2009

Predicting performance on high stakes testing: validity and accuracy of curriculum-based measurement of reading and writing

Marie Young Henderson
Louisiana State University and Agricultural and Mechanical College

PREDICTING PERFORMANCE ON HIGH STAKES TESTING: VALIDITY AND ACCURACY OF CURRICULUM -BASED MEASUREMENT OF READING AND WRITING

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Psychology

by
Marie Henderson
B.S., Louisiana State University, 2002
M.A., Louisiana State University, 2005
August, 2009

DEDICATION

I would like to dedicate this dissertation to my children. It is through your inspiration and love that I was able to persevere. May you learn the importance of education and forever be a student, learning from life's course. Pursue your dreams through all obstacles and challenges. I will always believe in you, just as you have believed in me.

ACKNOWLEDGEMENTS

I would like to thank Dr. George Noell for his continuous guidance and support throughout the conceptualization and preparation of this document. I would also like to thank Dr. Kristin Gansle for support with technical questions and editing. Additionally, I would like to thank Dr. Frank Gresham and Dr. Jeffery Tiger for being members of my dissertation committee. I would like to thank Andrea Ruckman, Dwaine Henderson, and Eric Penalber for helping with data collection. Finally, I would also like to thank Warren and Susan Young, and Lyn and Kaye Henderson for all the babysitting, love, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The purpose of the current investigation was to determine which curriculum-based measures of written expression demonstrated adequate technical characteristics and provided useful information towards predicting performance on a state-standardized assessment. Data collected from 124 third grade students was used for the study. Curriculum-based measures of reading and writing collected three times within the school year were utilized as the independent variables for predicting the dependent variables. Writing samples were scored using 9 indices of writing. Results from a state standardized assessment (*i*LEAP) were used as the dependent variables. The study found reliability coefficients for writing indices to be consistent with previous investigations. Principle components analysis revealed a consistent three component solution for the writing indices across benchmark periods. Regression analyses revealed percent correct word sequences, fall words spelled correctly, and winter complete sentences to be significant predictors; however, only fall words spelled correctly and winter complete sentences contributed to fall oral reading fluency for predicting the passing status of students.

INTRODUCTION

Early identification and prevention of academic problems has been found to be efficient and effective for increasing academic achievement (National Association of Sate Directors of Special Education (NASDSE; 2005). Recent legislation, No Child Left Behind (NCLB; 2004), supports the utilization of evidence based and scientifically validated instructional practices to improve learning outcomes for all students. NCLB, the reauthorization of Elementary and Secondary Education Act, has a preventative focus which complements and matches response to intervention (RTI) through the incorporation of evidence-based practices in systematic and data-driven application (Brown-Chidsey & Steege, 2005).

The National Association of State Directors of Special Education (NASDSE, 2005) outlined the core principles of RTI. These principles may be useful for guiding policy and regulations implemented by state education agencies (SEA). The first principle states that all children can be effectively taught. Therefore, it is our responsibility that educational conditions enable learning for all children. Early intervention is the second principle. By intervening early with academic and behavior concerns the problems are addressed while they are still relatively small. This will optimize the efficiency and effectiveness of the intervention. Service delivery should be deployed through a tiered model. To facilitate meeting the needs of all children instruction will necessarily vary in nature and intensity. The multi-tier model is a needs-driven, resource deployment system that enables practitioners to match instruction to ability. Decisions are made within the multi-tier model through problem solving. Four core questions must be answered through problem solving (NASDSE, 2005): Is there a

problem and what is it?; Why is it happening?; What are we going to do about the problem?; and Did our intervention work?

The last four principles center around the use of data and scientifically-based decisions. As stated in NCLB (2001) and the Individuals with Disabilities Education Act (IDEA, 2004), scientifically-based instruction and interventions should be used. Data should be gathered through frequent collection of sensitive measures to monitor student progress. This will allow data-based decisions on the effectiveness of the instruction and intervention being used. This point is central to the RTI model. Data-collection should be on-going and provide adequate information on targeted student progress to enable informed instructional decisions. Within the RTI model assessments are used for three main purposes: to identify children not making adequate progress; to determine what the children can and cannot do in the targeted area of concern; and to progress monitor intervention effectiveness (NASDSE, 2005).

Identifying assessment systems that are reliable, valid, and sensitive enough for the purposes previously described is essential. One such measure is Curriculum-Based Measurement (CBM). CBM is a brief, standardized measure of academic skills (Shinn, 1995) that is gaining prominence in schools for use within a problem-solving framework (Malecki & Jewell, 2003). All steps of the problem-solving process involving academic concerns utilize CBM technology. These steps include problem identification, instructional placement, goal-setting and intervention planning, progress monitoring, and eligibility decisions (Fuchs & Fuchs, 1997).

CBM assesses the effects of instruction efficiently and accurately, while utilizing a methodology that allows for formative assessment of student performance (Fewster &

MacMillian, 2002; Hintze, Christ, & Methe, 2006). Initial development of CBM was directed at assessing the effectiveness of a special education intervention model called data-based program modification (DBPM; Deno & Mirkin, 1977; Deno, 2003). The model was based on the idea that teachers could use repeated measurement of student performance to formatively evaluate the effectiveness of and improve their instruction. Key research conducted during the development of CBM addressed the necessary requirements of technical adequacy and ease of implementation. CBM measures were constructed to be short samples of work that would reflect performance on key skills or "vital signs" of academic performance (Deno, 1985; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). Curriculum-based measurement was designed to address the need for an instrument that could provide frequent, informative feedback about student progress. Traditional published psychological and educational tests were problematic for this type of decision making for multiple reasons. Psychometrically developed tests lacked testing-teaching overlap, the utility for instructional decision making did not exist, test items measured skills indirectly, fluency of responding was not considered, and the pre-post test design for evaluating change was inadequate and insensitive to pupil growth (Marston, 1989).

The characteristics necessary for monitoring student progress form the foundation of curriculum-based measures (Jenkins, Deno, & Mirkin, 1979). The criteria require that measures have to be (1) capable of having several forms, (2) sensitive to student progress over time, (3) overlapped with students' curricula, (4) inexpensive, and (5) time efficient to facilitate frequent administration (Jenkins et al., 1979; Marston, 1989). Earlier investigations and recent studies have provided sufficient evidence for utilizing the

number of words read in one minute as a reliable and valid indicator of reading proficiency over time. This information is currently being used to address reading concerns early in the school year to improve student performance. Alternatively, studies addressing the validity and reliability of CBM for written expression have produced inconsistent results. Measures that were found in early investigations to have adequate reliability and validity failed to provide evidence for adequate levels when subjected to further investigations. Recent studies have revealed promising evidence for newer indices of written expression, but continued investigation of the generalization of the validity of these measures for a variety of populations is necessary. Furthermore, the relationship of these measures with performance on major standardized state assessments is needed for establishing their utility as an indicator of student progress throughout the school year. The goal of the current study is to examine which indices produce the most useful information obtained from writing samples and their relationship to a major standardized state achievement test.

