, R. C. (2001). Exploratory and confirmatory factor analyses of the Woodcock Reading Mastery Tests-Revised with special education students. *Psychology in the Schools*, 38(6), 561-567.
- Woodcock, R. W. (1987). *Woodcock reading mastery tests-Revised*. Circle Pines, MN: American Guidance Service.
- Woodcock, R. W. (1998). *Woodcock reading mastery tests-Revised/ Normative Update*. Circle Pines, MN: American Guidance Service.

**Hye Kyeong Pae**, Ph.D., is a research associate in the Department of Psychology at Georgia State University and an adjunct instructor at Georgia Perimeter College.

**Justin C. Wise**, Ph.D., is a research associate in the Department of Psychology at Georgia State University.

**Paul T. Cirino**, Ph.D., is a research assistant professor at the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston.

**Rose A. Sevcik**, Ph.D., is an associate professor in the Department of Psychology at Georgia State University.

**Maureen W. Lovett**, Ph.D., is the director of the Learning Disabilities Research Center and a senior scientist in the Brain and Behavior Program at the Hospital for Sick Children and a professor in the Departments of Pediatrics and Psychology at the University of Toronto.

**Maryanne Wolf**, Ed.D., is a professor in the Eliot Pearson Department of Child Development at Tufts University, director of the Center for Reading and Language Research, and a research scientist in the Department of Psychiatry at McLean Hospital at Harvard Medical School.

**Robin D. Morris**, Ph.D., is vice president for research and Regents professor in the Department of Psychology and Department of Educational Psychology and Special Education at Georgia State University.