Reading Progress Monitoring for Secondary-School Students: Reliability, Validity, and Sensitivity to Growth of Reading-Aloud and Maze-Selection Measures

Renata Tichá, Christine A. Espin, and Miya Miura Wayman

University of Minnesota

The validity and reliability of curriculum-based measures in reading as indicators of performance and progress for secondary-school students were examined. Thirty-five grade 8 students completed reading aloud and maze-selection measures weekly for 10 weeks. Criterion measures were the state standards test in reading and the Woodcock-Johnson III Test of Achievement. Different time frames for each measure were compared. Most alternate-form reliability coefficients were above .80. Criterion-related validity coefficients ranged from .77 to .89. No differences related to time were found. Only maze selection reflected significant growth, with an average increase of 1.29 correct choices per week. Maze growth was related to the reading performance level and to change on the Woodcock-Johnson III from pre- to posttest.

No Child Left Behind (NCLB) requires that states set high standards in academic areas such as language arts (including reading), math, and science and assess and account for yearly progress of all students (U.S. Department of Education, 2008). In the reading/language arts area, states must have assessments in grades three through eight and in one grade of high school. The state assessments are high stakes not only for schools but also for teachers and students. When schools do not meet annual targets, they can be labeled and penalized with sanctions. When students do not achieve target scores, they can be denied the chance to graduate or to be promoted to the next grade (Warren & Edwards, 2005).

One way for teachers to help struggling students meet state standards is to monitor student progress and use the data to evaluate the effectiveness of students' instructional programs. A tool uniquely fitted to this purpose is Curriculum-Based Measurement (CBM; Deno, 1985). CBM is a systematic set of procedures for monitoring student progress and making instructional decisions (Deno, 1989). When using CBM, teachers sample student work on a frequent basis and graph student scores. If students progress at an adequate rate, the instruction remains the same. If students do not progress at an adequate rate, the instruction is changed. Via this recursive process of instruction, monitoring, evaluation, and modification, teachers systematically build effective and powerful instructional programs (Stecker, Fuchs, & Fuchs, 2005).

The need for careful and systematic design of instruction is essential for students with learning disabilities (LD), whose progress can be slow and incremental without implementation of powerful and effective interventions (Deno, Fuchs, Marston, & Shie, 2001; Fuchs, Fuchs, Hamlett, Walz, & Gerrmann, 1993). The need for careful and systematic design of instruction is imperative for secondary-school students with LD, who run the risk of failing to graduate if they are unable to pass state tests.

Over 30 years of research has supported the technical adequacy of CBM measures in reading (Marston, 1989; Madelaine & Wheldall, 2004; Wayman, Wallace, Wiley, Ticha, & Espin, 2007), and the theoretical and conceptual bases underlying the relationships between CBM and criterion measures (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Recently, interest in progress monitoring systems such as CBM has grown with the introduction of a response to intervention (RTI) approach for identifying students for special education services (see Bradley, Danielson, & Halihan, 2002). An integral part of RTI is measuring student progress within a given instructional setting and using the data to determine whether that setting—or tier of instruction—is effective for the student.

Despite the large research base, and recent interest in CBM, there remain large gaps in research and practice. One of these gaps is the development and use of the measures for secondary-school students, especially in the area of reading (see Wayman et al., 2007). The limited research that has been done at the secondary-school level has focused primarily on reading in the content areas rather than on general reading proficiency (see Espin & Tindal, 1998); however, with the passing of NCLB, there has been renewed interest in the reading proficiency of secondary-school students and a corresponding interest in the development of reading progress measures for these students.

Two measures have been examined in CBM reading research at the secondary-school level: reading aloud and maze selection. For the reading aloud measure, students read aloud from text for 1 minute, and the number of words read correctly is scored and graphed (see Deno, 1985). For...
maze selection, students read silently from a text in which every seventh word has been deleted and replaced with a multiple-choice item. The number of correct choices made in 1-3 minutes is scored and graphed (see Deno, Maruyama, Espin, & Cohen, 1989; Espin, Deno, Maruyama, & Cohen, 1989; Fuchs & Fuchs, 1992). Maze selection presents some advantages over reading aloud. It can be administered in groups and via the computer, and it is viewed by teachers as a reflecting decoding, fluency, and comprehension (Fuchs & Fuchs, 1992).

Muyskens and Marston (2006) examined the technical adequacy of a 1-minute reading-aloud measure in a study with 209 eighth-grade students. Three reading passages at approximately a ninth-grade reading level created from newspaper articles were used in the study. The relationship between the number of words read correctly and scores on the Minnesota state standards test in reading was examined. Results revealed a correlation between the CBM and criterion measure of .70.

