FURTHER EVIDENCE OF ORAL READING FLUENCY’S UTILITY IN PREDICTING STATEWIDE STUDENT READING PROFICIENCY

Paul Mooney¹, Kevin S. McCarter, & Jodie Schraven
Louisiana State University

John M. Hintze
University of Massachusetts at Amherst

Emily Mooney, Darlene Landry, & Terri Colclough
South Louisiana Primary School*

Kristin A. Gansle, R. Kenton Denny, & Gerlinde Grandstaff-Beckers
Louisiana State University

Gregory J. Benner
University of Washington at Tacoma

Jill H. Allor
Southern Methodist University

Abstract. Background, purpose. Authors replicated and extended research on the relationship between formative and summative reading assessment with a diverse student sample. Material and methods. Participants included all third-grade students (N = 84) in an urban K-5 elementary school in south Louisiana during the 2005-06 school year. Two types of measures were used to examine the relationship between progress monitoring and statewide assessments: (a) ORF benchmark probes from the DIBELS web-based system; and (b) the criterion-referenced reading portion of the third grade integrated Louisiana Educational Assessment Program (iLEAP)

¹ Address for correspondence: Paul Mooney, 219A Peabody Hall, Baton Rouge, LA 70803, (225) 578-2360, (225) 578-9135, pmooney@lsu.edu
Results and conclusions. Test data collected for 84 third grade students from an urban Louisiana elementary school demonstrated: (a) correlations between student scores on the reading portion of the integrated Louisiana Educational Assessment Program (iLEAP; Louisiana Department of Education, 2006b) and the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002) oral reading fluency (ORF) were moderate to strong across multiple ORF scores; (b) cut scores for identifying students at risk of failing the iLEAP met desired levels of diagnostic accuracy; and (c) ORF change score patterns were different from previously reported findings. Implications, limitations, and future research are discussed.

Key words: oral reading fluency, integrated Louisiana Educational Assessment Program, Dynamic Indicators of Basic Early Literacy Skills, student reading proficiency

School districts now must involve a majority of their students in statewide assessment systems. In the area of reading, for example, No Child Left Behind (NCLB; 2001) mandated that all students be tested annually in grades 3 through 8 and at least once between grades 9 and 12 inclusive. Test questions are often based on state-developed standards and designed to determine whether or not a student has demonstrated proficiency in grade level expectancies. In some states, including Louisiana, students must demonstrate mastery of grade level material in specified grades (e.g., fourth) in order to be promoted to the next grade [Louisiana Department of Education (LDE; 2006b)].

While well-developed statewide tests can provide educational stakeholders with information regarding a student population’s mastery of subject matter, they also present challenges. Among the hardships for educators are finding ways to meaningfully use such test data. With tests administered in the spring semester of a given academic year and results often reported at or near the end of the academic calendar, it becomes extremely difficult for school officials or anybody with a legitimate educational interest to respond to the results in a timely and efficacious manner. This problem has led some to advocate for supplementing annual accountability systems with formative assessment systems which allow for more frequent, efficient, and instructionally relevant data gathering efforts (Jiban & Deno, 2007). Crawford, Tindal, and Stieber (2001), for example, reasoned that ongoing monitoring was necessary in order to determine when instructional changes were warranted. Research ef-
forts to date support the assertion that combining annual accountability and ongoing progress monitoring systems has merit, with the present research building on this small but growing body of evidence.

Literature exploring the relationship between accountability and progress monitoring measurement systems in reading has been ongoing throughout the first decade of the 21st century. In discussing individual results as well as those findings within the broader burgeoning peer-reviewed literature context, researchers have drawn a number of conclusions. First, the magnitudes of the correlations between ORF measures and varied statewide testing systems have been both consistent and positive, ranging in strength from moderate to large. Correlations, predominantly between general outcome measures of fluency with connected text (i.e., oral reading fluency or ORF) and annual statewide tests, have ranged from 0.44 in Washington to 0.79 in Illinois (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006). Correlations reported in the peer-reviewed literature have been wide-ranging, as have been the types of analyses conducted. For example, some researchers (e.g., Good, Simmons, & Kame‘enui, 2001; McGlinchey & Hixson, 2004) have reported single spring semester correlations between statewide tests and ORF reading scores, whereas others (e.g., Crawford et al., 2001; Hintze & Silberglitt, 2005; Shapiro et al., 2006; Stage & Jacobsen, 2001) have reported multiple same-year correlations (e.g., fall; winter; spring) between statewide and ORF scores. Some researchers have focused on one grade (e.g., McGlinchey & Hixson; Stage & Jacobsen), whereas others (e.g., Shapiro et al.; Silberglitt, Burns, Madyun, & Lail, 2006) have targeted multiple grades. Longitudinal analyses of data (e.g., Good et al., 2001; Hintze & Silberglitt) have also been conducted.

Second, and related to the first conclusion, consistency in the correlations across states confirms the utility of CBM instruments such as the one-minute ORF measure. As Shapiro et al. (2006) note, “[c]onsidering that each state assessment measure is typically built to evaluate student progress toward competency on state curriculum standards, and that these standards vary considerably from state to state, CBM is indeed a very powerful measurement tool that appears to transcend the differences in state assessments” (p. 28). Third, considerably more research is warranted in the area of reading given the fact that state tests are often based on states’ own content standards and less than 20 per-
In the present research (e.g., McGlinchey & Hixson, 2004), the majority of data have been collected from populations that lack racial and ethnic diversity. Indicative of the research, demographic data reported in studies using Minnesota students included minority student data accounting for less than 10 percent of the sample (Hintze & Silbergliitt, 2005; Silbergliitt et al., 2006; Silbergliitt & Hintze, 2005). Demographic data from research in Oregon and Washington reported samples that were 10 percent minority or less (Crawford et al., 2001; Good et al., 2001). To date, McGlinchey and Hixson conducted the only published study reporting a diverse sample (i.e., a fourth grade sample that was 52 percent minority and 60 percent free or reduced lunch).

