2002

Psychometric Stability of Nationally Normed and Experimental Decoding and Related Measures in Children with Reading Disability

Paul Cirino
University of Houston, pcirino@uh.edu

Fontina Rashid
Georgia State University, frashid@atlantaspeechschool.org

Rose Sevcik
Georgia State University, rsevcik@gsu.edu

Maureen Lovett
University of Toronto, mwl@sickkids.ca

Jan Frijters
Brock University, jan.frijters@brocku.ca

See next page for additional authors

Recommended Citation

Cirino, P.T., P.T., Rashid, F.L., Sevcik, R.A., Lovett, M.W., Frijters, J., Wolf, M., & Morris, R. (2002). Psychometric stability of nationally normed and experimental decoding and related measures in children with reading disability. Journal of Learning Disabilities, 35(6), 525-538.

Authors
Paul Cirino, Fontina Rashid, Rose Sevcik, Maureen Lovett, Jan Frijters, Maryanne Wolf, and Robin D. Morris

Psychometric Stability of Nationally Normed and Experimental Decoding and Related Measures in Children with Reading Disability

Paul T. Cirino, Fontina L. Rashid, Rose A. Sevcik, Maureen W. Lovett, Jan C. Frijters, Maryanne Wolf, and Robin D. Morris

Abstract

Achievement and cognitive tests are used extensively in the diagnosis and educational placement of children with reading disabilities (RD). Moreover, research on scholastic interventions often requires repeat testing and information on practice effects. Little is known, however, about the test–retest and other psychometric properties of many commonly used measures within the beginning reader population, nor are these nationally normed or experimental measures comparatively evaluated. This study examined the test–retest reliability, practice effects, and relations among a number of nationally normed measures of word identification and spelling and experimental measures of achievement and reading-related cognitive processing tests in young children with significant RD. Reliability was adequate for most tests, although lower than might be ideal on a few measures when there was a lengthy test–retest interval or with the reduced behavioral variability that can be seen in groups of beginning readers. Practice effects were minimal. There were strong relations between nationally normed measures of decoding and spelling and their experimental counterparts and with most measures of reading-related cognitive processes. The implications for the use of such tests in treatment studies that focus on beginning readers are discussed.

Measurements of achievement and cognitive functioning critically affect decisions regarding the placement, treatment, and study of young children who are diagnosed with reading disabilities (RD). Considering their widespread use and potential impact, it is important that there be sufficient psychometric evidence and support for the use of these measures with such a special population.

One of the most basic aspects of psychometrically appropriate tests is the evidence that they possess sufficient reliability. Although general information on reliability is provided in many of the manuals that accompany nation- ally normed tests, these data frequently exclude individuals with specific reading disabilities from their sampling, which raises concerns about these measures' generalizability and effectiveness with this population (e.g., German, 1989; Markwardt, 1989). This is an especially important consideration given that many measures are de- signed to be administered to individuals across the range of abilities, yet their usage is commonly focused on identifying those children with difficulties in a particular area (whose performance clusters at the low end of the distribution).

There is much focus on internal consistency measures of reliability in test construction in the literature (e.g., Wilkinson, 1993; Woodcock, 1987); how- ever, as noted by Anastasi and Urbina (1997), several different types of reliability and related values (e.g., standard error of measurement) are available, and the choice of coefficients depends on the purpose of the test being studied. Test–retest reliability is particularly important for children with RD because repeat testing is often used not only in school settings for recur- ring placement issues but also in reading intervention studies to monitor progress and to measure meaningful growth in skills over time. Individuals who are experiencing difficulty with reading may exhibit a limited range of reading performance on nationally normed measures (a *floor effect*); that is, measures of this type may be useful only *after* study participants have become readers (Lombardino et al., 1999). Furthermore, it is recognized that intervention may be most effective when it begins at an early age (e.g., Foorman, Francis, Shaywitz, Shaywitz, & Fletcher, 1997). Given the characteristics of children who are often the focus of interventions for RD, one implication is that achieving adequate psychometric properties for a nationally normed test is difficult, which in turn can make it difficult to detect the processes important for making the none-to-some behavioral transition in early reading development. Test-specific practice effects can also be a complicating factor in the interpretation of intervention-induced change over time. Finally, these issues may also be relevant when assessing evolving early reading skills in *all* children, not just those targeted for intervention.

Several studies have examined the test stability over short time intervals of nationally normed achievement tests, such as the *Peabody Individual Achievement Test* (PIAT; Dean, 1979; Naglieri & Pfeiffer, 1983; Smith & Rogers, 1978) and the *Woodcock Reading Mastery Test* (WRMT; McCullough & Zaremba, 1979), in populations with RD. Some more recent studies that focused on stability indices of reliability for achievement measures unfortunately did not include a specific population of individuals with RD (e.g., Shull-Senn, Weatherly, Morgan, & Bradley-Johnson, 1995). Moreover, in the practical application of these measures in clinical and research settings, test–retest intervals cover a much longer period of time than was used in many of these studies. Also, previous test–retest studies have varied considerably in the size and representative- ness of their samples, and some studies have found relatively low levels of reliability. Other authors have explicitly suggested the need for further in- formation regarding the reliability and validity of commonly used achievement tests in exceptional populations (Costenbader & Adams, 1991) and for the development of better measures to address such specific needs.

In addition to those of nationally normed measures, test–retest reliability and practice effects of specifically targeted experimental measures of reading also merit attention. Some experimental tasks measure reading achievement and are frequently designed and used to monitor progress in interventions (e.g., Lombardino et al., 1999; O'Shaughnessy & Swanson, 2000; Stu- art, 1999; Torgesen & Davis, 1996), as they may provide greater sensitivity relative to nationally normed measures at evaluating the lower bounds of reading performance. Other tasks may tap the presumed underlying cognitive processing skills that are important for reading, such as phonology (Lundberg, Frost, & Peterson, 1988), rate and fluency (Wolf, Miller, & Don- nelly, 2000), or metacognitive strategies (Lovett et al., 1994; Lovett et al., 2000). Experimental measures (often designed for evaluating specific aspects of interventions or individuals) may be more sensitive to poor or beginning reading skills and to intervention effects than nationally normed measures, but their psychometric properties and convergent validity to traditional measures, which are the strengths of nationally normed measures, are rarely demonstrated. Practice effects may also be differentially prominent in nationally normed and experimental measures. One strategy to address this issue has been to tie experimental measures specific to an intervention to traditional, nationally standardized criterion measures of achievement, which serve as measures of generalization (see O'Shaughnessy & Swanson, 2000, for an example).

In sum, in order to measure reliable and meaningful change in the development of beginning reading skills in young children, one must consider the psychometric properties (e.g., test– retest reliability) of various types of measures (e.g., nationally normed versus experimental measures of achievement and reading-related cognitive processing skills), as well as considering their relationships. In order to maximize the usefulness of experimental measures, they should bear some relation to the nationally normed measures, but they should also tap independent variance in order to demonstrate their incremental utility.

The present study examined the test– retest reliability of a battery of commonly used, nationally normed tests, along with experimental measures, in a sample of young children with RD who were just learning to read. The stability of the performance of these children on measures of decoding, spelling, and reading-related cognitive skills was assessed during a double- baseline measurement period that began in late spring/early summer and was completed during the early fall before intervention was begun.

We hypothesized that both the nationally normed and the experimental measures would demonstrate adequate test–retest reliability. Some of these values, however, were expected to be reduced relative to populations of same-age children who did not experience reading difficulty, given the test–retest interval (3 to 4 months) and the nature of the sample (beginning, poor readers who had not received any intervention and, therefore, might not yet have developed any reading skills and might exhibit floor effects on some measures). Moreover, we hypothesized that practice effects would be minimal between the two testing times for both the nationally normed and the experimental measures. We did not believe that mere exposure to these tests would produce increases in performance, as the underlying behavior they assessed was not expected to change over the test–retest interval. Furthermore, we hypothesized that the experimental measures of decoding and spelling as well as of reading- related cognitive processing would correlate significantly with the nation- ally normed measures of decoding and spelling at the time just prior to reading intervention but that these two groups of measures would also exhibit independent variance. Finally, we also expected that measures of decoding and spelling in general would be related to measures of reading-related cognitive processes.