In a larger study, Silberglitt, Burns, Madyun, and Lail (2006) examined the relationship between reading-aloud and maze-selection measures and performance on Minnesota state reading tests. Silberglitt et al. aggregated data across a 7-year period. Measures included reading aloud, administered to 582 seventh-grade and 843 eighth-grade students, and maze selection, administered to 282 seventh-grade and 1028 eighth-grade students. The Minnesota Basic Skills Test for Reading (MBST) was used as the criterion measure for the eighth-grade students; the Minnesota Comprehensive Assessment for Reading (MCA) was used as the criterion measure for both seventh- and eighth-grade students. Results revealed correlations for reading aloud and maze selection of .60 and .54, respectively, for seventh-grade students and .50 and .48, respectively, for eighth-grade students.

The Muyskens and Marston (2006) and Silberglitt et al. (2006) studies both focused on the technical adequacy of CBM measures as performance measures. Two recent studies have examined the characteristics of CBM reading measures as progress measures. MacMillan (2006) examined changes on a 1-minute reading-aloud measure from fall to winter to spring for 1,691 students in grades two through seven. Results revealed that, although the reading-aloud measure reflected growth at every grade level, the amount of growth decreased as the grade level increased. Students in grade seven grew on average only 7 words across the year, compared to students in grade two who grew 54 words, and students in grade four who grew 16.50 words across the year.

Espin, Wallace, Lemke, Campbell, and Long (2009) compared the technical adequacy of reading-aloud and maze-selection measures as indicators of both performance and progress. They also examined the effects of different scoring systems and time frames for each measure based on the hypothesis that, for older students, longer time frames or more complex scoring procedures might be needed to produce reliable and valid samples of work. Participants in the performance study were 238 eighth-grade students in the classrooms of 17 English teachers. Participants in the subsequent progress study were 32 students from the original sample, who were monitored over a period of 10 weeks with both reading aloud and maze selection. Reading-aloud measures were scored for total words read and words read correctly in 1, 2, and 3 minutes of reading. Maze-selection measures were scored for both correct and correct minus incorrect word selections in 2, 3, and 4 minutes.

Results of the performance study revealed that alternate-form reliability coefficients were above .80 for both reading aloud and maze selection and were consistently stronger for reading aloud (in the 90s) than those for maze selection (in the 80s). Reliabilities were similar across scoring procedures, but not time frames. Alternate-form reliability for maze selection increased with time (e.g., .80 for 2 minutes to .88 for 4 minutes). With respect to validity, correlations with performance on the state standards' reading test ranged from .76 to .79 for reading aloud and from .75 to .81 for maze selection. Again, coefficients were similar across scoring procedures, but not time frame for maze selection. Validity coefficients for maze selection increased slightly with time (e.g., .75 for 2 minutes to .80 for 4 minutes).

Results of the progress study revealed differences between the reading aloud and maze selection as growth measures. Reading aloud produced minimal (.84 words correct per week for 1 minute) or no growth (for 2 and 3 minutes), and growth was not related to performance on the state standards test. Maze selection, in contrast, produced growth rates of 2.37 and 2.88 selections per week for 2- and 3-minute samples, and these growth rates were significantly related to performance on the state standards test.

Espin et al. (2009) highlighted three limitations of their study. First, the criterion variable used in the study was a state standards test in reading, a test that reported limited validity and reliability data, and a test that was unique to the state in which the study took place. Second, the weekly reading-aloud measures were always administered prior to the administration of the weekly maze-selection measure. It was possible that the initial exposure to the reading-aloud measure decreased the sensitivity of the maze-selection measure to change over time because the two measures were created from the same passages. Finally, the validity of the growth rates produced by reading aloud and maze selection was examined using a static measure, the state standards test, administered at the end of the study. A stronger test of the validity of the measures would be to examine the relationship between change on the CBM measures and change on a criterion measure.

The purpose of the present study was to replicate the Espin et al. (2009) study with special attention to the limitations outlined above. First, the current study included two criterion variables: a state standard reading test and a standardized reading measure. The standardized reading measure was the Woodcock-Johnson III (WJ-III) Broad Reading Clustering, a test with known technical adequacy and one that is used across states. Second, we reversed the order of passage administration and administered the reading-aloud measure first and maze selection second. Finally, we examined the relationship between change on the CBM measures and change on the WJ-III Broad Reading Clustering from winter to spring. As in the Espin et al. (2009) study, we compared different time frames for each measure to determine whether it was necessary for maze selection to be longer to be reliable. Given...
the lack of differences in scoring procedures found in the earlier study, we used only the most efficient scoring procedures for each measure: words read correctly for reading aloud and correct choices for maze selection. Four research questions were addressed in the study:

1. What are the alternate-form reliabilities of reading-aloud and maze-selection measures? Do reliabilities differ by time frame?
2. What are the validities of reading-aloud and maze-selection measures as indicators of performance? Do validities differ by time frame?
3. Are reading-aloud and maze-selection measures sensitive to growth over time?
4. Are the growth rates produced by reading aloud and maze selection valid with respect to the change in performance on the WJ-III for lower and higher performing students?