A fourth conclusion relates to the use of cut scores. Cut scores are essentially measures of predictive proficiency and can be used to make data-driven determinations as to whether or not a student is on his or her way to becoming a proficient reader (Silbergliitt & Hintze, 2005). Specifically, scores greater than or equal to those at a cut line could indicate proficient reading performance while those below that cut line could indicate some risk of reading difficulty. The architect of CBM, Stanley Deno (1985), originally proposed that local data be used to determine these cut scores, but Hasbrouck and Tindal (1992) argued that a larger and more diverse sample was warranted for determining these distinguishing scores. Hasbrouck and Tindal published their own cut scores in words correct per minute (WCPM) for end of first-, second-, and third-grade performance. Still other comparisons to data-driven benchmarks have been made available with access to the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; available on the web at http://dibels.uoregon.edu/benchmark.php) and commercially available systems such as AIMSweb. The DIBELS web site, for example, has readily available benchmarks for ORF across six elementary grades.

Current researchers have extended this available benchmark technology by connecting it with results from statewide accountability systems. That is, researchers have used CBM reading data to predict proficient or non-proficient performance on statewide tests. As Silbergliitt and Hintze (2005) report, findings across research studies indicate that “cut score values are remarkably consistent” (p. 320) across these studies. For example, Silbergliitt and Hintze reported end-of-year cut scores for Gra-
des 1, 2, and 3 at 49, 90, and 107 WCPM. Good et al. (2001) demonstrated the utility of those originally reported by Hasbrouck and Tindal (1992), which were 40, 90, and 110 WCPM for Grades 1-3. Ax and Bradley-Klug (2005) reported third-grade cut scores across nine states which ranged from 100 to 114 WCPM. Stage and Jacobsen (2001) established a fall fourth-grade cut score of 100 WCPM in Washington which was supported by data collected and analyzed by McGlinchey and Hixson (2004) in Michigan.

Similarly, Silberglitt and Hintze (2005) have conducted analyses which favor use of either logistic regression or receiver operating characteristic (ROC) curve analysis statistical techniques in determining cut scores. These analytical tools serve to increase the predictive power of ORF scores as indicators of proficiency or lack of proficiency on statewide tests. In practical terms, such use of statistical analyses to create cut scores provides school-based practitioners and anyone with a legitimate educational interest the chance to use statewide accountability data to inform instructional decision-making. That is precisely what Crawford et al. (2001) advocated for in identifying the shortfalls of informing instruction by relying solely on statewide accountability systems (Jiban & Deno, 2007). Research (e.g., McGlinchey & Hixson; Shapiro et al.; Silberglitt & Hintze, 2005; Stage & Jacobsen) examining the accuracy of pass or fail predictions on statewide reading tests based on calculated ORF reading cut scores is suggestive of ORF’s usefulness as a predictive screening tool. Shapiro and colleagues, for example, reported that the probability that students in two Pennsylvania school districts who failed the statewide test would be identified as not meeting the ORF cut score ranged from .83 to .94.

A final conclusion relates to patterns of student growth rates in reading. Silberglitt and Hintze (2007) examined the growth rates of a non-diverse sample of 6,642 elementary students in rural or suburban Minnesota using benchmark reading data. Findings demonstrated that the slope of ORF performance was significantly lower for students in the bottom and uppermost deciles of their fall benchmark distributions than those in the 50th-59th percentile. Moreover, Silberglitt and Hintze reported that the growth rate pattern “seemed to be more pronounced in the earlier grades (2 and 3) as compared to Grades 4 to 6” (p. 80).
Given the aforementioned conclusions, the goal of the present research was to replicate and extend findings reported in the peer-reviewed and published literature. The present research replicated the current evidence base by calculating correlations between statewide iLEAP scores and ORF measures, using these correlations and the area under the ROC curve to assess the accuracy of predicting iLEAP scores using the ORF scores, determining cut scores for each ORF measure, and examining ORF change scores. The present research extended the literature to a different statewide instrument (i.e., the Louisiana criterion-referenced reading test) as well as a more ethnically diverse population of students. Moreover, it focused specifically on third-grade students in one school for two reasons. One, third grade is the first grade in which NCLB has mandated statewide testing. Second, reporting of the analyses allowed the researchers to place the findings in a practical context in which benchmark data are used to determine who and who does not receive intervention services. This literature has yet to report on the types of instructional responses school personnel are making to benchmark data.

**METHOD**

**Participants and Setting**

Participants included all third-grade students \( (N = 84) \) in an urban K-5 elementary school in south Louisiana during the 2005-06 school year. In order to maintain as much confidentiality as possible, researchers agreed not to publish the names of either the school or the parish in which the school was located. Moreover, researchers were provided no student names, only an identification number. The present study’s sample characteristics were diverse. Specifically, 48 percent of the participants were minorities, all African American students. Additionally, 60 percent of the class received free or reduced cost lunches. Sixty-three percent of the sample was male. Ten percent of the sample had verified disabilities. There were no students for whom English was considered a second language. Fifty-seven percent of the sample scored at or above the basic level of proficiency on the statewide reading test. That compared to 42 percent of all Louisiana third graders and 43 percent of the parish’s third graders [Louisiana Department of Education (LDE, 2006b)].
Public school demographic data for the state and parish also indicated diversity. In Louisiana, 49 percent of the K-12 population of students was minority [Louisiana Department of Education, (LDE, 2006a)]. Additionally, 62 percent of Louisiana public school students during the 2005-06 academic year received free or reduced cost lunches, 13 percent were verified with disabilities, and 1 percent were limited English proficient (LEP). For the parish, which included 11 elementary or primary schools, 65 percent of the students were minorities, 45 percent received free or reduced cost lunches, 13 percent were verified with disabilities, and less than 1 percent were classified LEP (LDE, 2006a).