Method

Participants

Participants (*N* = 78) were recruited from three large metropolitan area schools (in Atlanta, Boston, and Toronto), identified by their teachers as falling behind their peers in reading, and selected for the study based on their performance on a screening battery that included the *Kaufmann Brief Intelligence Test* (K-BIT; Kaufman & Kaufman, 1990), the *Woodcock Reading Mastery Test–Revised* (WRMT-R; Woodcock, 1987), and the *Wide Range Achievement Test–Third Edition* (WRAT-3;

Wilkinson, 1993). The use of multiple sites was crucial for extending the generalizability of these findings to children of different regional cultures and dialects. There were no differences among sites in terms of their K-BIT or WRMT-R performance; therefore, future analyses combined children from all cities. Inclusion criteria were as follows: English as the primary language, chronological age between 6-6 and 8-6, Grade 1 or 2 at the time of screening, hearing and vision within typical limits, and ethnicity either White (51%) or Black (49%). Children were excluded if they had repeated a grade, achieved a K-BIT Composite score below 70, or had a serious psychiatric or neurological illness. The co-occurrence of a dis- order common in RD populations (e.g., attention-deficit/hyperactivity disorder) did not exclude a child. Children from average (58%) and below- average (42%) socioeconomic levels were included. Thirty percent of the participants were girls, and 17% were left-handed.

Children were considered for inclusion in the study if they met either low achievement (LA) or ability–achievement regression-corrected discrepancy (AA-D) criteria (Fletcher et al., 1994). The K-BIT Composite standard score (M = 92.1, SD = 12.3) was used as a screening measure of intellectual ability, and reading level was established on the basis of any one or more of the following combinations:

1. the average standard score of the WRMT-R Passage Comprehension, WRMT-R Word Identification, WRMT-R Word Attack, and WRAT-3 Reading subtests (Reading Total);
2. the combined standard score of the WRMT-R Word Identification and Word Attack subtests (Basic Skills Cluster);
3. the combined standard score of the WRMT-R Word Identification and Passage Comprehension subtests (Total Reading Cluster).

These different combinations were used to ensure a wide range of participants who met a variety of criteria that have been used in RD research. Participants were included under the LA criteria if their K-BIT Composite score was greater than 70 and their standard score on one or more of the three combinations of reading measures was 85 (approximately the 16th percentile) or less. Participants were included under the AA-D criteria if their actual reading performance was more than 1 standard error of the estimate (approximately 13 standard score points) below their expected achievement standard score (EASS), based on an average correlation of .60 between measures of reading performance and intellectual ability. Most children (64%) met both AA-D and LA criteria for RD; 12% met only AA-D criteria, and the remainder (24%) met only LA criteria.

Procedure

The children completed the measures of reading skills and cognitive ability in April through June (Time 1) and were retested on the same measures in September through October (Time 2). For the WRMT-R, the RAT-3, the Rapid Automatized Naming (RAN) and Rapid Alternating Stimuli (RAS) tests (Denckla & Rudel, 1974, 1976), and the Word Reading Efficiency subtest of the Comprehensive Test of Reading Related Phonological Processes (CTRRPP; Torgesen & Wagner, 1996), the test–retest interval was approximately 4 months. For the other measures, the test–retest interval was approximately 3 months.

Nationally Normed Measures of Decoding and Spelling

Woodcock Reading Mastery Test– Revised. Three subtests of the WRMT-R were used as measures of reading skill. Included were the measures of individual word decoding (Word Identification), nonword decoding (Word At- tack), and comprehension (Passage Comprehension), which requires the participant to supply a missing word in a short passage (cloze task). The standardization sample for the WRMT-R consisted of 6,089 individuals selected to approximate the population distribution of the United States based on 1980 census information. Internal consistency reliabilities (based on raw scores on the odd and even items for either Form G or Form H) for the Word Identification and the Word Attack subtests and Basic Skills Cluster (a de- coding composite) were .98, .94, and .98, respectively. The corresponding standard error of measurement values in *W*-score units was 5.2, 4.9, and 3.6, respectively.

McCullough and Zaremba (1979) found acceptable levels of reliability for the 1973 version of the *Woodcock Reading Mastery Test* (WRMT) in children with learning disabilities (LD) when examining the total test score. A sample of 384 boys with LD and 603 boys without LD (ages 12–17), some of whom had delinquent records, participated in this study. The WRMT total score was found to be reliable both for the LD group (*r* = .88) and for the non- LD group (*r* = .92). No individual sub- test data were reported

Wide Range Achievement Test–3. The WRAT-3 has subtests for individual letter identification and word decoding (Reading), writing letters and words to dictation (Spelling), and mechanically solving oral and written computations (Arithmetic). Each subtest has a pre-achievement section to ameliorate floor effects for younger children and was used in this study specifically because of this characteristic. A sample of 4,433 individuals stratified based on age, regional residence, gender, ethnicity, and socioeconomic level to approximate the 1990 U.S. Census data was used for standardization. The participants in the norming sample ranged in age from 5 to 65, with smaller age intervals between the ages of 5 and 16. A random selection of school-age participants was obtained from public schools, and children from special education classes were also included. The reliability of the WRAT-3 was estimated using four different methods (Wilkinson, 1993): coefficient alpha, alternate form, person separation, and test–retest coefficients. The median coefficient alpha across all three subtests, using both forms, ranged from .85 to .95. The alternate- form correlations for the total sample were .98 for all three subtests. Rasch-Pearson separation indices were also calculated and suggested excellent reliability (.98–.99). Test–retest coefficients corrected for attenuation (interval = 37 days) based on a sample of 142 individuals age 10.5 (4.0) years were .98 (Reading subtest), and .93 (Spelling subtest). No test–retest data were provided on children with RD.

Peabody Individual Achievement Test–Revised (PIAT-R). The PIAT-R Spelling subtest (Markwardt, 1989) is a four-option multiple choice task that measures two concepts: the ability to recognize letters from their names and their sounds, and later, the ability to recognize standard spellings of a spoken word, which can be construed as a measure of orthographic awareness. The PIAT-R was standardized using a population of 1,563 children between kindergarten and Grade 12. The sample was stratified based on geographic region, gender, parent education level, and race to reflect the 1985 U.S. Census data. The majority of the children in the sample were from public schools (91.4%); children from special education classes were not included. The manual for the PIAT-R (Markwardt, 1989) reported the use of four different methods for evaluating reliability: split- half, Kuder-Richardson, test–retest, and item response theory. All four methods yielded acceptable reliability coefficients (most above .90) when evaluating the Spelling subtest data based on age and grade. Reliability coefficients obtained using test–retest reliability (.78–.93) were slightly lower than those obtained by other methods. One sample of 45 randomly selected second graders achieved a test–retest coefficient of .91 for the Spelling subtest (interval = 2 to 4 weeks); the standard error of measurement for this subtest, based on split-half reliabilities for the 123 seven-year-olds in the standardization sample, was 2.3.

Studies of the original PIAT with participants with LD (Dean, 1979; Naglieri & Pfeiffer, 1983; Smith & Rogers, 1978) varied in the type of child identified by the term *learning disabled* or in the composition of the sample. For example, the children in Naglieri and Pfeiffer's study included children with learning disabilities, mental retardation, emotional disturbance, and neurological impairment, and children without a specific placement. Naglieri and Pfeiffer (1983) found a 1-year test–retest coefficient of .60 for the PIAT Spelling subtest, with no mean change in scores over time.