REVIEW OF LITERATURE

Identification of academic behaviors representing the necessary basic skill content areas that could be measured reliably and validly was first addressed by Stanley Deno and Phyllis Mirkin through the studies conducted by the Institute for Research on Learning Disabilities (IRLD). Initially, measures were identified though an extensive analysis of the literature and were subsequently reviewed to determine which measures adequately represented the established criteria. Measures determined to be representative of the criteria were field tested for criterion-related validity, reliability, and logistics of measurement (e.g. length of testing interval, size of the measurement domain) (Marston 1989).

Marston (1989) reviewed the early research conducted on curriculum-based measures. The first validity study for reading compared students' 1-minute oral reading performance utilizing their basal reader to published norm-referenced tests (Deno, Mirkin, & Chiang, 1982). The criterion measures selected for this study included the Stanford Diagnostic Reading Test (Karlsen, Madden, & Gardner, 1976), the Woodcock Reading Mastery Test (Woodcock, 1973), and the Reading Comprehension subtest from the Peabody Individual Achievement Test (Dunn & Markwardt, 1970). The 1-minute oral reading sample from students' basal readers was a valid measure, with correlation coefficients ranging from .73 to .91. Subsequent studies revealed oral reading from basal readers to have correlation coefficients ranging from .63 to .90 with different measures of global reading skills. Most coefficients were above .90. Subtests of these global measures yielded lower yet adequate correlation coefficients ranging from .53 to .91 with half of the coefficients exceeding .80. The lower correlations between some subtests and

oral reading fluency would be expected due to the low reliabilities of the subtests (Fuchs & Deno, 1981; Fuchs, Fuchs, & Maxwell, 1988; Marston, 1982; Marston, 1989; Marston & Deno, 1982).

Criterion-related validity was also ascertained by comparing reading fluency measures and criterion-referenced mastery tests from four different basal reading series. Correlations between reading from passages and total test scores ranged from .65 to .86 with three of four coefficients higher than .80 (Fuchs, Tindal, Fuchs, Shinn, Deno, & Germann, 1983; Fuchs, Tindal, Shinn, Fuchs, Deno, & Germann, 1983; Tindal, Fuchs, Fuchs, Shinn, Deno, & Germann, 1983; Tindal, Shinn, Fuchs, Fuchs, Deno, & Germann, 1983; Marston, 1989). These comparisons revealed correlations between curriculum-based measures and basal mastery tests directly proportional to mastery measures' correlation with more global measures of reading proficiency. Thus, curriculum-based measures shared more variance with basal mastery tests that were highly correlated with general measures of reading than measures less related to other measures of reading ability (Marston, 1989).

Construct validity has also been studied, beginning with the early investigations conducted by Deno and his colleagues. This was done by investigating discriminant validity and treatment validity. Discriminant validity was assessed by determining the degree to which oral reading samples distinguished between intact groups that differed theoretically in their reading skills. One-minute oral reading samples were found to reliably differentiate Chapter I and regular education first-, second-, and third-grade students from learning disabled students (Deno, Marston, Shinn, & Tindal, 1983). Shinn and Marston (1985) replicated this finding. They found words read aloud to differentiate

between regular education, Chapter I, and mildly handicapped students with learning difficulties (Marston, 1989).

Treatment validity has been addressed by employing longitudinal studies of reading growth. A valid measure of reading should show to be sensitive to growth as student skills improve (Marston, 1989). Reliable gains were established in a cross-sectional study of oral reading fluency across grades first though sixth utilizing a sample of 550 students (Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins, 1982). In another study, student progress was examined by administering standardized reading tests and CBM procedures (Marston, Fuchs, & Deno, 1986). Examination of the short-term reading progress of students across 10-week and 16-week intervals revealed student improvement on both measures; however, CBM procedures were more sensitive to gains (Marston, 1989).

Reliability studies have yielded impressive findings. Studies reviewed by Marston (1989) included test-retest reliability, parallel form estimates, and interrater agreement coefficients. Test-retest reliability assessed over a 2-week period and a 10-week period yielded reliability coefficients above .90 (Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983). Parallel form reliability assessed at the same time produced a correlation of .94, while alternate forms assessed over one week revealed a correlation of .89 (Tindal, Germann, et al., 1983; Tindal Marston, et al., 1983). Interrater agreement coefficients were found to be.99 (Tindal, Marston, et al., 1983).

Despite early evidence for technical adequacy, researchers and practitioners questioned the relation between reading aloud for 1-minute with reading proficiency, mainly the proficiency in reading comprehension (Mehrens & Clarizio, 1993; Yell, Deno,

& Marston, 1992; Wayman et al., 2007). Wayman et al. (2007) provided an updated review of the literature on CBM in reading. They determined that researchers have further examined the relation between reading aloud and general reading proficiency by focusing on two different approaches. First, researchers aimed to clarify the relationship between oral reading and reading comprehension by assessing alternative measures that might reflect reading comprehension performance more accurately and examining the theoretical relation between reading aloud and reading proficiency. A second approach aimed to address the relationship between oral reading and reading comprehension as a function of concomitant change for the individual student.

Fuchs et al. (1988) attempted to provide support for the validity of oral reading as more than just a measure of fluent decoding. They compared the validity of CBM oral reading measures to measures that are typically used to assess reading comprehension. These measures included story retell, cloze (every seventh word is deleted from the text and replaced with a blank), and question-answering measures. Participants included students in grades 4 through 8 with mild disabilities. Results revealed strong correlations for oral reading with scores on the word skills and comprehension subtests of a standardized achievement test (r = .80 and .91, respectively). These correlations were stronger than those found for the other typical measures of reading comprehension (rs = .76 to .82 for reading comprehension and .66 to .76 for word skills subtests). These results provided support for oral reading as more than just a measure of fluent decoding (Wayman et al., 2007).

Subsequent research attempted to address the theoretical nature of the relationship between oral reading and reading comprehension. One study utilized factor analysis to

determine the role of oral reading as it related to decoding, fluency, and reading comprehension skills (Shinn, Good, Knutson, Tilly, & Collins, 1992). Shinn et al. assessed the reading skills of students in grades 3 and 5. Results revealed a single-factor model of "reading competence" for third-graders that included significant contributions from all reading skills. However, for fifth-graders, a two-factor model was validated that included decoding and reading comprehension as two separate but highly related factors, with oral reading loading on the decoding factor. Other researchers also observed the changes associated with age in the relationship between reading aloud and reading proficiency. Hosp and Fuchs (2005) assessed children in first through fourth grades on oral reading and *Woodcock Reading Mastery Test-Revised* (WRMT; Woodcock, 1987). Relationships between the Decoding, Word Reading, and Comprehension subtests of the WRMT and CBM oral reading were similar for grades 2 and 3 (ranging from .82 to .88). In contrast, grade 4 correlations were lower for the Decoding and Word Reading subtests (*rs* = .72 and .73, respectively) than for the Reading Comprehension subtest (*r* = .82; Wayman et al., 2007).