In designing the study, we were faced with a choice of what type of a sample to include in our research. Our primary interest was the development of CBM measures for students with LD. However, if we were to include only eighth-grade students with LD, we would most likely obtain a skewed and truncated distribution of scores on both the CBM and criterion measures, adversely affecting the correlations calculated for Questions 1 and 2. One solution would be to include LD students across several grade levels (e.g., grades four through eight; see an example by Fuchs, Fuchs, & Maxwell, 1988), but this approach would not allow us to focus on the characteristics of the measures for secondary-school students. An alternative solution, and the one we chose, was to include a range of performance levels within a single grade (e.g., grade eight). This solution allowed us to focus on the characteristics of the measures specifically for secondary-school students. It also allowed us to break the groups into relatively lower and higher performing groups to compare the relationship between growth on the CBM and WJ-III (Question 4).

The approach we selected for determining our sample is not unique. In their review of the CBM reading literature, Wayman et al. (2007) reported on 29 technical adequacy studies conducted at the elementary-school level. Only 1 of the 29 used an exclusively special education sample. The remainder used samples of students in general education (13) or mixed samples of general and special education students (15).

**METHOD**

**Participants and Setting**

The study took place in an urban K-8 school in Minnesota with 740 students. Thirty-five percent of the students in the school were African American, 31 percent Asian American, 28 percent White, 3 percent Hispanic, and 3 percent Native American. Thirteen percent of the students in the school received special education services, and 74 percent qualified for free or reduced-price lunch.

Participants were 35 (15 male and 20 female) eighth-grade students. Forty-nine percent were White, 46 percent African American, 3 percent Asian, and 3 percent Hispanic. All students but one spoke English as their native language. One student spoke Hmong as a native language but, according to the school assessment, was proficient in English. Fourteen of the participants qualified for free or reduced-price lunch. Participants for the study were recruited from the classrooms of four eighth-grade teachers. At the beginning of the school year, the eighth-grade teachers grouped their students into four reading groups based on state test scores in reading and teacher judgment. The number of participants, and approximate reading level reported by the teachers for each group, were: Group 1 (n = 5), 2–3 years below grade level; Group 2 (n = 4), 1 year below grade level; Group 3 (n = 8), at grade level to 1 year above grade level; and Group 4 (n = 13), 1–2 years above grade level. In addition to the students in these four groups, five participants receiving services in special education—four for specific LD and one for emotional and behavioral disorders (EBD)—were included in the study.

**Measures**

**Predictor Variables**

The predictor variables in the study were scores on the reading-aloud and maze-selection measures. Differences in reliability and validity for various time frames were examined for both reading aloud (1, 2, and 3 minutes) and maze selection (2, 3, and 4 minutes). Scores for reading aloud were the number of words read correctly. Scores for the maze selection were the number of correct maze selections.

Reading-aloud and maze-selection passages were selected from the Espin et al. (2009) study. The passages were created from articles in the local newspaper. The following process was used to select the passages. First, passages were screened to eliminate those that might require specific background knowledge. For example, a passage about baseball was not included because students’ background knowledge about baseball might vary widely. Second, passages of equivalent difficulty were selected to the extent possible. All passages were examined for their readability level using the Degrees of Reading Power (Touchstone Applied Science and Associates, 2006) and Flesch–Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975). The Degrees of Reading Power formula is based on average word length (characters per word), average number of familiar words (proportion of words from the Dale–Chall list of 3,000 simple words), and average sentence length (words per sentence). The Flesch–Kincaid is based on the average number of syllables per word and average sentence length (words per sentence). The Flesch–Kincaid Grade Level was calculated via Microsoft Word. Passages that were most similar in readability were considered for inclusion. A final corpus of 10 passages was selected for the research. The passages were on average 750 words long. Readability, as measured by the Degrees of Reading Power scale, ranged from 31 to 61, representing approximately a sixth-grade level. For the Flesch–Kincaid, the readability level was between the fifth- and eighth-grade levels.
Maze-selection passages were created from the reading-aloud passages using the following procedures described by Fuchs and Fuchs (1992). The first sentence of the passage was left intact, after which every seventh word was deleted and replaced with three choices. Only one choice was semantically correct. Distracters were both auditorily and graphically different from the correct choice, but were approximately the same length as the correct choice. Students read silently through each maze-selection passage for 4 minutes. At 2 and 3 minutes, the administrator directed students to place a slash mark after the word they had just read and monitored students to ensure they followed the directions. For reading-aloud measures, numbered and unnumbered copies of each passage were created. The administrator used the numbered copy to score the passage during reading. Students read aloud for 3 minutes from the passage, and the administrator marked progress at 1, 2, and 3 minutes.