Experimental Measures of Decoding, Spelling, and Reading- Related Cognitive Processes

Given that the nationally normed measures of decoding and spelling de- scribed in the previous section do not comprehensively assess all aspects of reading ability for all populations (e.g., intervention effects, children with RD, underlying cognitive processes), the present study used a variety of experimental measures designed to be sensitive to these important issues. All of these tasks had standardized procedures, and some were locally normed. Some of these experimental tasks emphasized decoding and spelling per se, whereas other experimental tasks emphasized reading-related cognitive processes, including phonological aware- ness, naming skills, and metacognitive strategies. Reliability information was limited or absent for most of these tasks, although many have been widely used in research and clinical activities. Raw total scores (or scores corrected for age) were used for purposes of reliability analysis.

Word Reading Efficiency. Word Reading Efficiency (Torgesen & Wagner, 1996) includes two lists (A & B) of 104 real words that increase in difficulty. The mean number of words read on both forms in 45 seconds provides a measure of reading efficiency. Word Reading Efficiency is a subtest of the CTRRPP, the research version and fore- runner of the recently published *Test of Word Reading Efficiency* (TOWRE; Torgesen, Wagner, & Rashotte, 1999). The TOWRE was normed on 1,507 individuals ages 6 to 24, and the school- age sample was representative of the U.S. population in many respects, including the proportion of students with disabilities (Torgesen et al., 1999). Alternate-form reliability coefficients for the Sight Word Reading Efficiency subtest ranged from .93 to .97 for the age range sampled (6–9), and was .95 in a subgroup of 67 individuals with LD; test–retest reliability over 2 weeks in 29 students ages 6 to 9 was .97. The Word Reading Efficiency word list has some overlap with the WRMT Word Identification subtest but much more with the Sight Word Efficiency subtest of the TOWRE. However, the Word Reading Efficiency (CTRRPP) and Sight Word Efficiency (TOWRE) subtests are not the same, due to changes in the final version and different normative bases. The Word Reading Efficiency subtest used in this study utilized grade-level normative data in computing standard scores.

Normative data were provided by Torgesen and Wagner (1996) based on K, Grade 2, Grade 5, Grade 7, Grade 9, Grade 11, and Florida State University students in Tallahassee, Florida.

Timed Word Reading Tests. Three timed, computer-administered word identification tests (Lovett et al., 1994) measured word identification ability and its level of automaticity. The first of the three tests was the computer Keyword test (120 words), which included regular words with high- frequency spelling patterns; these words were to be explicitly taught to some children. Keywords were adapted from the keyword list introduced by the *Benchmark School Word Identification/ Vocabulary Development Program* (Gaskins, Downer, & Gaskins, 1986). The remaining two computer lists contained words assessing transfer of learning. The computer Test of Transfer included 120 words that were systematically related to words that were to be taught to some children. The computer Content Word test (117 words) contains a list of words carefully constructed to represent the full corpus of uninstructed content. Further descriptions of these measures and their construction can be found in Lovett et al. (1994). The measure used in this study was the raw number of words correctly read. Latency data were available for only a portion of the children in this sample and were not evaluated in the present study.

Challenge Test. The *Challenge Test* (Lovett et al., 2000) is an additional measure of learning transfer, composed of a 117-item word list presented to children on cards. Each challenge word encapsulates a keyword pattern (see Timed Word Reading Tests), along with multiple suffixes and affixes, and has been designed to present children with a difficult decoding task (Lovett et al., 2000). The total number of words correctly read was the measure used for analysis.

Spelling Transfer Test (STT). The *Spelling Transfer Test* (Lovett et al., 1994) measures spelling ability with five lists of words. The first list, Keywords, contains a selection of the 120 words of the computer Keyword test that were to be explicitly taught to some children in the remediation study. The Keyword subtest is followed by four subtests (Vowel, Initial, Final, and Affix) that systematically transform the key words to produce measures of spelling transfer that may indicate generalization of phonological knowledge and strategies to graphemic representations. The four transformations include modifying a vowel (15 items), initial (22 items) and final (18 items) consonants or clusters, and adding affixes or suffixes (21 items). The total number of words correctly spelled in each subtest was the measure used.

Homophone/Pseudohomophone Test. This task was adapted from one used by Olson, Wise, Conners, Rack, and Fulkner (1989) that contained 40 easy and 40 difficult word pairs (see also Olson, Forsberg, Wise, & Rack, 1994, for further details). The present task involved the selection of a correctly spelled word that is paired with a phonetic nonword that sounds the same as the target real word when read in a forced choice format. Because this task involves the recognition of a correctly spelled word, it may be construed more as a measure of orthographic awareness (like the PIAT-R Spelling task). Target words (25) were taken from the WRMT-R Word Identification test and were not the items listed in Appendix B of Olson et al. (1994). The total raw number of correct items was used for analysis.

Elision and Blending Phonemes- Words. The Elision subtest measures the ability to parse and synthesize phonemes, whereas the Blending subtest measures the ability to combine phonemes into words. These subtests are part of the CTRRPP (Torgesen & Wagner, 1996), the same experimental test battery from which the Word Reading Efficiency subtest, described earlier, was derived. This part of the CTRRPP is the prepublication research version and forerunner of the recently published *Comprehensive Test of Phono- logical Processing* (CTOPP; Wagner, Torgesen, & Rashotte, 1999). The CTOPP was normed on 1,656 individuals ages 6 to 24, and the school-age sample was representative of the U.S. population in many respects, including the proportion of students with disabilities (Wagner et al., 1999). Coefficient alpha reliability coefficients for the Elision and Blending Phonemes-Words subtests ranged from .79 to .92 for the age range sampled (6–9) and was .87 (Blending Phonemes-Words) to .91 (Elision) in a subgroup of 67 individuals with LD; test–retest reliability over 2 weeks in 32 students ages 5 to 7 was .88 for both subtests.

The items of the prepublication version and the published version of the CTOPP subtests overlap but are not the same, due to item changes and different normative bases. The Elision and Blending Phonemes-Words subtests used in this study used grade-level normative data to compute standard scores. Normative data were provided by Torgesen and Wagner (1996) and are based on K, Grade 2, Grade 5, Grade 7, Grade 9, Grade 11, and Florida State University students in Tallahassee, Florida.

Sound Symbol Test. The *Sound Symbol Test* (see Lovett et al., 1994, for a complete description of the first and last subtests of this measure) measures the awareness of phonology through the graphic presentation of letters and letter combinations in isolation that the participant is required to vocalize. The four subtests are as follows:

1. Letter-Sound Identification, which includes 37 items that target the sounds that the individual letters of the alphabet can make;
2. Onset Identification, which includes 15 letter combinations that frequently appear at the beginning of a word;
3. Rime Identification, which includes 25 letter combinations that commonly appear at the end of a word; and
4. Sound Combinations, which includes 30 letter strings that frequently occur together and can appear anywhere in a word.

The total score for each subtest was used in analysis.

Rapid Automatized Naming/Rapid Alternating Stimuli (RAN/RAS). The RAN/RAS tasks measure naming speed and accuracy for objects, numbers, letters, and a combination of letters and numbers. Measures include the latency to name 50 items arranged in a 10 × 5 format and the number of errors made. Normative data provided by Wolf, Bally, and Morris (1986) were used for the computation of standard scores for latency within each category. Raw scores were used for the number of errors within each category.

Test of Word Finding. The Picture Naming–Nouns subtest of the *Test of Word Finding* (German, 1989) measures accuracy and speed of confrontation naming of pictures. The standard procedure for this test is for children of different ages to receive different item sets (Items 1–22, 2–29, or 1–29). The data used in this study reflect the number of items correctly named and the total time for completion, recorded for the first 22 items (on which data were available for most participants). Although the *Test of Word Finding* produces standard scores, normative data for this subtest in isolation are not available. In a sample of 20 children over a period of 10 to 14 days, the entire Picture Naming Composite had a test–retest reliability coefficient of .85 (German, 1989).