While studies comparing patterns of results across groups are important, they do not provide information about the relationship between CBM oral reading and reading comprehension for the individual student (Wayman et al., 2007). Markell and Deno (1997) examined whether a concomitant change existed between oral reading and reading comprehension. Participants included third grade students for whom the difficulty level of reading material was manipulated. Students read passages that were two levels below, at, and two levels above grade level. Two comprehension tasks, maze and question answering, were also completed for each passage. On average, student performance

revealed significantly fewer words were read in 1 minute on the more difficult passages, selected fewer correct maze choices, and answered fewer questions correct. These results support a general relation between oral reading and reading comprehension.

Diagnostic Accuracy of Oral Reading Fluency

Research on oral reading fluency has provided strong evidence for oral reading fluency to be used as a measure of general reading proficiency (Wayman et al., 2007). As a measure of reading proficiency, oral reading fluency has been further examined to include its use as a predictor variable of high stakes testing performance. Specifically, investigators have conducted analyses to identify cut scores that could be used to help determine performance on high stakes tests, or state assessments (Good, Simmons, & Kame'enui, 2001; Hintze & Silberglitt; 2005; McGlinchy & Hixson, 2004; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Stage & Jacobsen, 2001; Wood, 2006). High stakes tests are used to assess students' level of progress in key academic areas. At a district and state level, results are used for holding educational systems accountable for the performance of all children to ensure utilization of effective instructional practices. For individual students, state assessments, or high stake tests, are used to determine if students are making appropriate educational progress and performing at a level necessary for advancement to more challenging instruction. However, these annual assessments typically provide parents and educators with too little information at a time of the year when it is too late to provide remediation (McGlinchey et al., 2004). Since curriculum-based measures such as oral reading fluency can be administered frequently and provide information that is sensitive to improved performance, understanding the relationship

between CBM and performance on state assessments is key to identifying students early in the school year who may need and benefit from academic remediation strategies.

Stage et al. (2001) calculated diagnostic efficiency statistics to identify students who were most likely to fail the Washington Assessment of Student Learning (WASL). Statistical measurements used in this study to assess the diagnostic accuracy of oral reading fluency included sensitivity, specificity, positive predictive power, and negative predictive power. Sensitivity referred to the percentage of students who had a score below the ORF cut score and failed the WASL. Specificity was the percentage of students who had a score higher than the ORF cut score and passed the WASL. Positive predictive power (PPP) was the probability that a student with a score below the ORF cut score would fail the WASL. Negative predictive power (NPP) was the probability that a student with an ORF score above the cut score would pass the WASL. Overall accuracy represented the measurement of agreement versus disagreement between the cut score and diagnostic criteria. Participants included 173 fourth graders who attended a school where 15% of the student population was eligible for free or price-reduced lunch. CBM reading fluency benchmark assessments were conducted at the end of September, January, and May. Passages used for measuring oral reading fluency were selected from the basal reading series used by the teachers. The WASL was administered in May. Pearson correlations between the ORF measures (words read correctly in one minute assessed in September, January, and May) and the WASL standard score for the reading assessment ranged from .43 to .44. Correlations with level of performance on the WASL for meeting reading proficiency standards ranged from .50 - .51.

The researchers utilized three analyses of variance (ANOVA) using WASL level scores to determine the three cut scores for the oral reading fluency measure (Stage et al., 2001). The number of correct words read per minute (wrpm) at each benchmarking period needed for passing the WASL was determined. The reading level at each benchmark was then assessed for its diagnostic accuracy. The base rate for failing the WASL was 20% and 80% for passing. The sensitivity of the September cut score (100 wrpm) for correctly identifying who failed was 66%. The specificity for correct identification of passing was 76%. The probability of correctly predicting (PPP) who would fail was .41 or correct identification of 41% as failing. This is above the base rate of 20%. The probability of correctly predicting who would pass (NPP) was .90, which is 10% higher than the actual passing rate. Results of these analyses revealed a difference between the statistics for September, January, and May to be 1% or .01. Thus, the findings were almost exact at each testing period. The authors concluded that these results supported Deno's (1985) assertion that ORF can be used as a "vital sign" of reading achievement. The correlations between ORF and WASL performance were .43 to .44. The authors attributed this medium effect (Cohen, 1992) to the WASL method for measuring reading. Specifically, short-answer and extended written responses were required components for the reading comprehension section which signifies that the WASL was not only measuring reading but also writing.

The Oregon Statewide Assessment (OSA) for reading/literature does not require written responses but utilizes a multiple-choice format to assess achievement level of individual students (Oregon Department of Education, 2000). In a study designed to assess the validity of a continuum of fluency based literacy skills, Good et al.(2001)

determined the strength of the relationship between CBM ORF and third grade high-stakes reading outcomes based on performance on the OSA. Passages used for assessing third grade ORF in this study were from the Test of Reading Fluency (Children's Educational Services, 1987). These are a standardized set of passages that are calibrated for grade level. The technical adequacy of the Test of Reading Fluency has been confirmed through test-retest and alternative form reliability studies (Tindal et al.1983); along with, criterion-related validity studies (Good & Jefferson, 1998; Good et al., 2001). Each student was administered three passages. The median correct words per minute from the three passages was selected as the ORF rate. A correlation of .67 existed between third grade ORF assessed in the spring and the OSA administered in the spring of third grade.

The level of proficiency on CBM ORF predictive of successful attainment on the state standard was described in terms of the probability of correctly predicting who would pass (NPP; Good et al., 2001). Of the 364 third grade students administered both the ORF and OSA assessments, 198 attained the May ORF goal of 110 wrpm. Of the students that met benchmark, 191 or 96% met or exceeded expectations on the OSA (scoring 201 or greater). Other diagnostic accuracy statistics (sensitivity, specificity, and positive predictive power) were not reported for the cut-score of 110. The authors noted that the likelihood of meeting expectations was less clear for students who read between 70 and 110 words per minute; however, no statistics were reported for this group. Almost half of the students fell within this group (n for students scoring between 70 and 110 wrpm = 166). Of the students who scored below 70 for ORF, 28% met expectations (13 of 46 students). The authors concluded that these results supported the utility of

accurately and fluently reading as an indicator of reading competence. However, while the correlation was large (Cohen, 1988) the diagnostic efficiency of the cut score is unclear. The authors noted the effectiveness of the cut score to predict who would pass the state test, but failed to provide information on the cut score's ability to predict those who would fail.