**Measures**

Two types of measures were used to examine the relationship between progress monitoring and statewide assessments: (a) ORF benchmark probes from the DIBELS (Good & Kaminski, 2002) web-based system; and (b) the criterion-referenced reading portion of the third grade integrated Louisiana Educational Assessment Program (iLEAP) [Louisiana Department of Education, (LDE, 2006c)].

**ORF.** DIBELS ORF is a standardized, individually administered test of accuracy and fluency with connected text (Good & Kaminski, 2002). Students read for one minute from a standardized reading passage written at a grade-appropriate level. Students are scored on WCPM. Technical adequacy research on CBM ORF measures has been favorable. Good and Kaminski, for example, reported test-retest reliabilities for elementary students above .9 and alternate-form reliabilities from .89 to .94. Additionally, studies of criterion-related validity have reported correlation coefficients between .52 and .91 (Good & Kaminski).

For the present study, progress monitoring assessments were conducted three times yearly in both the school and the parish, in September, January, and May of the academic year. Students were provided an ORF passage, which included a collection of sentences, and asked to read aloud from the top of the page for one minute. A school-based and district-trained team of DIBELS assessors administered the measure and conducted the scoring, counting the number of words read correctly in the minute that each student read. For each of the three progress monitoring assessment periods, students read from three ORF passages.
For each progress monitoring period, the student’s median score was reported for evaluation purposes. School officials then summarized the data by classes in order to identify at-risk students as well as gauge the effectiveness of the instructional program. ORF data were collected and analyzed for the following time periods: (a) Winter 2003 (January of students’ first-grade year); (b) Fall 2005 (September of students’ third-grade year); and (c) Winter 2006 (January of students’ third-grade year). For this particular grade, progress monitoring data were not collected at the school during the students’ second-grade year.

**iLEAP Grade 3 English/Language Arts Criterion-Referenced Test.** The iLEAP is Louisiana’s statewide assessment for students in grades 3, 5, 6, 7, and 9. The third-grade iLEAP measures students’ proficiency in reaching Louisiana academic standards in English/language arts (ELA), math, science, and social studies. According to the interpretive guide (LDE, 2006b), the iLEAP was developed in response to NCLB’s requirement for a state test that aligned with state content standards and included a scoring system that related directly to state performance standards (i.e., five achievement levels for Louisiana ranging from unsatisfactory to advanced). According to LDE (2006b), test items for the iLEAP came from two sources: (a) the Iowa Tests, a group-administered and norm-referenced achievement test battery for grades K-8; and (b) an LDE-administered process designed to create questions based on “certain content standards and grade level expectations” (LDE, 2006b, p. 1). According to LDE (2006b), for grade 3, items from the Iowa Tests were first matched to Louisiana grade level expectations. Where “gaps in coverage or instances of less than optimal alignment” (p. 1) occurred, then additional items were written to align to the Louisiana content standards, benchmarks, and grade level expectations (LDE, 2006b). Thus, the iLEAP ELA test consists of both a norm-referenced and a criterion-referenced portion. The norm-referenced test questions come from the Iowa Tests, whereas the criterion-referenced test includes Iowa Tests questions and those created to align with grade level expectations (LDE, 2006b). Data for the present study consist solely of criterion-referenced scores. Achievement levels and associated scaled scores were as follows: (a) unsatisfactory, 100-238; (b) approaching basic, 239-281; (c) basic, 282-337; (d) mastery, 338-382; and (e) advanced, 383-500. Students scoring at or above the basic level are considered to have passed the test. Students scoring at
the basic level are deemed to have demonstrated “only the fundamental knowledge and skills needed for the next level of schooling” (LDE 2006b, p. 3).

The criterion-referenced ELA test for third graders targets knowledge and skill development in reading and writing and consists of four subtests (i.e., writing, using information resources, reading vocabulary and comprehension, and language) which are administered over two days (LDE, 2006b). For the writing test, students are expected to write a narrative or descriptive composition on an assigned topic. Responses are scored from 1 to 4 points on each of two dimensions (i.e., composing, style/audience awareness). For the remaining three subtests, students are expected to answer multiple-choice questions. On the information resources subtest, students are provided resource materials from multiple sources such as newspapers or web pages and asked to locate specific information to answer questions. On the reading test, students answer questions in the areas of vocabulary and comprehension. In the vocabulary section, students are asked to read a word and then choose the word that has a similar meaning. In the comprehension section, students are required to read a statement and answer questions that follow the statement. On the language test, students are tested in areas pertaining to spelling, capitalization, punctuation, language usage, and language expression. All multiple-choice items are scored as correct (i.e., receiving 1 point) or incorrect (i.e., receiving 0 points) (LDE, 2006b). The iLEAP was administered in March 2006 to all Louisiana third-grade students. It was the first administration of the new iLEAP format in Louisiana.

Technical adequacy data for the iLEAP criterion-referenced, third-grade ELA test were provided by LDE staff. Reliability data were reported in terms of internal consistency. Two estimates of internal consistency (i.e., Cronbach’s alpha, Stratified alpha) resulted in high Pearson correlation coefficients of .93 for each method (LDE, 2006c). Salvia, Ysseldyke, and Bolt (2007) indicate that coefficients above .90 are recommended for tests in which educational decisions are made for individual students. Validity data were discussed in terms of content validity. The process of developing content validity was reported to have taken place in stages, including development of content standards for third-grade ELA by a committee consisting of educators, LDE curriculum and assessment staff, and an outside consultant, a test blueprint based on the standards,
and test items which were then reviewed for content alignment and field tested (LDE, 2006c). No validity coefficients were reported in the iLEAP technical summary.