Strategy Test. The *Strategy Test* (Lovett et al., 1994) is a measure of explicit strategy that a child might use for decoding words. Word identification strategies that are to be taught to children are described and demonstrated for the child and include sounding out, rhyming, using different vowel sounds, breaking up compound words into smaller words, and taking off beginnings and endings of words. The child is then presented with a novel word and is asked to "tell me how you would figure out this word." Children are scored on the following criteria:

1. ability to select an appropriate strategy;
2. proper application of any strategy;
3. presence of metacognitive monitoring of their progress; and
4. correct or partial identification of the target word.

The measure consists of six words that present a range of opportunities for using each of the strategies. The selection, application, and identification variables are scored on a 3-point scale (0 for *failure to use,* 1 for *incomplete usage,* and 2 for *complete and accurate usage*), whereas the monitoring variable is a binary score (0 for *use* and 1 for *nonuse*). The total summary score across words was used in the present analysis.

Test Revisions. Several of the tests described in this section (the *Spelling Transfer Test,* the *Timed Word Reading* Tests, the *Challenge Test,* and the *Strategy Test*) were subsequently revised by selecting a number of items from the original task in order to reduce testing time while maintaining reliability and predictiveness. Item selection was targeted to maximize item discrimination. The ordering of the items on some of these tests was also changed. The re- vised tasks retained overall performance levels, incorporated standardized ceilings, and correlated highly with the original version (mean $r = .97$ for all nine revised tasks, median $r = .98$, range $r = .89–1.00$). All of the analyses reported in this study were per- formed on the revised versions of these tests.

Results

Inspection of the univariate distributions of these measures revealed that many of the variables of interest were nonnormal in their distributions, as judged by the Shapiro-Wilk *W* statistic (SAS Institute, 1988), which ranges from 0 to 1 and is capable of detecting small departures from normality. Furthermore, stem-and-leaf, box, and nor- mal probability plots were examined for each variable, and many were found to be nonnormal. Of the 45 variables (examined at both time points), 13 (29%) were generally normally distributed, and an additional 13 (29%) were moderately nonnormal (typically positively skewed). Nineteen others (42%) were more severely nonnormally distributed (typically positively skewed and highly kurtotic). These significant floor effects and limited variances were not particularly surprising in this sample. The variables whose distributions departed most from normality were found nearly entirely in the *Sound Symbol Test,* the *Spelling Transfer Test,* the error scores from the RAN/RAS, the *Timed Word Reading* Tests, and the *Challenge Test.* The significant floor effects and limited variances observed in many of the variables used in this study clearly would be expected to affect correlation coefficients using these measures.

For all the nationally normed and experimental measures, test–retest reliability coefficients were computed for raw and standard scores (if both were applicable). Standard scores may result in more extreme floor effects when a group is selected for their extreme scores on a measure—as are children with RD—and raw scores may at times offer more range in the data. For most procedures, the average duration be- tween testing times was approximately 85 days, with the exception of the WRMT-R, WRAT-3, and RAN/ RAS tasks, which were readministered approximately 130 days after initial testing.

Change in Performance From Time 1 to Time 2

To examine the change in performance that resulted from the passage of time, means and difference scores for all measures are provided in Tables 1 (raw scores) and 2 (standard scores, if avail- able). In general, raw and standard score performance on the nationally normed decoding and spelling measures (WRMT-R, WRAT-3, PIAT-R) changed from Time 1 to Time 2 on all 11 variables, except raw scores for WRMT-R Word Attack and PIAT-R Spelling. However, where gains occurred on raw score measures, these gains were rather small in absolute, educationally relevant terms (e.g., approximately one raw score unit). Despite these small raw score increases, standard scores on all six measures actually decreased over the same time period, a finding explained by age norm changes in expectations.

Mean change scores from Time 1 to Time 2 for experimental measures of decoding and spelling (CTRRPP Word Reading Efficiency, *Timed Word Reading* Tests, *Challenge Test, Homophone/ Pseudohomophone* test, and *Spelling Transfer Test*) are also provided in Tables 1 and 2. The Word Reading Efficiency subtest of the CTRRPP exhibited the same pattern as the nationally normed measures (small increases in raw score units and a corresponding decrease in standard score units); standard scores were not available for any of the other 10 measures. However, there was very little evidence of increases in performance (only 3 of 10 measures); the raw score improvements that did occur were again very small.

Mean change (difference) scores for the experimental measures of reading- related cognitive processes, including phonological awareness (CTRRPP Elision and Blending Phonemes-Words subtests and the *Sound Symbol Test*), naming speed (RAN/RAS and the Picture Naming–Nouns subtest of the *Test of Word Finding*), and metacognitive processes (*Strategy Test*), are also pro- vided in Tables 1 and 2. As with the de- coding and spelling measures, there was little change overall from Time 1 to Time 2. Twelve of the 16 measures evidenced no significant change in raw score performance over time. Some measures did show significant positive differences in performance from Time 1 to Time 2; however, all of these gains were again rather small in absolute and educational terms (e.g., one to two raw score units). Of the measures that had standard scores, only the Blending Phonemes-Words subtest of the CTRRP demonstrated a significant increase in standard score units from Time 1 to Time 2 (see Table 2). None of the other measures exhibited a significant change over time.

As is clear from Tables 1 and 2, there were few mean changes in students' scores from summer (Time 1) to fall (Time 2). The largest changes involved a *decrease* in standard score units on academic measures. To further demonstrate the lack of systematic increases in scores (i.e., practice effects), the distribution of the difference scores over time for each variable of interest was examined. These analyses revealed that 19 out of 45 (42%) difference scores were normally distributed according to the box and stem-and-leaf plots, and the stringent Shapiro-Wilk *W* statistic in SAS and as presented in Tables 1 and 2. The difference scores were centered close to or below zero, indicating no systematic improvement over time. The remaining variables, which were not normally distributed, in almost every case were highly kurtotic (peaked) and again centered near zero. For the nationally standardized and the experimental measures of decoding and spelling and for the experimental reading-related cognitive process measures, there was generally minimal change over time. Where substantial differences did occur, these were exclusively concerned with decreases in standard score units on nationally standardized decoding and spelling measures, suggesting that these students failed to keep pace with their peers without RD, at least when not involved in a classroom setting (during their summer break).

Test–Retest Reliability

To examine the consistency of scores for participants over time, test–retest reliability coefficients were computed for the nationally normed measures and the experimental measures of de- coding and spelling and reading- related cognitive processes. Table 3 includes test–retest coefficients for the current sample ($N = 78$), in raw score and standard score units. We consider test–retest coefficients of .80 or higher to indicate *excellent* reliability and coefficients between .60 and .79 to indicate *good* reliability; coefficients below .60 are considered *weak*.

Test–retest coefficients for the nationally normed measures (WRMT-R, WRAT-3, and PIAT-R) were all highly significant for both raw and standard scores. Of the 11 values, 6 (55%) exhibited excellent reliability, and the other 5 (45%) measures had good reliability; none showed only weak reliability. The mean test–retest reliability coefficient of these nationally normed measures of decoding and spelling was $r = .82$ (median $r = .82$).

Each of the test–retest coefficients for the experimental decoding and spelling measures (CTRRPP Word Reading Efficiency, *Timed Word Reading* Tests, *Challenge Test, Spelling Transfer Test,* and *Homophone/Pseudohomophone* test) also were significant. Of these 12 values, 5 (42%) exhibited excellent reliability, and 4 others (33%) showed good reliability ($r = .62$–.69); the mean test–retest reliability value for the experimental measures of decoding and spelling was $r = .69$ (median $r = .68$). Clearly nonnormal distributions negatively affected the three remaining measures (e.g., the Initial List of the *Spelling Transfer Test,* the *Challenge Test,* and the standard score from the CTRRPP Word Reading Efficiency subtest), whose test–retest coefficients were weak. When the three measures with non- normal distributions were removed, the mean test–retest value for the experimental measures of decoding and spelling was $r = .78$ (median $r = .80$), highly similar to the value for the nationally normed measures of decoding and spelling.