McGlinchey et al. (2004) noted the importance and utility of curriculum-based measurement's (CBM) sensitivity to predicting future performance on the high stakes tests. CBM could be used to monitor progress toward, and predict future performance on, the state assessment, while providing teachers with invaluable information throughout the year regarding the effectiveness of instruction. Establishing appropriate benchmarks for CBM is necessary to achieve these goals. McGlinchey et al. sought to replicate and extend the findings of Stage et al. (2001) by investigating the predictive validity of ORF CBM in relation to the Michigan Educational Assessment Program's (MEAP) fourth grade reading test, utilizing a sample of scores across 8 years (1994-2002; n = 1,362). Passages were selected from the basal fourth grade reading text and screened using the Fry (1977) readability formula to ensure that all passages were at a fourth grade reading level. Students were administered the reading passages once a year, two weeks prior to taking the MEAP. The authors determined the accuracy of the reading rate by analyzing the data and utilizing diagnostic efficiency statistics. They set the reading rate cut score at 100 wrpm (based on research conducted by Fuchs & Deno, 1982; Hasbrouck & Tindal, 1992; Stage et al., 2001). Concurrent, criterion-related validity of ORF was calculated each year by correlating the reading score with MEAP raw scores. Correlations were

fairly consistent for all years (range = .63 to .81) except for the 1998-1999 school year (r = .49).

Diagnostic efficiency statistics revealed the specificity for identifying those who did achieve Satisfactory on the MEAP to be 74% and the sensitivity for identifying those who did not was 75% (McGlinchey et al., 2004). The probability of correctly identifying those who achieved Satisfactory (NPP) was 72% while the probability of correctly identifying those did not was 77%. These statistics provide an improvement in prediction above base rate. The base rate of achieving a Satisfactory score was 46%, and for not achieving the Satisfactory score was 54%. The authors noted that cut scores can be raised and lowered depending on the level of confidence in which the school district is interested; however, utilization of a higher cut score that achieves a higher probability of predicating a passing score also decreases the probability of predicting failure. For example, the cut score recommended by Howell and Nolet (2000) is 140 wrpm. Application of this cut score would have revealed 84% of students reading at or above this rate achieving a Satisfactory score, but 39% of students reading less than 140 wrpm would have also received a Satisfactory score. Results from this study support the utility of ORF CBM for predicting performance on state-mandated high stakes testing. However, the ORF samples were only taken at a single point during the school-year (2 weeks prior to state testing). The benefit of CBM is it's ability to be administered on multiple occasions, over time, while helping to guide and improve instruction.

Utilizing CBM to predict performance over longer time durations has implications for directing instruction at time periods well in advance of administration of high stakes testing. Hintze et al. (2005) analyzed the concurrent and predictive validity of ORF CBM

over longer time durations with the Minnesota Comprehensive Assessment (MCA). The first purpose of their study was to compare commonly used statistical approaches to standard setting and cut score determination. (i.e., discriminant analysis, logistic regression, and receiver operator characteristic (ROC) curves). This was done to provide information on which would be most advantageous for determination of ORF CBM performance standards. The second goal of the study was to compare two different approaches for making prediction over time. The authors wanted to determine which method would produce the most sensitive cut-scores to help guide early instruction and intervention. The differential procedures compared for establishing cut-scores were the use of constant prediction of performance on high-stakes test or prediction to successive ORF CBM benchmarking periods. Thus, they manipulated which variable, performance on high stakes test or CBM performance, was used as the predictive measure for establishing the most sensitive cut-scores.

Data for the 1,766 participants were collected over three years in seven elementary schools (Hintze et al., 2005). The standard benchmark reading assessment passages for first through third grade were purposively developed with controlled vocabulary and difficulty. Each participant was assessed eight times with the ORF measure starting in the winter of grade 1 and continuing each Fall, Winter, and Spring until the Spring of grade 3. The reading portion of the MCA was administered in the Spring of third grade. The predictive validity of the ORF measure was assessed through correlational analyses. Results suggested strong validity with a range between the lowest coefficient at .49 for the first grade ORF winter score to .69 for the third grade spring ORF score.

The analysis of the diagnostic accuracy of ORF CBM was conducted by comparing the three statistical methods using two different criterion methods (Hintze et al., 2005). Results from the *discriminative analysis* revealed that cut scores derived using ORF CBM in a successive fashion across grades to ultimately predict MCA performance lead to improved precision for identifying those students likely to fail the MCA, as compared to using the MCA as the criterion. Using this method (ORF CBM as the criterion) resulted in consistently higher cut scores. Ranges (across benchmarking periods) for the diagnostic accuracy statistics were as follows: sensitivity, .82 to .95; specificity, .77 to .93; PPP, .87 to .97; and NPP, .69 to .91. When using MCA as the criterion variable, the ranges for diagnostic accuracy statistics were as follows: sensitivity, .50 to .65; specificity, .86 to .87; PPP, .79 to .81; NPP, .52 to .75.

Results of the *logistic regression analysis* revealed cut scores to be roughly equal across the two different criterion measures (differences ranged from 1 to 6 points; Hintze et al., 2005). Similar to results from discriminative analyses, using ORF CBM in successive fashion as the criterion resulted in higher levels of sensitivity, specificity, PPP, and NPP. Results from the *ROC curves* resulted in consistently lower cut scores when using ORF CBM as the criterion, which is in contrast to results from discriminative analysis. Also, consistent with logistic regression, ORF CBM resulted in higher sensitivity, specificity, PPP, and NPP.

The authors concluded that using ORF CBM in a successive manner to establish cut scores at each benchmark assessment for ultimately predicting MCA performance appeared to be more accurate and efficient than using MCA as the criterion (Hintze et al., 2005). Also, each statistical procedure set cut scores that yielded adequate levels of both

diagnostic accuracy and efficiency. Thus, ORF CBM appears to be a measure that is efficient for predicting performance on high-stakes tests as far back as first grade. In regards to which statistical analysis and prediction method to use, the authors suggest that the resources, expertise, and data collection abilities of the school district will make the determination. Districts with a high level of expertise may consider using ROC curves because of their flexibility for providing cut scores across a variety of assessment decisions. Logistic regression or discriminant analysis may be best used for districts that desire a single set of cut scores. However, logistic regression may be more desirable because it produced cut scores that are (a) both accurate and efficient, and (b) highly similar regardless of the criterion used (Hintze et al., 2005).

The technical adequacy and diagnostic efficiency of CBM for reading has been well established. However, determining an effective progress monitoring tool for writing is more complex. A curriculum-based measure for writing must be able to be administered frequently, be sensitive to growth in performance, and essentially quantify a skill that is typically assessed for its quality.

Curriculum-Based Measurement for Written Expression

Earlier research conducted at the University of Minnesota's Institute for Research on Learning Disabilities (IRLD) attempted to validate the use of measures of written expression in the 1970s and early 1980s (Marston, 1989). Initial studies focused on the concurrent validity of written expression measures with standardized achievement test performance (Deno, Mirkin, & Marston, 1980; Deno, Marston, Mirkin, 1982). Six measures of written expression were correlated with the following standardized criterion measures: Test of Written Language (TOWL; Hammill & Larsen, 1978), Developmental

Sentence Scoring System (DSS; Lee & Canter, 1971), and the language subtest of the Stanford Achievement Test (SAT; Madden et al., 1978). The six measures of written expression assessed in these studies included (1) number of words written, (2) number of words spelled correctly, (3) number of correct letter sequences, (4) number of mature word choices, (5) number of large words written, and (6) Hunt's (1965) average *t*-unit length. Of these measures, words spelled correctly, correct letter sequences, mature words, and total words written were highly related to the criterion measures (Marston, 1989). Results from Deno et at. (1980) also indicated that compositions could be written using a topic sentence, picture stimuli, or story starters and vary in length from 2 to 5 minutes with equivalent results. Correlations for words spelled correctly ranged from .45 to .92, with most above .70. Correct letter sequence correlations ranged from .57-.86. Total words written had correlations that ranged from .41 to .84 when compared to the criterion measures.