**Procedures**

Each third-grade participant completed the iLEAP ELA test in March of the spring semester of the 2005-2006 academic year. Each participant was also administered ORF probes during the fall and winter school-wide benchmarking periods. Students who had either attended that school since first grade or who had winter first grade ORF measures available for analysis were also included. Demographic and assessment data were collected and entered into an Excel file by school personnel who removed student names and replaced them with student identification numbers prior to providing the database to the university research team.

**RESULTS**

Before conducting any statistical analyses, the data were examined for accuracy of data entry, missing values, outliers, and distributional properties. Data entry analyses by a member of the school-based assessment team yielded no changes to what had been entered. Four cases were missing third grade ORF fall scores; three cases were missing third grade ORF winter scores; 29 cases were missing winter first grade ORF data; no cases were missing iLEAP scores. Using a $p < .005$ criterion for Mahalanobis distance, no outliers among the cases were found. No gross departures from normality were detected.

**Testing Descriptives**

Table 1 provides descriptive test data. For all students, the average WCPM on the third grade fall ORF was 75.01 (95% confidence interval; CI [69.30, 80.72]), whereas for the winter ORF it was 82.89 (95% CI [76.66, 89.12]). The average increase from fall to winter ORFs is estimated to be 7.98 (95% CI [5.52, 10.45]). At each time point, there were no significant differences found in the average ORF scores across race, gender, or SES. For students entering third grade and assessed during the fall
benchmarking period, for example, $p$-values for race, gender, and SES were 0.3795, 0.2173, and 0.8284, respectively.

### Table 1. Descriptive Statistics for Reading Scores

<table>
<thead>
<tr>
<th>Measure and Testing Season</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF Fall Grade 3</td>
<td>80</td>
<td>75.0</td>
<td>25.7</td>
<td>25</td>
<td>133</td>
</tr>
<tr>
<td>ORF Winter Grade 3</td>
<td>81</td>
<td>82.9</td>
<td>28.2</td>
<td>31</td>
<td>151</td>
</tr>
<tr>
<td>ORF Winter Grade 1</td>
<td>55</td>
<td>26.2</td>
<td>18.1</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>iLEAP ELA-Reading Spring Grade 3</td>
<td>84</td>
<td>275.9</td>
<td>60.7</td>
<td>100</td>
<td>368</td>
</tr>
</tbody>
</table>

*Note:* ORF = oral reading fluency curriculum based measurement scores; iLEAP = integrated Louisiana Educational Assessment Program. ORF scores constitute the number of correct words read correctly per minute; iLEAP scores are standard scores. A passing score on the iLEAP is 282 and above.

The average student fell just short of demonstrating a basic level of proficiency on the iLEAP (i.e., a mean score of 275.9 versus a basic range of 282-337). Average iLEAP scores were not significantly different across race ($p = 0.2870$), gender ($p = 0.3519$), or SES ($p = 0.4385$).

At each time point, there were highly significant differences found in the average ORF scores between those who passed and those who failed the iLEAP. Students who passed the iLEAP averaged 16.5 WCPM better on the first grade winter ORF than did those who failed iLEAP ($p < 0.0001$). Similarly, those who passed the iLEAP averaged 32.8 points better on the third grade fall ORF than did those who failed iLEAP ($p < 0.0001$), and students who passed the iLEAP averaged 32.3 points better on the third grade winter ORF than did those who failed iLEAP ($p < 0.0001$).

### Relationship Between ORF and iLEAP

Table 2 presents the degree of association among the ORF and iLEAP variables. Pearson correlation coefficients were used to measure the strength of the linear relationship between the scores on the ORF and iLEAP tests. Pearson correlations between the ORF and iLEAP scores were statistically significant ($p < .01$) at all time periods. As expected, the fall third-grade ORF- iLEAP correlation (0.718, 95% CI [0.614,
0.805]) and winter third-grade ORF- iLEAP correlation (0.655, 95% CI [0.537, 0.753]) were stronger in magnitude than the first-grade winter ORF- iLEAP correlation (0.450, 95% CI [0.279, 0.601]). Unexpectedly, the fall third-grade ORF-iLEAP correlation was stronger in magnitude than the winter third-grade ORF-iLEAP correlation though not significantly stronger at the \( p < .05 \) level. Finally, the ORF measures were strongly related to each other (see Table 2 for point estimates and 95% confidence intervals).

**Table 2. Pearson Correlations Between ORF and iLEAP Scores for Third-Grade Sample**

<table>
<thead>
<tr>
<th>Reading Performance Measures</th>
<th>N</th>
<th>Correlation</th>
<th>95% Confidence Interval Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>iLEAP Winter 1st Grade ORF</td>
<td>55</td>
<td>0.45</td>
<td>[0.279, 0.601]</td>
</tr>
<tr>
<td>iLEAP Fall 3rd Grade ORF</td>
<td>80</td>
<td>0.72</td>
<td>[0.614, 0.805]</td>
</tr>
<tr>
<td>iLEAP Winter 3rd Grade ORF</td>
<td>81</td>
<td>0.66</td>
<td>[0.537, 0.753]</td>
</tr>
<tr>
<td>Winter 1st Grade ORF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter 1st Grade ORF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter 3rd Grade ORF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iLEAP Winter 3rd Grade ORF</td>
<td>53</td>
<td>0.65</td>
<td>[0.499, 0.777]</td>
</tr>
<tr>
<td>Winter 3rd Grade ORF</td>
<td>81</td>
<td>0.93</td>
<td>[0.898, 0.949]</td>
</tr>
<tr>
<td>Fall 3rd Grade ORF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter 3rd Grade ORF</td>
<td>77</td>
<td>0.93</td>
<td>[0.898, 0.949]</td>
</tr>
<tr>
<td>Difference Between Correlations</td>
<td></td>
<td>-0.06</td>
<td>[-0.125, 0.005]**</td>
</tr>
</tbody>
</table>

*Note: ORF = oral reading fluency; iLEAP = integrated Louisiana Educational Assessment Program.*

* Correlations significant at \( p < 0.01 \).