Overall test–retest coefficients for experimental language processing measures related to reading (CTRRPP Elision and Blending Phonemes-Words subtests, *Sound Symbol Test,* RAN/ RAS, TOWF Picture Naming–Nouns, and the *Strategy Test*) appeared slightly lower than those obtained for the nationally normed measures, although most were again highly significant. Five of 22 (23%) measures exhibited excellent reliability, and an additional 12 (55%) showed good values. The mean test–retest coefficient for the experimental measures of reading-related cognitive processes was $r = .67$ (median $r = .73$). Weak values were associated with the RAN Objects Latency (raw and standard scores) and with the error scores from the RAN/RAS (Numbers, Objects, and Numbers/ Letters); however, test–retest coefficients may not be an appropriate index for these measures, given the low overall base rate of errors on several of the RAN/RAS tasks and their consequent highly nonnormal distributions. When the raw error scores from the RAN/ RAS tests were removed from the experimental cognitive process measures, the mean value for the remain- der was $r = .74$ (median $r = .75$).

Correlations Among Measures

To more fully understand the nature of the experimental measures used in the present study, their correlations with nationally normed measures of spelling and decoding were examined. Specifically, the correlations of the WRMT-R Word Identification and WRAT-3 Reading subtests with the CTRRPP Word Reading Efficiency and *Timed Word Reading* tests were examined. Overall, there were strong correlations between these measures (eight correlations, mean $r = .74$, median $r =.73$, range $r = .54$–$.95$). Significant but somewhat lower correlations were obtained between nationally normed measures of spelling (WRAT-3 and PIAT-R Spelling subtests) and experimental spelling tests (*Spelling Transfer Test* and *Homophone/Pseudohomophone* test; 12 correlations, mean $r = .51$, median $r = .49$, range $r = .35$–$.73$), which may reflect the different formats employed (e.g., forced choice, multiple choice, or written spelling) as well as the complexity of spelling skills and the possibility that different measures tapped diverse skills.

Correlations between nationally normed and experimental measures of decoding and spelling and measures of cognitive processing related to reading can be found in Table 4. All correlations were in the expected direction (i.e, with better or faster language processing skills related to better decoding and spelling performance). The correlations of rapid naming skills (RAN, TOWF Picture Naming–Nouns) with both nationally normed and experimental decoding skills were significant, with similar significant correlations with spelling performance (for all correlations, mean absolute $r = .32$, median $r = .29$). An exception was RAN object naming, which had low correlations with nearly all measures of decoding and spelling (without this variable, mean absolute $r = .37$, median $r = .42$). The correlations of the *Strategy Test,* designed to tap metacognitive processes, with the nationally normed and experimental measures of decoding and spelling were mostly significant and strong (mean $r = .41$, median $r = .46$). The correlations between measures of phonological awareness (CTRRPP Elision and Blending Phonemes-Words, *Sound Symbol Test*) and decoding and spelling (both nationally normed and experimental) were also examined; in general, these correlations were stronger than for rapid naming or metacognition (mean $r = .55$; median $r = .57$).

TABLE 1

Raw Scores on Nationally Normed and Experimental Measures for Time 1, Time 2, and Difference Scores

		Time 1		Time 2		Time 1–Time 2			
Measure	***n***	***M***	***SD***	***M***	***SD***	***M***	***SD***	**Range**	***SE*** ^{a}
WRMT-R									
Word Identification	78	13.96	11.2	15.56	11.9	1.60*	3.87	–4.0–14.0	2.66
Word Attack	78	2.03	3.4	1.88	3.3	–0.14	2.00	–6.0–9.0	1.36
WRAT-3									
Reading	78	17.60	2.7	18.10	3.1	0.50*	1.63	–3.0–6.0	1.19
Spelling	77	15.64	2.7	16.25	2.3	0.61*	1.73	–3.0–6.0	1.07
PIAT-R									
Spelling	71	25.56	8.0	26.00	8.0	0.44	4.86	–12.0–14.0	3.39
CTRRPP									
Word Reading Efficiency	78	7.33	5.8	9.26	7.8	1.94**	3.48	–4.5–14.5	2.35
TWR									
Keywords	65	8.63	8.2	10.08	9.5	1.45**	3.19	–6.0–12.0	2.12
Content Words	65	0.29	1.0	0.58	2.0	0.29	1.59	–2.0–11.0	1.24
Test of Transfer Words	65	1.75	4.6	2.34	5.7	0.59*	2.05	–3.0–9.0	1.39
Challenge Test									
Total	56	0.50	2.6	0.55	3.5	0.05	3.49	–15.0–21.0	2.76
STT									
Key List	69	1.39	2.3	1.29	2.1	–0.10	1.39	–5.0–6.0	0.96
Initial List	71	0.69	1.1	1.21	2.1	0.52*	1.96	–4.0–11.0	1.63
Vowel List	71	0.86	1.6	1.27	2.5	0.41	1.83	–4.0–12.0	1.40
Affix List	71	0.13	0.5	0.08	0.4	–0.04	0.31	–2.0–1.0	0.16
Final List	71	0.75	1.5	0.94	1.8	0.20	1.35	–3.0–8.0	1.01
HP/PHP									
Total	71	16.46	4.2	16.99	3.8	0.52	3.46	–7.0–11.0	2.33
CTRRPP									
Elision	73	8.05	3.6	7.74	4.2	–0.32	2.78	–6.0–7.0	2.08
Blending Phonemes-Words	72	8.03	4.8	8.89	4.7	0.86*	2.52	–6.0–8.0	1.77
Sound Symbol Test									
Letter-Sound Identification	72	19.06	7.9	19.93	8.4	0.88	5.49	–14.0–20.0	4.00
Onset Identification	71	2.90	4.4	2.96	4.5	0.06	3.32	–12.0–8.0	2.40
Rime Identification	65	2.80	4.6	3.35	5.1	0.55	3.31	–8.0–14.0	2.45
Sound Combinations	72	3.67	3.1	3.89	4.1	0.22	3.14	–7.0–10.0	2.42
RAN Numbers									
Latency	78	54.41	16.3	53.34	16.3	–0.87	13.16	–39.0–33.0	9.35
Errors	78	0.83	1.9	0.54	1.1	–0.29	1.93	–12.0–3.0	0.96
RAN Letters									
Latency	78	58.64	20.5	58.22	21.5	–0.42	15.50	–34.0–68.0	11.15
Errors	78	1.19	2.6	1.90	4.2	0.71*	2.46	–7.0–11.0	1.67
RAN Objects									
Latency	78	63.56	12.3	64.13	16.3	0.56	13.90	–50.0–65.0	10.80
Errors	78	0.68	1.0	0.53	0.8	–0.15	1.74	–4.0–3.0	0.76
RAS Numbers/Letters									
Latency	76	80.08	37.4	80.71	43.8	0.63	30.36	–137.0–117.0	22.77
Errors	77	2.35	3.1	2.32	2.7	–0.03	3.33	–16.0–11.0	2.18
TOWF Picture Naming									
Nouns—Items 1–22	61	12.08	3.8	13.59	4.1	1.51*	2.20	–3.0–7.0	1.58
Strategy Test									
Total	57	5.79	7.4	7.56	7.4	1.77*	4.64	–8.0–16.0	3.30

Note. Nationally normed measures: WRMT-R = *Woodcock Reading Mastery Test–Revised* (Woodcock, 1987); WRAT-3 = *Wide Range Achievement Test,* 3rd ed. (Wilkinson, 1993); PIAT-R = *Peabody Individual Achievement Test–Revised* (Markwardt, 1989). Experimental measures: CTRRPP = *Comprehensive Test of Reading Related Phonological Processes* (Torgesen & Wagner, 1996); TWR = *Timed Word Reading Tests* (Lovett et al., 1994); *Challenge Test* (Lovett et al., 2000); STT = *Spelling Transfer Test* (Lovett et al., 1994); HP/PHP = *Homophone/Pseudohomophone Test* (Olson et al., 1994); *Sound Symbol Test* (Lovett et al., 1994); RAN = *Rapid Automatized Naming* (Denckla & Rudel, 1976); RAS = *Rapid Alternating Stimuli* (Denckla & Rudel, 1976); TOWF = *Test of Word Finding* (German, 1989); *Strategy Test* (Lovett et al., 1994).