**Criterion Variables**

Criterion variables in the study were the WJ-III Tests of Achievement, Broad Reading Cluster (Woodcock, McGrew, & Mather, 2001) and the Minnesota Basic Skills Test (MBST) in reading. The WJ-III Broad Reading Cluster consisted of three subtests: Letter-Word Identification, Reading Fluency, and Passage Comprehension. In the Letter-Word Identification subtest, students read individual words that gradually increased in difficulty. In the Reading Fluency subtest, students read sentences silently and marked whether the sentence was true or false by circling YES or NO in their booklet. The sentences became increasingly more complex. In the Passage Comprehension subtest, students filled missing key words into short passages they read silently.

The WJ-III was normed on 8,818 subjects from over 100 U.S. communities from northeast, midwest, south, and west regions. The stratified sample was representative of urban and rural communities of different races and from various schools. The authors reported that the WJ-III was designed to minimize test bias due to gender, race, or origin. Internal reliability coefficients for the standard battery ranged from .81 to .94. Median test–retest reliabilities were .91 for the Letter–Word Identification subtest, .90 for the Reading Fluency subtest, and .83 for the Passage Comprehension subtest for students aged 5–19. Overall test–retest reliability was reported as .76 to .94 for students aged 7–11 and .80 to .89 for students aged 14–17 (Mather & Woodcock, 2001). The Broad Reading Cluster score used in our study correlated .60 to .70 with other reading measures, including the Kaufman Test of Educational Achievement (KTEA) and the Wechsler Individual Achievement Test (WIAT) (McGrew & Woodcock, 2001). Standard scores for the WJ-III were used for analysis. Because we were interested in individual changes in scores on the measure (rather than changes in relative rankings), and because the measure was given twice over a relatively short period of time, we calculated the standard scores on both pre- and post-test using the winter conversion tables. In this way, we could calculate the extent to which student scores changed from the first to the second testing administration.

The MBST (Minnesota Department of Education, 2004; Minnesota Department of Children, Families, and Learning, and NCS Pearson, 2001–2002) in reading was a high-stakes assessment that students needed to pass in order to graduate from high school and was aligned with Minnesota content standards. The first statewide mandatory administration of the test was in 1998. The MBST was a paper-and-pencil test that was not timed. At the time of the study, students in the state took the test in eighth grade. Students could retake the test each year until they passed. The MBST was composed of one narrative and three expository nonfiction articles selected from newspapers. Each article was at least 500 words long. Students were required to answer 40 multiple-choice questions based on the passages. Questions were both literal (approximately 65 percent) and inferential (approximately 35 percent). Students had to receive a percentage score of 75 to pass the test. Percentage scores were used for the analysis.

**Procedure**

**Data Collection**

Data were collected across a period of 10 weeks. The MBST was given at the beginning of the study. The WJ-III Broad Reading Cluster was individually administered to students at the beginning and end of the study by two graduate students trained in test administration for students with disabilities. Following the administration of the WJ-III pretest, students completed one reading-aloud and one maze-selection passage each week for 10 weeks. Students completed the maze-selection task prior to the reading-aloud task each week. Maze selection was group administered by the classroom teachers every Thursday. The classroom teachers were trained in maze administration procedures prior to the beginning of the study. The reading-aloud measure was individually administered by graduate students on the Monday or Wednesday following the maze. The graduate students administering the reading-aloud and maze measure were trained in CBM data collection procedures.

**Scoring**

Reading aloud was scored by the graduate students collecting the data. Words read correctly in 1, 2, and 3 minutes were recorded. Maze selection was scored by four graduate students who were trained prior to the beginning of the study. The students demonstrated 90 percent or above scoring accuracy on practice maze passages prior to beginning scoring. The maze was scored by counting the number of correct choices in 2, 3, and 4 minutes. To control for guessing, if students made three consecutive incorrect choices, scoring was stopped and the number of correct choices were counted up to that point.

**Data Analysis**

To address Research Questions 1 and 2 regarding reliability and validity of the reading-aloud and maze-selection measures (Howell, 2002), Pearson–Product Moment correlations
were computed. To address Research Questions 3 and 4 regarding the sensitivity of the two CBM measures to growth in reading and the relationship between growth on the CBM measures and change on the WJ-III, hierarchical linear modeling was used (HLM; Bryk & Raudenbush, 1987, 2002).

RESULTS

Interscorer Agreement

Interscorer agreement was calculated for every 20th reading-aloud and maze-selection passage. Agreement was calculated by dividing agreements by the total responses. Average agreement between the two reading-aloud scorers was above 95 percent. Average agreement among the four maze-selection scorers was 97 percent.

Alternate-Form Reliability

Alternate-form reliabilities were calculated by examining correlations between scores on pairs of passages administered at pretest. Reliabilities were calculated separately for each time frame (see Table 1). Because a large number of correlations were generated for each set of analyses, a Bonferroni correction was used to establish the levels used to determine significance (Howell, 2002). The $p$ value for significance was $p < .006$ ($0.05$ divided by nine sets of correlations for each measure). Alternate-form reliabilities for both reading aloud and maze selection were all statistically significant and generally above .80. Reliability coefficients were consistently higher for reading aloud than those for maze selection, although even maze-selection reliabilities were typically above .80. No effects of time frame were seen for reading aloud; a small increase in average alternate-form reliability was seen for maze selection across time frame, but these differences were small ($r = .84$ to .88) and were not consistent across pairs of passages.