** Estimated using students that had scores on the iLEAP, the Fall 3rd grade ORF, and the Winter 3rd grade ORF.
Diagnostic Accuracy of ORF

Table 3. Cut Scores and Diagnostic Accuracy Statistics for Oral Reading Fluency Benchmark Periods

<table>
<thead>
<tr>
<th>Benchmark Period</th>
<th>Oral Reading Fluency Cut Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Winter First Grade (N = 55)</td>
</tr>
<tr>
<td>Area under the ROC Curve</td>
<td>.79</td>
</tr>
<tr>
<td>Cut Score</td>
<td>19</td>
</tr>
<tr>
<td>Sensitivity (True Positive Rate)</td>
<td>.78</td>
</tr>
<tr>
<td>Specificity (True Negative Rate)</td>
<td>.73</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>.27</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>.22</td>
</tr>
<tr>
<td>Positive Predictive Power (PPP)</td>
<td>.58</td>
</tr>
<tr>
<td>Negative Predictive Power (NPP)</td>
<td>.87</td>
</tr>
<tr>
<td>Hit Rate</td>
<td>.76</td>
</tr>
<tr>
<td>Kappa</td>
<td>.52</td>
</tr>
<tr>
<td>Phi</td>
<td>.56</td>
</tr>
</tbody>
</table>

Table 3 provides diagnostic information for evaluating the diagnostic accuracy of the ORF at each of the three time points as a predictor of failing iLEAP. Diagnostic accuracy refers to the ability of an instrument to distinguish between two diagnostic alternatives and select the correct one (Swets, Dawes, & Monahan, 2000). In the present study, researchers evaluated the ability of the ORF test (i.e., predictor) to successfully distinguish between students who would eventually fail the iLEAP (i.e., criterion) and those who would eventually pass. A pass was considered an iLEAP score at or above the basic level (i.e., ≥282), whereas a fail was a score below 282 (i.e., at the unsatisfactory or approaching basic levels). ORF scores at the three time points were considered separately. For each time point, the ROC curve was constructed and the area under this curve was calculated. The area under an ROC curve is one measure of diagnostic accuracy when using one variable as a predictor of another, and can be used to compare different variables with respect to their predictive accuracy.
ability. In the present context, given two randomly-selected students, one who would fail the iLEAP and the other who would pass, the area under the ROC curve can be interpreted as the probability that the ORF screening test correctly identifies the student who would fail the iLEAP (Rosner, 2000). For a given predictor variable, the closer the area under its ROC curve is to one, the better the predictor variable. For the ORF measured during the winter of the first grade, the area under the ROC curve was 0.79. For the ORF measured during the fall of the third grade, the area under the ROC curve was 0.86, and for the ORF measured during the winter of third grade, the area under the ROC curve was 0.84. From these values it is clear that the ORF scores measured during the third grade are superior to the ORF score measured during the first grade in predicting who will pass and who will fail the third grade iLEAP.

The next task in evaluating the screening test was to identify the appropriate ORF cut point. For a given cut point, diagnostic accuracy is characterized by the sensitivity (true positive rate) and the specificity (true negative rate) of the resulting screening test. How these measures are interpreted depends on what the predictive models are set up to predict; in the present context, their interpretations depend on whether the models are set up to predict students who will pass iLEAP, or set up to predict students who will fail. In this study models were set up to predict students who would fail the iLEAP. Therefore sensitivity referred to the probability that the ORF score correctly identified those students who would fail the iLEAP reading portion (i.e., scored below the basic level). Specificity referred to the probability that the ORF score accurately identified those students who would pass the iLEAP.

The positive predictive power (PPP) and the negative predictive power (NPP) are related to sensitivity and specificity and are useful when evaluating the reliability of predictions of a screening test in a given population. As in interpreting the sensitivity and specificity, the interpretations of the PPP and NPP depend on what the screening test is set up to predict. In the context of this study, PPP referred to the probability that those students who were predicted to fail the iLEAP actually did fail. NPP referred to the probability that those students who were predicted to pass the iLEAP actually did pass. Similar to ROC analysis, the stronger instruments are those whose PPP and NPP are as close to 1.0 as possible. It is important to note that the PPP and NPP depend not only on the ac-
accuracy of the screening test (i.e., the test’s sensitivity and the specificity), but also on the proportion of the population possessing the characteristic to be identified, in this case the proportion of students who will fail the iLEAP. A screening test with a given sensitivity and specificity, applied to different populations with different proportions of people possessing the given characteristic, will have different values of PPP and NPP when applied to the different populations. Because of this, the PPP and NPP are not fundamental measures of the accuracy of a screening test. This is not to diminish the importance of the PPP and NPP, but instead to clarify how and when these various measures are used when developing and using screening tests. Sensitivity and specificity are fundamental measures of accuracy which are used when developing a screening test. Once a screening test has been established and is being used in a particular population, the PPP and NPP become of primary importance, because those statistics indicate how much confidence assessors can have in a particular prediction.