Videen, Deno, and Marston (1982) extended this research by assessing the validity of correct word sequences (CWS; any two adjacent, correctly spelled words that are acceptable within the context of the writing sample). Writing samples from Deno et al. (1980) were selected randomly and scored for this measure. Results revealed weak to moderate correlations with the TOWL (r = .69) and the DSS (r = .49). Teacher holistic ratings of the writing samples were also used as a criterion measure in this study. Correlations between these ratings and CWS were relatively strong (r = .85). The total words written measure also correlated with teacher holistic ratings of writing skill at .85 (Videen et al., 1982).

Several types of reliability were also assessed by the IRLD researchers for total words written (WW), words spelled correctly (WSC), and correct letter sequences (CLS) (Marston, 1989; McMaster & Epsin, 2007). Reliability estimates were determined by analyzing interrater agreement, test-retest, alternate form, and internal consistency. Interrater agreement, or interscorer reliability, was reported in most studies and found to be generally strong (Deno et al., 1982; Marston & Deno, 1981; Marston et al., 1983; Marston, Lowry, Deno, & Mirkin, 1981; Tindal, Marston, & Deno, 1983; Videen et al., 1982). The mean agreement for all three measures was .98 (Marston, 1989).

Test-retest reliability for WW and CLS written in 5 minutes had relatively strong correlations when assessed over a 1-day interval (r = .91 for WW, .81 for WSC, and .92 for CLS) but moderate over a 3-week interval (r = .64 for WW, .62 for WSC, and .70 for CLS; Marston et al., 1981; McMaster et al., 2007). Student longitudinal growth was also studied (Deno, Marston, Mirkin, Lowry et al., 1982; Marston et al., 1981; Tindal et al., 1983). Deno et al. (1982) termed this "growth stability." They looked at the reliability from fall to spring for first through sixth graders. Coefficients for first graders (r = .20-.47) and third graders (r = .37) were weak, while moderate to strong for second- through sixth-graders (r = .60-.86). In another study, fall to spring coefficients for fifth graders was r = .56 for both WW and CLS (McMaster et al., 2007; Tindal et al., 1983).

The IRLD researchers also assessed the reliability of administering alternate forms. Reliability between two 5-minute story prompts was strong for CLS (r = .96), WSC (r = .95), and WW (r = .95; Marston et al., 1981). Tindal et al. (1983) used a 3-minute writing sample with a story prompt. They obtained moderate to strong coefficients (r = .73 for WW, .72 for WSC, and .93 for CLS; McMaster et al., 2007). A

3-minute writing sample was also assessed by Shinn, Yesseldyke, Deno, and Tindal (1982) with children identified as learning disabled or low achievers. They found weaker coefficients for WW (r = .51 - .71). When general education fourth- and fifth-graders were assessed using a 3-minute writing sample, moderate correlations were found for WW (r = .71) and CLS (number of letters, r = .70).

Aggregating scores across days has been found to increase reliability (Fuchs, Deno, & Marston, 1983). Specifically, students that were considered low achievers were assessed with a 3-minute writing sample weekly for 10 weeks. Correlations were then calculated between scores on adjacent measures (week 1 and 2), across 4 sessions (mean of weeks 1 and 3 compared to mean of 2 and 4), and across 6, 8, and 10 sessions. Results revealed stronger reliabilities with aggregations across more days(r = .55 across 2 days, .72 across 4 days, .85 across 6 days, .88 across 8 days, and .89 across 10 days; Fuchs et al., 1983; McMaster et al., 2007).

Findings from the IRLD studies that laid the groundwork for future researchers in the area of CBM written expression. Promising results revealed moderate to strong criterion validity coefficients for countable indices of writing, such as WW, WSC, and mature words. Coefficients were strongest between these indices and the TOWL and DSS (r = .67 - .88). IRLD studies also indicated that valid measures of written expression could be obtained from brief, 3-5 minute writing samples and relatively efficient, quantitative scoring procedures (McMaster et al., 2007).

The IRLD studies also demonstrated inconsistent results for the reliability of the measures (McMaster et al., 2007). While interscorer reliability had consistently strong coefficients (r > .90 for most measures), alternate form reliability results were

inconclusive. Reliability was lower within grade level (Tindal et al., 1983), and for students with LD and low achievers (Marston et al., 1981; Shinn et al., 1982). Aggregation of scores across writing samples collected over time (weekly) revealed an improvement in reliability coefficients (from r = .55 to .89; Fuchs et al., 1982). However, when using CBM to identify students at risk or evaluate a student's progress, it is crucial to make efficient, timely decisions. Waiting weeks or months to obtain reliable information is contradictory to one of the key goals of using CBM data. Results from studies assessing sensitivity to growth were somewhat limited (Deno et al., 1980; Deno et al, 1982; Marston et al., 1981; Shinn et al. 1982; Tindal et al., 1983). Growth was examined across grades or from fall to spring. Growth for WW, CLS, and WSC was found from first to sixth grade, within grade across 10 weeks and from fall to spring. The authors noted that this growth was not dramatic but was evident, whereas growth on the standardized measure (Language subtest of the SAT) was absent (Marston et al., 1983). Growth was not examined for monitoring progress on a weekly basis (McMaster et al., 2007).

Tindal and Parker (1991) addressed the reliability and validity of written expression measures across student populations and grade levels for qualitative and quantitative measures. They describe four criteria that must be met for any writing assessment despite its purpose (program evaluation, progress monitoring, screening-eligibility, or accountability). The four minimal measurement criteria included: (a) consistent administration and reliable scoring; (b) discrimination among students at different skill levels; (c) demonstrate at least low-moderate relation to other accepted assessment methods; and (d) exhibit student score improvement over the course of a year.

Participants included students from an elementary school, in grades three through five, at various skill levels (Tindal et al., 1991). Skill levels included those receiving specialized services (learning disabled and Chapter 1 students) versus those not receiving special services (low regular education students and other regular education students; Tindal et al., 1991). Students were administered a writing task in the fall (November) and spring (May). Quantifiable indices of writing included WW, WSC and words sequenced correctly. Fifth grader writing samples were also scored for exploratory measures: number of incorrect word sequences, total number of word sequences, and percentage of correctly sequenced words. The Stanford Achievement Test (SAT; Gardner, Rudman, Karlsen, & Merwin, 1982) was administered in April; thus, scores from the language portion (spelling, language, and subtotal) and a reading subtotal were included for the fifth graders. Three subjective measures were included: (a) story idea, (b) organization-cohesion, and (c) conventions-mechanics.