<table>
<thead>
<tr>
<th>Variable</th>
<th>2 minutes</th>
<th>3 minutes</th>
<th>4 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maze Selection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passages 1 and 2</td>
<td>.83</td>
<td>.83</td>
<td>.87</td>
</tr>
<tr>
<td>Passages 1 and 3</td>
<td>.79</td>
<td>.87</td>
<td>.89</td>
</tr>
<tr>
<td>Passages 2 and 3</td>
<td>.90</td>
<td>.91</td>
<td>.88</td>
</tr>
<tr>
<td>Average</td>
<td>.84</td>
<td>.87</td>
<td>.88</td>
</tr>
<tr>
<td>Reading Aloud</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passages 1 and 2</td>
<td>.96</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>Passages 1 and 3</td>
<td>.96</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>Passages 2 and 3</td>
<td>.97</td>
<td>.97</td>
<td>.96</td>
</tr>
<tr>
<td>Average</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
</tr>
</tbody>
</table>

Note: All correlations significant at the $p < .006$ level.

TABLE 1

Alternate-Form Reliability for Reading Aloud and Maze Selection

Validity

Means and standard deviations for the CBM and criterion measures are reported in Table 2. For each student, the median score of the three passages was used. The means reported in Table 2 represent the mean of the median scores. Prior to conducting the analyses, one outlier was removed from the dataset (see Figure 1 for scatterplot and outlier). Students read on average 139 words in 1 minute. Although their average rate slowed slightly between 2 and 3 minutes, distributions were fairly normal at each time frame. The skewness and kurtosis for the distribution of scores on the 3-minute CBM reading-aloud measure were $\gamma_1 = -.89$ and $\gamma_2 = -.15$. Students made an average of seven correct maze choices in 1 minute. This rate did not decrease with time, indicating that students did not slow down over time. Examination of the distributions at each time point showed fairly normal distributions. The skewness and kurtosis for maze selection in 4 minutes were $\gamma_1 = -.22$ and $\gamma_2 = -.73$. The mean percentage score on the MBST was 49 percent with a range from 6 percent to 98 percent. The mean standard score on the WJ-III was 101.40 with a range of 65-138.

Correlations between pretest scores on the CBM measures, the WJ-III pretest, and the MBST are reported in Table 3. Correlations reflect concurrent validity. Due to the fact that multiple correlations were calculated for each measure, a Bonferroni correction was used to determine significance levels. The $p$ value for significance was $p < .008$ ($0.05$ divided by six sets of correlations for each measure). All correlations were statistically significant. Correlations between reading aloud and MBST ranged from .77 to .78 and between MBST and the WJ-III from .87 to .89. Correlations between maze selection and MBST ranged from .80 to .85 and between maze selection and the WJ-III from .86 to .88. Although correlations increased slightly with time, these differences were minimal, and even the shortest times for reading aloud and maze selection (i.e., 1 and 2 minutes) yielded coefficients close to or above .80. Correlations for reading aloud and maze selection were similar for the WJ-III, but small differences were found for the MBST, with slightly

<table>
<thead>
<tr>
<th>Measures</th>
<th>$M$</th>
<th>(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maze Selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 minutes</td>
<td>13.90</td>
<td>(6.12)</td>
</tr>
<tr>
<td>3 minutes</td>
<td>21.94</td>
<td>(9.23)</td>
</tr>
<tr>
<td>4 minutes</td>
<td>29.57</td>
<td>(11.96)</td>
</tr>
<tr>
<td>Reading Aloud</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 minute</td>
<td>138.94</td>
<td>(62.04)</td>
</tr>
<tr>
<td>2 minutes</td>
<td>280.59</td>
<td>(82.37)</td>
</tr>
<tr>
<td>3 minutes</td>
<td>412.66</td>
<td>(120.71)</td>
</tr>
<tr>
<td>MBST reading</td>
<td>49.03</td>
<td>(27.34)</td>
</tr>
<tr>
<td>WJ-III broad reading</td>
<td>101.44</td>
<td>(17.14)</td>
</tr>
</tbody>
</table>

Note: CBM scores represent the mean of the individual median scores.
FIGURE 1 Scatterplot of Curriculum-Based Measurement reading aloud 3 minutes median score and Minnesota Basic Skills Test raw score revealing an outlier.