In developing a diagnostic test it is obviously desirable that the test be as accurate as possible; that is, it is important for the sensitivity (true positive rate) and the specificity (true negative rate) of the test to be as close to 1.0 as possible. It must be kept in mind, however, that in a diagnostic test based on a cut score, sensitivity and specificity are inversely related. Therefore, changing the cut score in order to increase sensitivity can result in a decrease in specificity, and vice-versa. In addition, in some situations it may be more important to avoid false positives than to avoid false negatives, while in other situations the opposite may be true. In general, the criteria used in selecting a cut score must take into account the relative costs associated with false positives and false negatives, and is therefore dependent on the particular situation. In this study researchers used a procedure similar to that outlined in Silberglitt and Hintze (2005). Specifically, researchers first identified all cut scores for which the empirical sensitivity and specificity were both at least 90 percent. If no such cut scores were available, then all cut scores for which the empirical sensitivity and specificity were both at least 80 percent were identified. If no such cut scores were available, then all cut scores for which the empirical sensitivity and specificity were both at least 70 percent were identified. From the selected cut scores, the cut score providing the largest empirical sensitivity was chosen. For the first grade
winter benchmarking period, this procedure resulted in a cut score of 19 with an empirical sensitivity of 0.78 and empirical specificity of 0.73. For the third grade fall benchmarking period, the resulting cut score was 73 with an empirical sensitivity of 0.82 and empirical specificity of 0.80. For the third grade winter benchmarking period, the resulting cut score was 82, with an empirical sensitivity of 0.79 and empirical specificity of 0.72. These values, along with the empirical PPP and NPP and other measures assessing the predictive ability of these tests, can be found in Table 3.

**Change in ORF**

Students were categorized by fall third grade quintile groups, and the average increase in ORF scores from the fall to winter time points were calculated for each group (see Table 4). Analysis of variance was used to compare these average increases, which just missed being significant at the 0.05 level ($p = 0.0528$). For students in the lowest quintile group, there was a significant increase in average ORF scores from fall to winter of 11.06 points (95% CI [6.00, 16.12]). The average increases in ORF scores for the second and third quintile groups, 3.23 and 3.89, were not significant, however. The average increases in ORF scores of 9.00 and 12.38 for the last two quintile groups were statistically significant.

**Table 4. Changes in ORF Words Correct Per Minute (WCPM) from Fall to Winter Third Grade by Fall Third Grade Quintile Group**

<table>
<thead>
<tr>
<th>Quintile</th>
<th>N</th>
<th>Fall ORF 95% Confidence Interval Estimate</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>25 to 49 6.00 16.12 0.41</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>51 to 69 -2.56 9.01 0.12</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>71 to 83 -1.03 8.81 0.14</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>84 to 97 3.21 14.79 0.33</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>99 to 133 7.16 17.59 0.46</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>80</td>
<td>25 to 133 5.52 10.45 0.30</td>
<td></td>
</tr>
</tbody>
</table>

*Change scores significant at $p < .05$. Note: ORF = oral reading fluency curriculum based measurement scores.
DISCUSSION

Literature regarding the relationship between CBM and statewide measures supports the assertion that there is a strong positive association between the two measurement systems, particularly in the area of reading. Such a relationship is important given the fact that the way statewide tests are administered it is hard to use the data to improve instruction in a timely manner. The purpose of the present study was to replicate and extend findings reported in the peer-reviewed and published literature. Researchers hypothesized that the Pearson correlations, diagnostic accuracy statistics, and change score patterns for one third-grade body of students would be consistent with previous research in spite of the fact that a more diverse and urban sample from a different state was being analyzed. Use of third-grade data was important because federal legislation (i.e., NCLB) now mandates annual assessment beginning in third grade. Results generally supported the researchers’ hypotheses regarding the correlations and diagnostic accuracy statistics, but not the hypothesis concerning comparable change score patterns.

In placing the present findings in the context of the larger literature, both similarities and differences are worth highlighting relative to the correlation and diagnostic accuracy data. In terms of similarities, findings again demonstrated that readily available ORF probe scores were moderately to strongly correlated with scores from a statewide instrument, in this case Louisiana’s annual criterion-referenced reading test for third graders. Estimates of the Pearson correlations between the third grade iLEAP and the third-grade fall and winter ORF measures were 0.72 (95% CI [0.614, 0.805]) and 0.66 (95% CI [0.537, 0.753]), respectively, with the 95% confidence intervals affirming the relative strength of both associations, particularly the fall benchmarking period. The estimate of the correlation between third grade iLEAP scores and winter first grade ORF scores was 0.45 (95% CI [0.279, 0.601]), the confidence interval indicating a strength of relationship from small to large. Correlations in two previous non-diverse third-grade samples (i.e., Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005) ranged from 0.66 to 0.70, whereas coefficients in the diverse multi-year sample of fourth graders in McGlinchey and Hixson (2004) ranged from 0.49 to 0.81. These findings highlight the continued consistency with respect to the strong positive relationship.
between CBM and statewide assessment systems in third grade (Ax & Bradley-Klug, 2005; Crawford et al., 2001; Good et al., 2001; Hintze & Silberglitt, 2005; Powell, 2004; Silberglitt & Hintze, 2005; Silberglitt et al., 2006).

Second, the average fall ORF score of 75 WCPM was comparable to scores in related studies (i.e., 78 in Hintze & Silberglitt; 73 in Silberglitt & Hintze). Third, fall cut scores were also alike, with the present 73 similar to scores of Hintze and Silberglitt (i.e., 68) and Silberglitt and Hintze (i.e., 70). Finally, for this population, the third grade fall and winter cut scores obtained give estimated PPP of 0.76 and 0.68, respectively, and estimated NPP of 0.86 and 0.83. These values are within the ranges reported in previous studies (Hintze & Silberglitt; Silberglitt & Hintze) and in the same pattern (i.e., NPP greater than PPP).