Interscorer reliability ranged from .92 to .99 for countable indices (WW, WSC, and words in correct sequence) and .73 to .88 for the qualitative measures (Tindal et al., 1991). Significant differences were found for all measures between the four student groups. Post hoc comparisons revealed significant differences for all measures between general education and learning disabled students; on most measures between general education and Chapter 1; and on some measures between general and low general education. Quantitative measures were not highly correlated with any of the SAT tests. Correlations between quantitative and qualitative measures were weak to moderate, which contrasts with results from the IRLD studies (McMaster et al., 2007). A factor analysis revealed a three factor model accounting for 81% of the variance. The first

factor accounted for 37% of the variance and consisted of four production variables: number of words written, correctly spelled, correctly sequenced, and total number of sequenced words. The other factors consisted of (Factor 2) the percentage of correctly sequenced words, subjective judgment of conventions-mechanics, (Factor 3) subjective judgment of story idea, and organization-cohesion. Growth for all students was observed for the number of words written, words spelled correctly, and words in correct sequence. An analysis of individual student growth indicated that some students improved quantitatively and qualitatively, while the other students improved qualitatively but not quantitatively. Tindal et al. (1991) concluded that significant differences were found within grade levels with different student groups. However, in terms of program evaluation, a multi-faceted effort may be necessary, with focus on determining what to measure, how to change performance, and how to document changed performance. In support of this, Tindal and Hasbrouck (1991) noted that quantitative scoring procedures provide useful progress monitoring information, while qualitative scoring may be used for diagnostic and instructional decisions (McMaster et al., 2007).

While studies discussed thus far have provided valuable information on countable writing indices, including their ability to represent various levels of performance across student populations and grade levels, they have not addressed the selection of cut off scores for screening-eligibility decisions. Parker, Tindal, and Hasbrouck (1991a) examined the suitability of five quantitative or countable indices of writing quality for decisions involving screening-eligibility, focusing on sensitivity in the low cutoff score range. The five countable indices included total words written, correctly spelled words, correct word sequences, percentage of correctly spelled words, and percentage of correct

word sequences. Teachers' holistic ratings on the communicative effectiveness of the writings were also used as a criterion variable. Data were collected in the fall and spring for 1,917 students in grades 2 through 5. Students in grades 6, 8, and 11 (n = 243) were assessed in the spring.

Consistent with previous studies, writing skills, as measured by the countable indices, increased over grades (Parker et al., 1991a). From fall to spring, mean scores increased for each countable index for grades 2 through 5. Also, scores generally increased from fall to fall and spring to spring (across grade levels), except for total words, correctly spelled words, and correct word sequences from fourth to fifth grade. For grades 2 through 5, fall scores were plotted on histograms with a superimposed normal curve derived from the distribution mean and standard deviation. They also used percentile line graphs and standard error of measurement bands to aid in identification of the best-suited measures for screening-eligibility decisions. In conclusion, the authors recommended use of percentage of correctly spelled words, for all grades, due to its reasonable validity scores with teacher holistic ratings and display of suitable distribution in the lower score range. If grade 2 is excluded, the percentage of correct word sequences was recommended as an alternative scoring method. However, the authors describe these findings as offering barely sufficient validity and measurement sensitivity to serve as gross measures for written expression. Rate-based indices (correct word sequences and correctly spelled words) were not recommended based on low reliability and unacceptable distribution. Total words written was also not recommended due to its low validity coefficients with teacher ratings and marginal distribution properties.

In addition to utilization for screening-eligibility purposes, CBM for written expression should be able to provide information through progress monitoring. Thus, the measure should be sensitive to small increments of growth over relatively short periods of time (Tindal, 1989). This information is valuable for formative adjustment of instruction (Moran, 1987). Therefore, the measure needs to be technically adequate. Parker, Tindal, and Hasbrouck (1991b) investigated seven objective indices of writing quality of 36 middle school students with mild disabilities. They sought to establish the technical adequacy of these measures by collecting writing assessments four times over a 6 month period, and utilizing holistic measures and the Test of Written Language (TOWL; Hammill & Larsen, 1983) as criterion comparisons. The seven measures of written expression include four production dependent and three production independent measures (Rafoth & Rubin, 1984). The production dependent measures include the following: (1) total number of words written (TWW), (2) number of correctly spelled words(CSW) , (3) correct word sequences (CWS), and (4) number of letter groupings recognizable as real English (Leg. Wd.). These measures partly depend on the length of the writing sample, unlike production-independent measures (Parker et al., 1991). The three remaining indices are considered production-independent: (5) average length of all continuous strings of CWS, (6) percentage of total words written that are legible as English, and (7) percent of words that are correctly spelled.

Results revealed that only TWW and number of legible words written appeared to increase regularly over the six months (Parker et al., 1991b). However, the strongest predictors of holistic ratings were percent of legible words, CWS, and mean length of CWS. These correlations were moderately strong (Parker, et al., 1991b). The weakest

predictor of holistic ratings was TWW. In regards to performance on the TOWL, of the six subtests, Word Use was correlated with all direct scoring indices and holistic ratings. Thematic Maturity and Spelling were also significantly correlated to percent of legible words, CWS, mean length of CWS, and holistic ratings. The remaining three subtests (vocabulary, handwriting, and style) were not significantly related to any direct measure. Stability estimates between adjacent assessment time periods were at least moderate size and uniform for TWW, CSW, CWS, and legible words. Also, pronounced linear growth was noted for TWW, number of legible words, and CSW. The authors concluded that even though these measures' stability estimates and linear trends suggest "sensitivity to growth," a key element for progress monitoring, that conclusion was not supported by the lack of corroboration with criterion measures and informal observations by the research team. Thus, none of the measures appeared sufficiently valid for the purpose of measuring skill growth for writing-deficient middle school students when using holistic judgments as the criterion (Parker, et al., 1991b).

In a study designed to extend the research on measuring written expression with middle school students, researchers assessed the utility of correct minus incorrect word sequences (CWS-IWS); along with, words written, words correct, words incorrect, characters, sentences, characters/words, words/sentences, CWS, incorrect word sequence, and mean length of correct word sequence (Espin, Shin, Deno, Skare, Robinson, & Benner, 2000). They compared the writing indices to teacher ratings and a district writing test taken by the eighth grade sub-sample. For 6^{th}, 7^{th}, and 8^{th} graders' writing samples, correlations between the potential indicators and holistic ratings were generally moderate, except for CWS-IWS, which had moderately strong correlations. Correlations

between predictors and the 8th graders' scores on the district writing test were slightly larger than for the teachers' ratings. Also, moderately strong correlations were not only found for CWS-IWS, but included two sentence measures: sentences and words per sentence. Espin et al. (2000) also sought to determine if a combination of predictors would predict student performance better than a single measure through regression analyses. Teacher holistic ratings were used as the criterion variable. Predictors selected for the analysis included CWS-IWS, characters, and sentences. Only CWS-IWS entered the equation (*R* values ranged from .65 - .69). With these findings, the authors concluded that CWS-IWS may serve as the best indicator of student's general writing proficiency for secondary school students.