TABLE 3
Correlations Between Curriculum-Based Measures and Criterion Measures

<table>
<thead>
<tr>
<th>Variable</th>
<th>WJ-III Pretest</th>
<th>MBST Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maze Selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 minutes</td>
<td>.86</td>
<td>.80</td>
</tr>
<tr>
<td>3 minutes</td>
<td>.88</td>
<td>.82</td>
</tr>
<tr>
<td>4 minutes</td>
<td>.88</td>
<td>.85</td>
</tr>
<tr>
<td>Reading Aloud</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 minute</td>
<td>.87</td>
<td>.77</td>
</tr>
<tr>
<td>2 minutes</td>
<td>.88</td>
<td>.77</td>
</tr>
<tr>
<td>3 minutes</td>
<td>.89</td>
<td>.78</td>
</tr>
</tbody>
</table>

Note: All correlations significant at the p < .008 level.
WJ-III = Woodcock-Johnson III; MBST = Minnesota Basic Skills Test for Reading.

larger correlations obtained for maze selection than for reading aloud. For both CBM measures, correlations tended to be larger with the WJ-III than with the MBST.

Growth

Research Question 3 addressed the sensitivity of the measures to growth. In essence, this question addressed whether or not the mean growth rates for the two measures were statistically different from zero; that is to say, it addressed whether the measures reflected change across the 10 weeks of the study. The assumption of this analysis was that students as a group did improve across the 10 weeks of the study; thus, if the measures were valid indicators of growth, they should reflect this improvement.

Sensitivity to growth was examined using HLM procedures. Given that there were minimal differences in the slopes for the various sample durations, we focused on the sample durations examined in Espin et al. (2009) to allow for comparisons across studies: 1-minute reading aloud and 3-minute maze selection. Descriptive statistics for both measures across the 10 weeks are reported in Tables 4 and 5.

We examined whether the mean initial status and the mean growth rate of all the students were statistically different from zero using the following model of unconditional random effects. The subject-specific Level 1 model was

\[ Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}, \]

where \( Y_{ij} \) was the \( i \)th person's score at the \( j \)th time point, \( i = 1, \ldots, n \) and \( j = 1, \ldots, t_j \), \( x_{ij} \) was the linear time coding used to fit a linear trend to the \( i \)th person's data across time, \( e_{ij} \) was the \( i \)th person's residual at time \( j \), and \( \beta_0 \) and \( \beta_1 \) were the person-specific intercept and linear coefficient. The Level 2 group equations were

\[ \beta_0 = \gamma_{00} + u_0, \]

\[ \beta_1 = \gamma_{10} + u_1, \]
examined the relationship between growth on the maze selection to a more stringent test of validity. We conducted an analysis of variance (ANOVA) to determine if there were significant differences in the maze selection in terms of its sensitivity to growth. To control for the initial status of maze, reading, and growth on a criterion variable, the WJ-III, we entered the student reading performance level as a variable so that we could examine the nature of the relation for students who were relatively lower and higher performing. The results revealed that, whereas the reading aloud measure was not sensitive to growth, the maze selection was.

Results of the analyses are illustrated in Figures 2 and 3. The first model included all interactions between initial status on maze, growth on maze, reading group, and growth on WJ-III. The difference score on the WJ-III was used as a covariate. All nonsignificant terms were dropped from the model. We examined the final model using a model of conditional random effects. The final subject-specific Level 1 model was

\[ Y_{ij} = \beta_{0i} + \beta_{1i}X_{ij} + e_{ij}. \]

The final Level 2 group equations were

\[ \beta_{0i} = \gamma_{00} + \gamma_{01}(\text{group}_i) + u_{0i}, \]

\[ \beta_{1i} = \gamma_{10} + \gamma_{11}(\text{group}_i) + \gamma_{12}(\text{WJ difference}_i) + u_{1i}, \]

where \( \gamma_{0i} \) was the relationship between the reading group and the intercept controlling for WJ-III difference score, \( \gamma_{1i} \) was the relationship between the reading group and the slope controlling for WJ-III difference score, and \( \gamma_{12} \) was the relationship between WJ-III difference scores and the slope controlling for the reading group. The group and WJ difference represented the ith person’s reading group level and WJ-III difference score.

Results of the analyses are illustrated in Figures 2 and 3. Intercept, or initial, scores for the higher performing students were significantly different from those of lower performing students, confirming performance differences between the groups.
FIGURE 2 Estimated maze growth for the higher and lower performing students, controlling for change on the Woodcock-Johnson III.

FIGURE 3 Estimated maze growth for the higher and lower performing students based on more and less change on the Woodcock-Johnson III.
two groups (23.32 vs. 11.74 correct choices for the higher and lower performing groups, respectively; when controlling for differences in the WJ-III; \( \gamma_{0.1} = -11.58, t = -4.08, p < .001 \)). Results also revealed differences in growth rates between groups when controlling for differences in the WJ-III. Although students in both the higher and lower performing groups demonstrated improvements on the maze-selection measure, growth rates were significantly greater for the higher performing group (1.33 correct choices per week) than those for the lower performing group (.41 correct choices per week; \( \gamma = -.92, t = -3.90, p < .001 \); see Figure 2).