Noteworthy differences in the present findings relative to correlational and cut score analyses centered on the winter third grade data. Unlike other studies in which longitudinal third-grade data were analyzed (e.g., Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005), the winter correlations in the present data were statistically equivalent to those for the fall benchmarking period. Significant differences have been noted in previous studies (Hintze & Silberglitt; Silberglitt & Hintze). One would hypothesize that the ORF measure closest in time administered to the statewide assessment – the winter benchmark score – would be the strongest of all the correlations as has been demonstrated in the literature. Also, the average ORF score for the winter benchmarking period was lower than those reported in other studies. Additionally, the winter cut score calculated for the present data was smaller when compared to cut scores calculated for other third grade studies. Whereas the fall cut scores were comparable (i.e., 73 vs. 68 or 70), the winter cut scores were disparate (i.e., 82 vs. 91 or 93)( Hintze & Silberglitt; Silberglitt & Hintze). These differences may be directly related to the present sample’s growth rates. When comparing average expected levels of growth based on norms tables available for the DIBELS and AIMSweb systems (both of which were used by this school’s staff), the WCPM increases from fall to winter are generally around 15 words across the DIBELS levels of risk and between 19 and 22 words when considering AIMSweb’s 25th-, 50th-, or 75th-percentiles. In the present study, the mean change in WCPM from third grade fall to winter ORF is estimated to be 7.98 (95% CI [5.52, 10.45]). And as
Table 4 indicates, a full 40 percent of the sample had mean growth rates of fewer than one word correct per month (or 3 to 4 WCPM from fall to winter benchmarking periods).

Growth pattern differences between the present study and Silberglitt and Hintze (2007) may be the most intriguing findings when viewed in the context of school’s districts’ responses to response to intervention (RTI). Silberglitt and Hintze demonstrated that in a large sample of third graders that average growth rates differed significantly between the group of students in the fifth decile and those groups of students in the first and second deciles, for example. Moreover, their analyses pointed to lower indices of weekly growth for those students whose fall ORF benchmark scores placed them in the bottom 20 percent. The present findings demonstrated non-significant (p = .0528) differences across quintiles. And while non-significant, larger mean WCPM scores were noted for the lowest 20 percent than for the middle 20 percent of the sample.

Two factors may contribute to an explanation of the growth patterns exhibited in the present data. The first factor may be the level of reading delays noted in the sample. Nearly half of the students (i.e., 37 of the 80 with fall ORF data) were at increased risk of failing the statewide test based on an application of the 73 WCPM cut score. The core curriculum and instruction relative to that core curriculum in these students’ first three years of schooling was apparently not meeting their early literacy needs. That noted, the second factor may have been the potential positive impact on WCPM of the school’s intervention efforts, which can be viewed as the school’s response to RTI. The school’s diagnostic team, which included its educational diagnostician and prereferral team coordinator, identified and initiated intervention efforts with all third graders whose fall ORF score placed them as at risk according to DIBELS guidelines (< 53 WCPM). That group, which essentially mirrors the first quintile in Table 4, received additional one-on-one and/or small group activities such as remedial phonics instruction and repeated reading which was provided by each student’s homeroom teacher. This group’s average WCPM growth pattern can be contrasted with the second quintile, which essentially consists of those students whose fall ORF score placed them at high risk of failing the iLEAP test based on this study’s cut score of 73. Yet, while the growth rates were about 8 WCPM greater for the first quintile as compared to the second, those increases did not
translate into higher passage rates on the iLEAP. As Silberglitt and Hintze (2007) reiterated, this body of findings confirms the critical function of early intervention, with ORF-informed efforts beginning as early as midway through first grade and other CBM-informed efforts being initiated in kindergarten.

**Implications**

Practitioner- and researcher-oriented implications can be drawn from the present study. We highlight two implications for practitioners and one for researchers. First, the present study was designed to determine whether the strong relationship reported in the literature between formative and summative reading assessment could be duplicated given a more diverse sample. The answer to that question was yes. The present sample was 48 percent minority. The present sample included a majority of third graders (i.e., 60 percent) who received free or reduced cost lunches. Yet in spite of demographics that were largely different than those reported in most previous research, particularly when it came to third-grade samples, the correlations between ORF and statewide test scores were well in line with those reported in the larger literature. The present findings add to the fourth-grade findings of McGlinchey and Hixson (2004) which indicated moderate to strong and statistically significant relationships between ORF and the state of Michigan’s statewide reading test. Overall, the collective research appears to support the assertion that it is the one-minute reading score that proves predictive of reading proficiency and not variables such as demographics or the individual statewide test that are outside the control of a school’s administrators, teachers, and staff. Once again, ORF “appears to transcend the differences in state assessments” (Shapiro et al., 2006, p. 28). School-based practitioners, no matter what make-up of the school or district, can be confident that the use of ORF as a screening tool beginning as soon as midway through the first grade year will allow them to accurately identify students who are likely targets of preventative reading intervention.

Two, the importance of districts gathering and continuously evaluating their own data cannot be overstated. In the case of the present school, for example, intervention team members relied on the DIBELS national benchmarks in order to determine which students needed
additional targeted intervention. In the case of this school, during the 2005-06 year, third-grade students scoring below 53 WCPM during the fall benchmarking period \( (n = 18, 23 \text{ percent}) \) began receiving supplemental reading intervention once the data were analyzed. In an RTI context, then, nearly 1 in 4 third graders began receiving Tier 2 intervention. However, if school officials had instead used their own cut score of 73 as the gate keeper for entry into Tier 2 intervention, then an additional 19 students (another 24 percent) would have begun receiving supplemental intervention services, accounting for nearly half the class. With supplemental intervention being provided to additional students identified as in need of such services, one might expect the number of students who reached proficiency on the statewide test to increase. Such an assertion is supported by a post hoc analysis of the present data, which indicates that the average gain between the fall and winter benchmarking period was nearly 11 WCPM for the Tier 2 group whereas those who were not chosen but could have been based on a 73 WCPM cut score only grew by a little more than 3 WCPM.