[a]Standard error of measurement using the *SD* of Time 2 raw scores and test–retest reliability coefficients (see Table 3).
**p* < .05. ***p* < .0001.

TABLE 2
Standard Scores on Nationally Normed and Experimental Measures for Time 1, Time 2, and Difference Scores

		Time 1		Time 2		Time 1–Time 2			
Measure	***n***	***M***	***SD***	***M***	***SD***	***M***	***SD***	**Range**	***SE*[a]**
WRMT-R									
Word Identification	78	80.65	9.3	75.11	10.0	–5.44**	3.95	–14.0–6.0	2.83
Word Attack	78	76.82	8.1	71.49	10.3	–5.33**	6.55	–32.0–7.0	4.96
Basic Reading	78	78.91	9.2	71.79	10.0	–7.11**	4.20	–14.0–5.0	3.00
WRAT-3									
Reading	78	80.99	8.4	76.76	7.9	–4.23**	6.20	–18.0–13.0	4.25
Spelling	77	82.55	9.3	79.96	7.3	–2.58*	6.55	–16.0–16.0	3.95
PIAT-R									
Spelling	71	82.17	8.9	79.85	8.4	–2.32*	6.17	–17.0–14.0	4.22
CTRRPP									
Word Reading Efficiency	78	69.40	8.1	66.70	7.6	–2.70*	8.25	–20.3–13.5	5.62
Elison	73	75.43	13.5	74.28	15.6	–1.18	10.43	–22.5–26.3	7.79
Blending Phonemes-Words	72	84.68	21.0	88.48	20.9	3.80*	11.14	–26.5–35.3	7.80
RAN Latency									
Numbers	78	61.50	23.1	60.12	22.6	–1.38	17.32	–41.0–35.0	12.15
Letters	78	62.41	25.5	61.59	24.5	–0.82	20.50	–57.0–49.0	12.75
Objects	78	77.67	17.2	75.83	19.3	–1.83	17.45	–56.0–62.0	12.95
RAS Latency									
Numbers/Letters	76	69.25	22.3	66.11	24.3	–3.14	16.75	–38.0–45.0	12.17

Note. Standard scores were available only for the measures listed here. Nationally normed measures: WRMT-R = *Woodcock Reading Mastery Test–Revised* (Woodcock, 1987); WRAT-3 = *Wide Range Achievement Test,* 3rd ed. (Wilkinson, 1993); PIAT-R = *Peabody Individual Achievement Test–Revised* (Markwardt, 1989). Experimental measures: CTRRPP = *Comprehensive Test of Reading Related Phonological Processes* (Torgesen & Wagner, 1996); RAN = *Rapid Automatized Naming* (Denckla & Rudel, 1976); RAS = *Rapid Alternating Stimuli* (Denckla & Rudel, 1976).
[a]Standard error of measurement using the *SD* of Time 2 raw scores and test–retest reliability coefficients (see Table 3).
**p* < .05. ***p* < .0001.

TABLE 3

Test–Retest Coefficients for Raw and Standard Scores on Nationally Normed and Experimental Measures, with Mean Testing Intervals

		Testing interval[a]		Test–retest coefficient	
Measure	***n***	***M***	***SD***	**Raw score**	**Standard score**
WRMT-R		132	22.9		
Word Identification	78			0.95**	0.92**
Word Attack	78			0.83**	0.77**
Basic Skills Cluster	78				0.91**
WRAT-3		132	27.1		
Reading	78			0.85**	0.71**
Spelling	77			0.78**	0.71**
PIAT-R		85	16.0		
Spelling	71			0.82**	0.75**
CTRRPP					
Word Reading Efficiency	78	133	23.7	0.91**	0.46**
TWR					
Keywords	65	94	30.5	0.95**	
Content Words	65	94	60.5	0.62**	
Test of Transfer	65	94	30.5	0.94**	
Challenge Test		94	30.5		
Total	56			0.38*	
STT		85	16.7		
Key List	69			0.80**	
Initial List	71			0.41**	
Vowel List	71			0.69**	
Affix List	71			0.82**	
Final List	71			0.67**	
HP/PHP		85	18.4		
Total	71			0.63**	
CTRRPP					
Elision	73	85	17.2	0.75**	0.75**
Blending Phonemes-Words	72	85	17.2	0.86**	0.86**
Sound Symbol Test		77	31.2		
Letter-Sound Identification	72			0.77**	
Onset Identification	71			0.72**	
Rime Identification	65			0.77**	
Sound Combinations	72			0.66**	
RAN		133	22.3		
Numbers					
Latency	78			0.67**	0.71**
Errors	78			0.29*	
Letters					
Latency	78			0.73**	0.66**
Errors	78			0.84**	
Objects					
Latency	78			0.56**	0.55**
Errors	78			0.16	
RAS Numbers/Letters		133	22.3		
Latency	76			0.73**	0.75**
Errors	77			0.34**	
TOWF Picture Naming		83	18.2		
Nouns—Items 1–22	61			0.85**	
Strategy Test		82	15.8		
Total	57			0.80**	

Note. Standard scores were not available for some experimental measures due to a lack of normative data. The Basic Skills Cluster of the WRMT-R is a composite standard score; therefore, no raw score was available for this measure. Nationally normed measures: WRMT-R = *Woodcock Reading Mastery Test–Revised* (Woodcock, 1987); WRAT-3 = *Wide Range Achievement Test,* 3rd ed. (Wilkinson, 1993); PIAT-R = *Peabody Individual Achievement Test–Revised* (Markwardt, 1989). Experimental measures: CTRRPP = *Comprehensive Test of Reading Related Phonological Processes* (Torgesen & Wagner, 1996); TWR = *Timed Word Reading Tests* (Lovett et al., 1994); *Challenge Test* (Lovett et al., 2000); STT = *Spelling Transfer Test* (Lovett et al., 1994); HP/PHP = *Homophone/Pseudohomophone Test* (Olson et al., 1994); *Sound Symbol Test* (Lovett et al., 1994); RAN = *Rapid Automatized Naming* (Denckla & Rudel, 1976); RAS = *Rapid Alternating Stimuli* (Denckla & Rudel, 1976); TOWF = *Test of Word Finding* (German, 1989); *Strategy Test* (Lovett et al., 1994).

[a] Mean interval in days between testing at Time 1 and testing at Time 2.

*$p < .05$. **$p < .0001$.

TABLE 4

Correlations Between Raw Scores at Time 2 of Decoding/Spelling and Reading-Related Cognitive Process Measures