Our primary interest was on whether the growth rates would be related to the change in performance on the WJ-III for the higher and lower performing groups. Results revealed that growth on the maze selection measure was related to change on the WJ-III when controlling for group differences (\( \gamma_{1.2} = .05, t = 2.60, p = .01 \)), with students with larger WJ-III change scores growing more rapidly on the maze-selection measure than students with smaller WJ-III change scores. There were no differences in the relationship between growth on the maze-selection measure and change on the WJ-III for higher and lower performing students. We illustrate the results in Figure 3 by presenting the mean growth rates for students within each group divided into those with larger (at or above the 50th percentile) and smaller (below the 50th percentile) WJ-III change scores. From the graph, one can see that growth rates for students with larger WJ-III change scores are steeper than those with smaller WJ-III change scores and that this pattern is evident for both higher and lower performing students.

**DISCUSSION**

The purpose of this study was to examine the reliability and validity of reading aloud and maze-selection measures as indicators of performance and progress in reading. Specifically, we examined alternate-form reliability, concurrent and predictive validity, sensitivity to growth, and the validity of the growth rates for the two measures. Additionally, we investigated differences in patterns of results related to different time frames for the measures. Our study was a replication and extension of a study by Espin et al. (2009), which found that both reading aloud and maze selection had good reliability and validity as performance measures, but only maze selection produced substantive growth rates and produced growth rates related to performance on a criterion measure. In our study, we address three limitations of the earlier Espin et al. study: (1) we reversed the order in which reading aloud and maze-selection measures were administered; (2) we employed two criterion measures, one a standardized achievement test with known reliability and validity and one a state standards test; and (3) we examined the validity of the growth rates produced by CBM measures with respect to change on a criterion measure.

Our results replicated the results of Espin et al. (2009) in terms of both the performance and progress characteristics of the measures. With respect to performance, results of our study revealed that both reading aloud and maze selection had good reliability and validity as indicators of performance on the state standards and standardized achievement tests. The two CBM measures had relatively strong alternate-form reliability coefficients with most coefficients above .80. As with the Espin et al. (2009) study, alternate-form reliability coefficients were consistently higher for reading aloud (all above .95) than those for maze selection (all but one above .80). In addition, there were no effects of time frame on alternate-form reliability for the reading-aloud measure. For the maze-selection measure, the present study revealed small increases in average reliabilities with time frame, yet the increase was minimal (.84 to .88) and was not seen for every passage pair. In contrast, Espin et al. (2009) found a somewhat larger increase with time frame (.80 to .88).

The validity coefficients obtained in our study were surprisingly similar to those obtained in the earlier study (Espin et al., 2009) for the MBST. In our study, correlation coefficients between reading aloud and the MBST ranged from .77 to .78 and between maze selection and the MBST between .80 and .85. In Espin et al., correlation coefficients between reading aloud and the MBST ranged from .70 to .78 and between maze selection and the MBST between .79 and .80. Similarly, Muyskens and Marston (2006) found correlation of .70 between a 1-minute reading aloud sample and the MBST; however, Silberglitt et al. (2006) found correlations of only .54 and .48 between the reading aloud and maze selection and the MBST. It is not clear why the Silberglitt et al. study differed from the others. Note that each of these studies used the MBST as a criterion variable, a factor that perhaps contributes to the similarity in results between this study and the others. An important question was whether such results would generalize to a different criterion measure.

Results of the current study supported the generalizability of outcomes across criterion variables. Correlation coefficients between the reading-aloud measure and WJ-III ranged from .87 to .89 and between maze selection and the WJ-III ranged from .86 and .88. Correlations with the WJ-III were consistently stronger than those with the MBST reading. We hypothesize that these differences reflect the broad nature of the WJ-III test, which included letter and word reading, reading fluency, and reading comprehension. The MBST focused on reading comprehension only.

With respect to the progress characteristics of the measures, results of the present study were again similar to the Espin et al. (2009) study; both sets of results revealed that on repeated, weekly measurements, maze selection reflected substantial growth over time (2.88 and 1.29 for a 3-minute sample for the original and replication study, respectively), but reading aloud showed minimal or no growth (.84 and 0 because the growth rate was not significantly different from zero) for a 1-minute sample for the original and replication study, respectively. In addition, results of Espin et al. revealed a significant relationship between growth on the maze selection and performance on the state test, but no such relationship between growth on reading aloud and state test performance. In our study, we used the WJ-III as our criterion measure, and examined change on the measure as our outcome. Our results indicated that students who improved more on the WJ-III also grew more on the maze selection measure and that this relationship held true for students.
in both lower and higher performing groups. In addition, students in the higher performing group grew faster on the maze-selection task than students in the lower performing group, even after differences related to change on the WJ-III were controlled for.