Ongoing analyses of local data may have also contributed to school administrators and curriculum specialists taking a closer look at both the core and supplemental curricula being implemented at a schoolwide level. As one indication, more than 40 percent of the class began third grade at some risk of reading failure whether the benchmark was the local cut score or the DIBELS’ national norms. Such a figure might have been an indicator that a more evidence-based curriculum or specialized teacher training could be beneficial in the earlier grades. As a second indication, comparisons of fall to winter benchmarking periods in third grade demonstrated that the average gains being reported for the grade (i.e., about 8 WCPM) lagged behind national growth expectations from DIBELS and other sources and relevant research (e.g., Hintze & Silberglitt, 2005). Evaluation of these data may have also facilitated school administrator decision-making regarding curriculum and/or professional development. Overall, while national benchmarks can provide a starting point for evaluative discussion, it is important for educational stakeholders to use their own formative and summative assessment information to make instructional decisions. In the case of the present data, a local perspective on Silberglitt and Hintze's (2005) statement that cut scores
were “remarkably consistent” (p. 320) across data bases would be accurate for the fall and inaccurate for the winter benchmarking periods.

For researchers, the use of ORF in particular and CBM in general as a screening tool in the educational field opens the door to myriad questions about how we apply diagnostic accuracy in school classrooms, buildings, and districts. In this study, this research team followed the lead of Silberglitt and Hintze (2005) and essentially applied decision rules that chose to balance the highest levels of diagnostic tool sensitivity and specificity. In this manner, we balanced ORF's ability to accurately identify those who passed and those who failed the iLEAP reading test. But just what is the ‘best’ criterion. Swets (1992) noted that there have been three historical approaches to determining decision rules, “one very strict, one very lenient, and one in the middle” (p. 524). In describing the three approaches, Swets noted that the strict criterion yields low false positive rates, the lenient criterion yields high true positive rates, and the medium criterion maximizes true positives and true negatives. The education literature to date has relied upon the latter. Optimal criterions, according to Swets, take into account a number of variables, including the benefits and costs of correct and incorrect decisions that arise as a result of the application of cut scores. Educational research in future years might inform judgments as to just what is the optimal set of decision rules for not only Louisiana schools, their professional personnel, students, and parents, but also those nationwide.

Limitations and Future Research

Despite the fact that statistical analyses demonstrated significantly strong predictive correlations between ORF and iLEAP reading assessments for students of diverse socioeconomic, cultural, educational, and gender backgrounds, some limitations in the present research are noteworthy. One, students with missing data required for a particular analysis were excluded from that analysis. This most directly impacted the first-grade analyses, wherein nearly 35 percent of the total sample did not have first-grade CBM data available. It is possible that a complete data set from that group of students could have statistically significantly impacted the presented findings. Moreover, it is possible that the subset of students without first-grade scores could provide important in-
tervention-related information to educational stakeholders considering that this group of students likely has had multiple school placements in the primary grades. Research is needed that compares both the benchmark reading scores, growth rates, and relationships between formative and summative assessment measures relative to these two groups of students. Two, the present study only included data from one school’s third grade which did not include students for whom English was their second language. Results may differ depending on the makeup of any given grade, school, and/or district. Additional research is warranted. Three, since this study only focused on one measure of reading fluency, other formative reading assessment measures might yield differing results.

**CONCLUSION**

Bushell and Baer (1994) suggested that “measurably superior instruction” (p. 3) results from “close, continual contact with relevant outcome data” (p. 7). Findings from this study and the larger applicable literature demonstrate that ORF provides relevant outcome data that can influence local instruction in ways that allow for increased numbers of students to demonstrate reading proficiency by earning passing marks on annually administered statewide assessments. In today’s accountability climate, educators need more data than that provided by end-of-the-year summative assessments. Research is continuing to demonstrate that the marriage of progress monitoring and statewide accountability systems might adequately serve the needs of educational stakeholders. The present findings affirmed the strong association between ORF and statewide tests by adding Louisiana to the list of states with statewide tests that are highly correlated with the CBM fluency measure. The present findings also affirmed the utility of cut scores by demonstrating the continued predictive accuracy as well as the consistency in scores across grades. The present research findings lend credence to the field’s ongoing efforts to build and refine RTI frameworks to ameliorate the academic deficits of those students that data determine to be at risk for current and future school failure. These data affirm a school teacher’s, principal’s, and/or system’s responsibility for regularly gathering and cri-
tically evaluating – that is, maintaining close, continual contact with – meaningful outcome data.

Acknowledgment

The authors greatly appreciate the dedication and effort demonstrated by the principal, educational diagnostician, school building level committee chair, and lead special education teacher at the school from which these data were gathered. Procedures approved by the school district and university Institutional Review Board prohibited the researchers from accessing student names in order to maintain student confidentiality. Therefore, school officials were entirely responsible for data gathering and entry. Their efforts were outstanding.

*Note

The real name of the school and town in which the school is located were withheld from the title page in order to maintain confidentiality and comply with an agreement between the researchers and district administrators.

References


Good, R. H., Simmons, D. C., & Kame‘enui, E. J. (2001). The importance and decision-


Review, 30, 407-419.

PAPILDOMI SKAITYMO GARSIAI ĮGŪDŽIŲ NAUDINGUMO,
PROGNOZUOJANT MOKINIŲ BENDRUOSIUS SKAITYMO GEBĖJIMUS, ĮRODYMAI


Pagrindiniai žodžiai: skaitymo garsiai įgūdžiai, Luizianos švietimo programos metodai skaitymo įgūdžių vertinti, dinaminis pagrindinių ankstyvojo raštingumo įgūdžių vertinimas, bendrieji skaitymo gebėjimai.

Received November 3, 2008
Accepted December 3, 2008