	Reading-related cognitive process measure[a]											
Decoding/spelling measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
WRMT-R												
Word Identification	–.44**	–.44**	–.14	–.47**	.50**	.42**	.54**	.68**	.65**	.73**	.68**	.55**
Word Attack	–.28*	–.29*	–.14	–.28*	.53**	.60**	.72**	.64**	.67**	.91**	.70**	.51**
WRAT-3												
Reading	–.48**	–.50**	–.18	–.56**	.55**	.46**	.58**	.70**	.64**	.72**	.59**	.48**
CTRRPP												
Word Reading Efficiency	–.44**	–.45**	–.22*	–.43**	.46**	.39*	.48**	.61**	.59**	.64**	.60**	.52**
TWR												
Keywords	–.41*	–.44**	–.10	–.44**	.47**	.41**	.56**	.65**	.63**	.77**	.75**	.47**
Content Words	–.19	–.24*	–.02	–.20	.32*	.35*	.29*	.40*	.41**	.57**	.52**	.19
Test of Transfer	–.30*	–.35*	–.13	–.28*	.39*	.50**	.52**	.55**	.67**	.80**	.65**	.30*
Challenge Test	–.16	–.18	–.01	–.15	.20	.22	.19	.28*	.40*	.53**	.37*	.14
WRAT-3												
Spelling	–.52**	–.51**	–.19	–.51**	.45**	.36*	.58**	.61**	.53**	.67**	.59**	.44**
PIAT-R Spelling	–.39*	–.38*	–.11	–.44**	.48**	.28*	.49**	.59**	.45**	.55**	.53**	.49**
HP/PHP	–.28*	–.29*	–.15	–.31*	.38*	.23*	.28*	.46**	.49**	.45**	.50**	.43**
STT												
Key List	–.29*	–.29*	–.12	–.28*	.47**	.57**	.68**	.57**	.64**	.79**	.61**	.54**
Initial List	–.24*	–.29*	–.11	–.27*	.56**	.53**	.72**	.58**	.60**	.79**	.47**	.48**
Vowel List	–.23*	–.22	–.11	–.24*	.47**	.59**	.69**	.52**	.60**	.79**	.52**	.39*
Affix List	–.13	–.16	–.01	–.14	.33*	.33*	.36*	.36*	.33*	.55**	.49**	.26*
Final List	–.22	–.23*	–.12	–.26*	.50**	.54**	.73**	.53**	.52**	.73**	.42**	.33**

Note. WRMT-R = *Woodcock Reading Mastery Test–Revised* (Woodcock, 1987); WRAT-3 = *Wide Range Achievement Test,* 3rd ed. (Wilkinson, 1993); CTRRPP = *Comprehensive Test of Reading Related Phonological Processes* (Torgesen & Wagner, 1996); TWR = *Timed Word Reading Tests* (Lovett et al., 1994); *Challenge Test* (Lovett et al., 2000); PIAT-R = *Peabody Individual Achievement Test–Revised* (Markwardt, 1989); HP/PHP = *Homophone/Pseudohomophone Test* (Olson et al., 1994); STT = *Spelling Transfer Test* (Lovett et al., 1994).

[a] Coded as follows: 1 = *Rapid Automatized Naming* (RAN; Denckla & Rudel, 1976) Numbers Latency; 2 = RAN Letters Latency; 3 = RAN Objects Latency; 4 = *Rapid Alternating Stimuli* (RAS; Denckla & Rudel, 1976) Numbers/Letters Latency; 5 = *Test of Word Finding* (TOWF; German, 1989) Picture Naming—Nouns; 6 = CTRRPP Blending; 7 = CTRRPP Elision; 8 = *Sound Symbol Test* (Lovett et al., 1994) Letter-Sound; 9 = Sound Symbol Test Onset; 10 = Sound Symbol Test Rime; 11 = Sound Symbol Test Sound Combinations; 12 = *Strategy Test* (Lovett et al., 1994).

**p* < .05. ***p* < .0001.

Discussion

A primary result of this study was that most of the tests used, whether nationally normed or experimental, demonstrated good to excellent test–retest reliability in this sample of children with RD (see Table 3) despite the presence of several factors that may have attenuated stability values. For example, these children were struggling just to begin to read, as evidenced by their overall weak performance not only on reading measures but also on spelling and cognitive processing measures in areas typically related to RD, such as phonological awareness (e.g., CTRPP Elision, Blending Phonemes-Words) and rapid naming (e.g., RAN/RAS). Moreover, the time period after which the retesting of the children in the present study took place was longer than that typically reported in most test manuals that report test–retest reliability (e.g., German, 1989; Markwardt, 1989; Wilkinson, 1993), a factor known to reduce stability estimates on most tests. On the other hand, this test–retest time interval would be considered common for repeated testing for research or educational purposes, particularly for children in the developing stages of reading, whose progress needs to be closely monitored. Similar stability values were evidenced on the nationally normed and experimental measures as a group, providing evidence for the psychometric soundness of the experimental tests and suggesting that they can be employed effectively in intervention studies that deal with improvement over time.

Some test–retest reliabilities were below expected levels based on general practice, which is of some concern although not entirely surprising. Such attenuation of the test–retest reliability coefficients appears to be due to several interrelated factors, including positive distribution skewness, kurtosis, floor effects on a number of the cognitive measures used in the present study, and restriction of participants' range in age and performance. In this study, the individual variables that produced weak test–retest reliability coefficients (e.g., error scores on the RAN/RAS, latency scores for the RAN Objects subtest, CTRRPP Word Reading Efficiency standard score, Initial List subtest of the *Spelling Transfer Test,* and the *Challenge Test*) yielded the most nonnormal distributions at both time points. The somewhat lower test–retest reliability values obtained on these measures were mitigated by good *SE* values (typically less than 3 raw score points for raw scores, and slightly higher for standard scores), which were generally consistent with those reported for the nationally standardized measures, suggesting that the degree to which these measures tap the underlying construct of interest is high.

Another conclusion from this study was that the reading performances of students with RD do not change significantly over time, at least when they are not attending school; there is no indication that practice effects operated as a whole. There also is no evidence for differential change over time be- tween nationally normed measures of decoding and spelling as a group and the experimental measures of decoding and spelling as a group or the experimental measures of reading- related cognitive processing skills as a group. These results indicate that reading and reading-related performance do not improve based solely on the passage of time, or with repeated exposure to reading lists or related stimuli. Therefore, if a significant and substantively meaningful change were to occur over time on these measures, this change could likely be attributed to the effectiveness of the instruction received by the children.

For several measures on which participants showed a statistically significant increase in raw score points, they simultaneously showed a statistically significant decrease in standard score units for these same measures. This may occur because older children typically outperform younger ones on measures of academic achievement and because normative samples are based on separate groups of individuals at different ages (or grades) collected simultaneously. If a child or group of children who are followed longitudinally fail to improve their raw score performance at a rate consistent with the difference in performance of two independent cohorts who differ in age (or grade), their standard score will decline.

The last major result of this study concerns the relation of the nationally normed measures of decoding and spelling to the experimental measures of decoding and spelling and the relation of all measures of decoding and spelling to the measures of cognitive processes presumed to be related to reading. Correlations between nation- ally normed decoding measures and experimental decoding measures were robust but not isomorphic, and correlations between nationally normed spelling measures and experimental spelling measures revealed more modest relations. Both the nationally normed and the experimental measures of decoding and spelling evidenced broadly similar correlations with the measure of cognitive processing skills related to reading; however, stronger correlations were found with phonological processes than with naming speed processes. There was, however, some variability, in that rapid naming of letters and numbers evidenced stronger correlations with reading and spelling measures than did rapid object naming. Furthermore, phonological correlations may have been higher in general, given this study's emphasis on decoding and spelling rather than on comprehension, which may involve rapid naming and retrieval to a greater extent. Overall, the overlap between the nationally normed and experimental measures of decoding and spelling in this study and the overlap of all the decoding and spelling measures with reading-related component cognitive processing skills suggest some evidence for their criterion validity.

On the other hand, experimental measures of decoding and spelling and related cognitive skills may possess unique variance in their relation with variables important in the reading process. For example, whereas phono- logical awareness is commonly identified as important for reading, other measures (e.g., keyword acquisition, Frijters et al., 2002; rapid naming skills, Wolf et al., 2002) add to the ability to predict reading gains and performance over and above phonological skills. As such, these and other measures de- scribed in this study may be useful tools for measuring the components critical for change early in the course of reading development, particularly for children experiencing difficulty at these stages.

For children with difficulties in learning to read, several factors (e.g., young age, floor effects, and the necessity to assess change over long periods of time) increase the chances of obtaining weak reliabilities, large practice effects, or nonsignificant relations between different types of measures. The practical effect of obtaining such poor psychometric properties would be to mitigate their validity and utility, particularly when attempting to deter- mine treatment effectiveness. Despite the obstacles that were present (and their consequences), good to excellent reliability estimates were obtained for nearly all measures of reading and reading-related cognitive processing used in this study, whether nationally normed or experimental. Furthermore, low measurement error, minimal practice effects, and significant correlations among different types of measures were obtained. Overall, the results of this study provide strong evidence to suggest that changes observed on these measures have utility in documenting the effectiveness of various intervention techniques.