We were surprised by the replication in growth results. It is difficult to imagine that reading aloud could produce good reliability and validity data as an indicator of performance, but not be sensitive to progress in reading over time. Espin et al. (2009) raised the possibility that such seemingly contradictory results might be due to the fact that students in their study completed the weekly reading-aloud task first and maze-selection task second; completion of the reading-aloud task might have first familiarized students with the content of the passages and thereby increased growth on the maze-selection measure. Although not a direct examination of order effects, in the current study, maze selection was administered prior to reading aloud, yet we found a similar pattern of results as Espin et al. with respect to sensitivity to growth for the two measures. Espin et al. also considered the fact that their results were particular to the sample of students included in their study; yet our study included different students from a different school (who were in the same school district and albeit both studies included students only in eighth grade).

It is possible that the results are unique to the specific materials used to develop the CBM measures because the same set of materials were used in both our study and the Espin et al. (2009) study. However, recall that the reading-aloud and maze-selection passages given each week were created from the exact same passages; thus, it is difficult to attribute the differences in growth rates between the two types of measures to differences in materials. Perhaps the most salient explanation for the differences in the characteristics of the two measures as growth measures is that offered by Espin et al. (2009). Each student reaches a natural level of individual reading fluency that does not change greatly over a short period of time (e.g., 10 weeks) but does serve to rank order students in terms of general reading skills.

Whatever the reason for the progress results, we must at this time caution secondary-school teachers about the use of reading aloud as a progress measure, at least over short periods of time. We would instead recommend consideration of a 2- or 3-minute maze-selection procedure scored for correct choices. We recommend, however, that the research be replicated with different materials and at different grade levels before concluding that reading aloud is not appropriate for monitoring progress of secondary-school students. Our study was conducted with a small sample of eighth-grade students. It is not clear whether these results can be generalized to students of other ages or to other sets of reading passages. We would also recommend that the administration order of the reading-aloud and maze-selection measures be counterbalanced to ensure that order does not play a role in the results. We were unable to counterbalance in our study due to logistical reasons, but a counterbalanced order would eliminate questions about order effects.

In sum, results across the Espin et al. and our replication study supported the reliability and validity of both reading-aloud and maze-selection measures as indicators of general reading performance. In terms of time frame, a 1-minute reading-aloud sample was as reliable and valid as 2- or 3-minute samples. For maze selection, some differences in alternate-form reliability were found related to the time frame, with increases in reliability seen with increases in the time frame; however, these differences were not consistent across passages within our study and were small in the current study compared to those found in the Espin et al. (2009) study. With respect to the measures as progress measures, results of both studies revealed that a 3-minute maze-selection measure was more sensitive to change over time than a 1-minute reading-aloud measure.

The current and previous research studies provide information on the technical adequacy of CBM reading measures for secondary-school students. However, what has not yet been addressed is the effect of measurement implementation on the performance and progress of students with LD. In other words, does a teacher's use of progress monitoring contribute to improved instruction and subsequent improved performance for students with LD? Determination of "treatment validity" is essential to the determination of the overall construct validity of the measures (see Messick, 1989a, b) and must be addressed in future research.

ACKNOWLEDGMENTS

We wish to thank Tharushi Nanayakkara for her help with data collection, Jeffrey Long and Seungsoo Yeo for their guidance and feedback on data analysis, and Stanley Deno for his insights on data interpretation. We also wish to thank the Netherlands Institute for Advanced Study in the Humanities and Social Sciences for its support in the preparation of this manuscript.

NOTES

1. HLM analysis was also conducted for 3-minute reading aloud in order to have a direct comparison with the results for 3-minute maze selection. The results for 3-minute reading aloud showed the same pattern as the results for 1-minute reading aloud.

2. HLM accounts for variance not explained by the slopes causing intercept estimates to differ slightly from the Week 1 raw means.

REFERENCES


About the Authors

Renáta Tichá, Ph.D., works as a research associate at the Institute on Community Integration, University of Minnesota. She received her Ph.D. and M.Ed. in Educational Psychology/Special Education at the University of Minnesota. Her research focuses on developing formative assessments for students with significant cognitive disabilities, and on studying ways to help teachers become more effective instructional decision makers.

Christine Espin, Ph.D., is a Universitair Hoofddocent in Special Education at Leiden University and an adjunct Professor in Cognitive Sciences at the University of Minnesota. Prior to moving to the Netherlands, she was a professor at the University of Minnesota in the department of Educational Psychology/Special Education. Her primary research interests are the development of progress monitoring procedures in reading, written expression, and content-area learning for secondary-school students with learning difficulties. She began her career as a special education teacher for high school students with learning disabilities.

Miyaka Miura Wayman, Ph.D., works as a research specialist at the Institute on Community Integration, University of Minnesota. She earned her Ph.D. and M.A. in Educational Psychology/Special Education from the University of Minnesota and her B.S. in Special Education and Mathematics Education from Winona State University. Her current research interests include developing progress monitoring measures for high-stakes assessments.