While CWS-IWS appeared to be a potentially promising measure of writing for middle school students, there was not enough literature to claim that all potential variables had been studied. Recently, researchers were unsatisfied with the restrictive range of writing indices that had been studied for elementary students (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002). Identifying variables that are important for screening and progress monitoring in the elementary grades is crucial for providing additional support and practice when students writing skills are emerging. Gansle et al. (2002) sought to determine whether additional variables would share more variance with criterion measures than correct words written or correct word sequences. Their study included several variables that had been previously studied: total words written, long (large) words, words spelled correctly, and words in correct sequences. In addition, they also included the following variables for exploratory purposes: number of nouns, number of verbs, number of adjectives, total punctuation marks, correct capital letters, complete

sentences, sentence fragments, and simple sentences. Variables easily scored by Microsoft Word were also included, as these variables would be easily available if found to be reliable and valid. Those variables included Flesch reading ease (Word), Flesch-Kincaid grade level (Word), Flesch-Kincaid grade level (WordPerfect), Sentence Complexity (WordPerfect), and Vocabulary Complexity (WordPerfect). The study included 179 third and fourth graders. They were assessed on two consecutive days with a 3-minute writing probe. The criterion variables included teacher rankings of students according to their writing skill and performance on standardized assessments. For third graders, the language-related subtests on the Iowa Test of Basic Skills (ITBS) was used as the criterion variable. Fourth graders take the Louisiana Educational Assessment Program (LEAP), which students must pass for promotion to the fifth grade. Write competently and use conventions of language, the writing relevant scales of the LEAP, are scored based a writing sample and multiple choice questions (Gansle et al., 2002).

Interscorer agreement for all but four variables met the 80% criterion (Gansle et al., 2002.) The four variables that did not meet this criterion included simple sentences, sentence fragments, words in complete sentences, and complete sentences (range from .70-.76). Alternate form reliability was calculated for all variables between the two probe administrations. Positive correlations for hand-scored variables ranged between .006 (long words) to .62 (total words). Correlations for computer-scored variables ranged between .09 (vocabulary complexity) to .55 (Flesch-Kincaid grade level, WordPerfect).

A series of stepwise, forward, multiple-regression analyses were conducted between all probe variables and criterion variables (Gansle et al., 2002). For third graders, scores from the ITBS were used. Variables that entered the regression equation

for the language usage subscale included the words in correct sequence, number of correct punctuation marks, and number of verbs. Total number of verbs had a negative relation, whereas the other two were positively related. The variables that entered the regression equation for the language total scale included correct punctuation marks, words in correct sequence, words written, and long words. Words written was negatively related to the ITBS language total subscale while the others had a positive relationship.

For fourth graders, scores from the LEAP were used as the criterion variable (Gansle et al., 2002). Variables entering the regression equation for the write competently subscale included number of verbs and the vocabulary complexity score from WordPerfect. The regression equation for the use of conventions of language subscale included the following variables: words in correct sequence, total words written, and nouns. Once again, total words written had a negative relationship while all other variables were positively related.

A regression analysis was also conducted for all third and fourth grade probe variables with teacher rankings of students' writing skills (Gansle et al., 2002). The variables found to be significantly related to teacher rank included correct word sequence, total words written, and correct punctuation marks. Total words written was negatively related. This relationship is a function of the variable being entered into the equation after correct word sequences. For all of the regression equations reported, total words written may have been functioning as a suppressor variable in that it has low correlations with the criterion variables used in the study (correlations ranging from .08 to .28). However, the intercorrelations with other predictors were not reported. The authors concluded that correct punctuation marks shows the most promise as an

additional index for writing skill. Results from this study also added validity for using words in correct sequence as an indicator. The results from total number of words written suggest that it is not as good an indicator as other variables.

In a follow-up study, the technical characteristics of total words written, words spelled correctly, correct punctuation, correct capitalization, complete sentences, words in complete sentence, and correct word sequences were analyzed for a larger group of students across grade levels (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). They also sought to establish the test-retest reliability, interrater reliability, and criterion-related validity of a commercially published assessment system that utilizes teacher ratings of writing quality, the Six Trait model (Northwest Regional Educational Laboratory, 2000). The criterion-related validity was assessed by investigating the measures' relationship with the *Stanford Achievement Test, Ninth Edition* (Stanford 9; Harcourt Brace Educational Measurement, 1996). Data reported in the study included 206 CBM student writing samples with 190 retest samples; 214 writing samples analyzed for the Six Trait model with 201 test-retest evaluations; and 169 Stanford-9 test results for second through 5th graders. The follow-up, retest samples were collected 1 week after the first administration.

Interrater agreement for CBM variables ranged between 81.8 to 97.7%, while exact agreement for trait measures ranged between 53.2 to 58.7% (Gansle et al., 2006). Test-retest reliability for indices of written expression, using Pearson correlations, resulted in a range between .44 for correct capitals and .82 for words spelled correctly. Test-retest reliability for Six Trait writing sample scores had a percent exact agreement range from 35 to 40% and the correlation ranged from .06 - .25. These results suggest

unacceptably low levels of temporal stability over a short interval (Gansle et al., 2006). Intercorrelations were calculated between CBM writing variables and the Stanford-9 standard scores. Intercorrelations among CBM indices resulted in two apparent clustering of scores. Total words written, words spelled correctly, and correct word sequences were intercorrelated at $r > .92$. Words in complete sentences, correct punctuation, and complete sentences were intercorrelated at $r > .77$. Validity coefficients with the total language scale of the Staford 9 were similar for all CBM indices and generally moderate. Of the intercorrelations between the Six Trait measures and the Stanford 9, the strongest relationships existed between the Six Trait measures themselves. Cronbach's alpha revealed correlations between .78 to .95, which suggests that the items contribute to a single dimension. This appears to contradict the publisher's assertions that it assesses distinct dimensions (Gansle et al., 2006).

In conclusion, the authors noted that this study demonstrated that some collections of CBMs may provide unique variance and maintain technical adequacy (Gansle et al., 2006). Thus, clusters of indices exist that are intercorrelated and may represent an element or category of writing. Also, in their previous study, Gansle et al. (2002) noted that a combination of variables to form new indices may contribute to the prediction of criterion variables. They specifically recommended the utilization of total words minus words in complete sentences to create a measure for number of error words. This measure may be similar to those studied by Tindal and Parker (1989) and Parker et al. (1991b), where ratios and percentages were included in their analyses as indices of writing skill (Gansle et al., 2002).

Jewell and Malecki (2005) extended the research on percentage measures to elementary students and utilized clusters of indices to create composite scores. They wanted to assess the validity of three categories of CBM indices for written expression: production-dependent, production-independent, and accurate-production indicators. Production-dependent indices included Total Words Written (TWW), Words Spelled Correctly (WSC), and Correct Writing Sequences (CWS). Production-independent indices used in this study included percentage of Words Spelled Correctly and percentage of Correct Writing Sequences. Correct Minus Incorrect Writing Sequences (CMIWS) was the accurate-production indicator. Composite scores were also calculated for each category that contained multiple indices. The production-dependent indices were average to produce the Production-Dependent Composite (PDC). The Production-Independent Composite (PIC) was the average of the production-independent indices. They utilized the Tindal and Hasbrouck (1991) Analytic Scoring System (THASS), the Stanford Achievement Test (SAT; Harcourt Brace Educational Measurement, 1997), and students' Language Arts classroom grades as the criterion comparisons. Participants included 203 second-grade (n = 87), fourth-grade (n = 59), and sixth-grade (n = 57) students.