ABOUT THE AUTHORS

Paul T. Cirino, *PhD, is a developmental neuropsychologist, associate director of the Regents Center for Learning Disorders, and an adjunct instructor at Georgia State University.* ***Fontina L. Rashid,*** *PhD, is a clinical psychologist and an assistant professor in the Department of Psychology at the University of the Pacific.* ***Rose A. Sevcik,*** *PhD, is a developmental psychologist and associate professor in the Department of Psychology at Georgia State University.* ***Maureen W. Lovett,*** *PhD, 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, University of Toronto.* ***Jan C. Frijters,*** *MA, is completing his PhD in applied developmental psychology at the University of Guelph, Ontario, Canada; he also consults for the Learning Disability Research Program at the Hospital for Sick Children in Toronto.* ***Maryanne Wolf,*** *EdD, 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, Harvard Medical School.* ***Robin D. Morris,*** *PhD, is a developmental neuropsychologist and Regents Professor of Psychology and also holds appointments in the Department of Educational Psychology and Special Education. He is the associate dean of research and graduate studies in the College of Arts and Sciences at Georgia State University. Address: Paul T. Cirino, Texas Institute for Measurement, Evaluation, and Statistics, Dept. of Psychology, 126 Heyne, Suite 204, Houston, TX 77204-5022.*

AUTHORS' NOTES

1. *Some of the experimental measures described in this study (*Word Reading *Efficiency, Elision, Blending) are now available on a nationally normed basis as the* Test of Word Reading Efficiency *(TOWRE) and the* Comprehensive Test of Phonological Processing *(CTOPP), respectively. Information regarding other experimental measures (*Sound Symbol Test, Strategy Test, Spelling Transfer Test, Timed Word Reading Tests, *and* Challenge Test*) is available from the fourth author (M.W.L.; e-mail: ldrp@sickkids.ca).*
2. *Support for this research was provided by the National Institute of Child Health and Human Development grant HD30970 to R.D.M., M.W., and M.W.L. at Georgia State University, Tufts University, and Hospital for Sick Children/University of Toronto, respectively, and from a Research Program Enhancement Grant from Georgia State University.*
3. *We thank the families of our participants and their schools for their cooperation and support.*

REFERENCES

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

Costenbader, V. K., & Adams, J. W. (1991). A review of the psychometric and administrative features of the PIAT-R: Implications for the practitioner. *Journal of School Psychology, 29,* 219–228.

Dean, R. S. (1979). The use of the PIAT with emotionally disturbed children. *Journal of Learning Disabilities, 12*, 629–631.

Denckla, M. B., & Rudel, R. G. (1974). Rapid automatized naming of pictured objects, colors, letters and numbers by normal children. *Cortex, 10,* 186–202.

Denckla, M. B., & Rudel, R. G. (1976). Rapid automatized naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14,* 471–479.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M. (1997). The case for early reading intervention. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 243–264). Mahwah, NJ: Erlbaum.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., et al. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86,* 6–23.

Frijters, J., Steinbach, M., De Palma, M., Temple, M., Lovett, M., Wolf, M., et al. (2002, February). *Word identification speed and learning transfer of reading disabled children.* Paper presented at the annual meeting of the International Neuropsychological Society, Toronto, ON, Canada.

Gaskins, I. W., Downer, M. A., & Gaskins, R. W. (1986). *Introduction to the Benchmark School word identification/vocabulary development program.* Media, PA: Benchmark School.

German, D. J. (1989). *Test of word finding: Technical manual.* Austin, TX: PRO-ED. Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman brief intelligence test.* Circle Pines, MN: American Guidance Service.

Lombardino, L. J., Morris, D., Mercado, L., DeFillipo, F., Sarisky, C., & Montgomery, A. (1999). The Early Reading Screening Instrument: A method for identifying kindergarteners at risk for learning to read. *International Journal of Language and Communication Disorders, 34,* 135–150.

Lovett, M. W., Borden, S. L., DeLuca, T., Lacerenza, L., Benson, N. J., & Brackstone, D. (1994). Treating the core deficits of developmental dyslexia: Evidence of transfer of learning after phonologically- and strategy-based reading programs. *Developmental Psychology, 30,* 805–822.

Lovett, M. W., Lacerenza, L., Borden, S. L., Frijters, J. C., Steinbach, K. A., & De Palma, M. (2000). Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes. *Journal of Educational Psychology, 92,* 263–283.

Lundberg, J., Frost, J., & Peterson, O. Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23,* 263–284.

Markwardt, F. C. (1989). *Peabody individual achievement test–Revised manual.* Circle Pines, MN: American Guidance Service.

McCullough, B. C., & Zaremba, B. A. (1979). Standardized achievement test used with learning disabled and non–learning disabled adolescent boys. *Learning Disability Quarterly, 2,* 65–70.

Naglieri, J. A., & Pfeiffer, S. I. (1983). Stability and concurrent validity of the Peabody Individual Achievement Test. *Psychological Reports, 52,* 672–674.

Olson, R., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyon (Ed.), *Frames of refer-*

ence for the assessment of learning disabilities (pp. 229–242). Baltimore: Brookes.

Olson, R., Wise, B., Conners, F., Rack, J., & Fulkner, D. (1989). Specific deficits in component reading and language skills: Genetic and environmental influences. *Journal of Learning Disabilities, 22,* 339–348.

O'Shaughnessy, T. E., & Swanson, H. L. (2000). A comparison of two reading interventions for children with reading disabilities. *Journal of Learning Disabilities, 33,* 257–277.

SAS Institute. (1988). *SAS procedures guide, release 6.03.* Cary, NC: Author.

Shull-Shenn, S., Weatherly, M., Morgan, S. K., & Bradley-Johnson, S. (1995). Stability reliability for elementary-age students on the Woodcock-Johnson Psychoeducational Battery–Revised (Achievement section) and the Kaufman Test of Educational Achievement. *Psychology in the Schools, 32,* 86–92.

Smith, M. D., & Rogers, C. M. (1978). Reliability of standardized assessment instruments when used with learning disabled children. *Learning Disability Quarterly, 1,* 23–31.

Stuart, M. (1999). Getting ready for reading: Early phoneme awareness and phonics teaching improves reading and spelling in inner-city second language learners. *British Journal of Educational Psychology, 69,* 587–605.

Torgesen, J. K., & Davis, C. (1996). Individual difference variables that predict response to training in phonological awareness. *Journal of Experimental Child Psychology, 63,* 1–21.

Torgesen, J. K., & Wagner, R. K. (1996). *The comprehensive test of reading related phonological processes.* Tallahassee, FL: Author.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of word reading efficiency (TOWRE): Examiner's manual.* Austin, TX: PRO-ED.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing (CTOPP): Examiner's manual.* Austin, TX: PRO-ED.

Wilkinson, G. S. (1993). *The wide range achievement test: Administration manual* (3rd ed.). Wilmington, DE: Jastak.

Wolf, M., Bally, H., & Morris, R. (1986). Automaticity, retrieval processes, and reading: A longitudinal study in average and impaired readers. *Child Development, 57,* 988–1000.

Wolf, M., Goldberg O'Rourke, A., Gidney, C., Lovett, M., Cirino, P., et al. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Reading and Writing, 15,* 43–72.

Wolf, M., Miller, L., & Donnelly, K. (2000). Retrieval, automaticity, vocabulary, elaboration, orthography (RAVE-O): A comprehensive, fluency-based reading intervention program. *Journal of Learning Disabilities, 33,* 375–386.

Woodcock, R. W. (1987). *Woodcock reading mastery tests–Revised: Examiner's manual.* Circle Pines, MN: American Guidance Service.