A three-minute writing sample was collected from each student with the standardized assessment subsequently administered within six days (Jewell et al., 2005). The fall Language Arts grade was used for analyses (this grade represents the time during which CBM samples were collected). MANOVAs were run to assess grade-level and gender differences. Correlations were computed between all variables; while a regression analysis was conducted using the composite scores (PDC and PIC), CMIWS, and grade level as predictors of students' scores on the THASS.

In regards to gender, girls outperformed boys on all fluency measures but not on production-independent or accurate-production indices (Jewell et al., 2005). Girls typically wrote more, produced more correct spelling, and correct word sequences. Across second, fourth, and sixth grade levels, scores were significantly different for all writing indices; except fourth and sixth graders not having significantly different production-independent indices. Results from the correlational analyses, conducted to assess the interrelationships among the writing indices, revealed that TWW was not related to production-independent measures at any grade level. However, the CMIWS scores were correlated with both production-dependent and production-independent indices. Correlations between CBM indices, composites, and criterion measures were also analyzed for significance. Grade level differences were detected in how the CBM indices related to the criterion measures. Fewer CBM scoring indices were related to the criterion measures as grade-level increased. Most production-dependent, production-independent, and accuracy production measures were significantly related for elementary aged students. However, for sixth –grade score, only CWS, production-independent, and accuracy measures were related (not TWW or WSC). Correlations were moderate to strong and ranged from .43 to .67 for SAT scores and from .34 to .56 for the THASS scores. Regression analyses revealed that the four predictor variables accounted for 53% of the variance for the total THASS scores. CMIWS was a significant predictor for both the THASS Total and convention-mechanics scores. Students' grade level was also a significant predictor for conventions-mechanics (Jewell et al., 2005).

In conclusion, it appears that production-independent and accuracy measures had consistently stronger relationships with the criterion measures (SAT, student grades,

THASS) than production-dependent measures (Jewell et al., 2005). All three categories were typically related to student performance in elementary grades while only production-independent and accuracy measures were significantly related at older grades. Correct minus incorrect writing sequences had reliability and validity evidence that suggests its use as an appropriate measure of students' overall writing proficiency.

Summary of Research Findings and Rationale for the Current Study

Curriculum-based measurement (CBM) was originally primarily viewed as a progress monitoring tool (Wayman et al., 2007). The premise of CBM was that it met the characteristics considered essential features desirable for monitoring student progress (Jenkins et al., 1979). These characteristics included the following: (1) the measure is tied to the curricula, (2) can be administered frequently and is time efficient, (3) has multiple forms, (4) inexpensive, (5) sensitive to student improvement, and (6) can be measured reliably and validly (Marston, 1989). Results from the IRLD studies on CBM for reading revealed that oral reading fluency was a valid indicator for overall reading ability and could be measured reliably (Marston, 1989). Subsequent research supported oral reading as a general measure of reading proficiency and a better indicator of reading comprehension than traditional measures (Waymen et al., 2007). Given the established validity and reliability of oral reading fluency, researchers have begun to investigate the utility of oral reading for screening purposes. Specifically, researchers focused on the diagnostic efficiency of oral reading fluency for predicting passing or failing status on state mandated tests (Good et al., 2001; Hintze et al., 2005; McGlinchey et al., 2004; Stage et al., 2001). These studies found that cut scores can be established that are able to predict performance on high stakes test at a level higher than base rates of performance

and these cut scores can be established for short and longer periods of time, as far back as first grade.

Research on CBM for written expression is more complex. Initially researchers for the IRLD studies found that total words written, words spelled correctly, correct letter sequence, and correct word sequence could be measured reliably, had moderate correlations with criterion measures, and were more appropriate than standardized tests for monitoring progress (Deno et al., 1982; Fuchs et al., 1982; Marston et al., 1981; Marston et al, 1983; Shinn et al., 1982; Tindal et al., 1983; Videen et al., 1982). These studies also laid the groundwork for establishing the types of procedures that could be used for collecting brief writing samples. However, questions still existed regarding the reliability of these measures for different skill levels, whether these measures could be used for screening, and the limited number of indices that had been studied as measures of written expression (McMaster et al., 2007).

Researchers continued to assess the validity and reliability of the measures addressed in the IRLD studies while also expanding the research. Results from studies revealed that groups differed when comparing students receiving special education services and those not receiving services (Tindal et al., 1991). In regards to the utility of writing indices for screening purposes, percentage of words spelled correctly was found to useful for grades 2 through 5, while percent correct word sequences was also found to be useful for grades 3 through 5 (Parker et al., 1991a).

Unsatisfied with the range of indices studied, researchers have investigated the technical adequacy for new measures of written expression. Correct minus incorrect word sequences, correct punctuation marks, and the use of composite scores have been

added to the list of measures that appear to be related to criterion measures and may provide unique information when assessing writing across grade levels (Jewell et al., 2005).

The ability of these measures to guide educators' decision making has yet to be established. The variety of purposes for assessment (screening, monitoring progress, eligibility-determination) is directly related to the types of decisions educators will need to make. Before utilizing these indices as a basis for decisions, it must be established that these measures are related to variables or measures already considered important and reflective of a student's general academic abilities. Measures that are utilized on a consistent basis for all students include district administered tests. These tests are typically administered and used to provide parents and teachers with information on students' abilities; however, results are not typically used to guide instructional decisions, but may be used as evidence for grade level placement or special education eligibility. While research on CBM oral reading fluency has produced numerous studies addressing the concurrent and predictive validity of this measure with performance on state tests, only two studies have assessed the validity of written expression CBMs for elementary students with state test performance (Gansle et al., 2002; Weissenburger & Espin, 2005). However, neither of these studies assessed the predictive and concurrent validity of indices of written expression over multiple data collection periods administered throughout the school year. This type of data collection method is currently being used by school systems. Specifically, schools are collecting CBM data three times throughout the school year to help identify students at-risk or struggling in order to help guide timely instructional decisions.

The current investigation will examine the degree to which it is possible to obtain more useful information from the writing samples than current practices yield and their relationship to major standardized achievement tests. Writing samples will be collected throughout the school year, scored, and assessed for the predictive and concurrent validity using the state standardized achievement test as the criterion measure. In line with previous research, the concurrent and predictive validity of oral reading fluency CBM will also be assessed. As of yet, no investigation has assessed simultaneously the validity of reading and writing CBMs with standardized state assessments.

The utility of indices of written expression found to be promising in previous studies will be assessed through a series statistical analyses. The design of this current investigation aims to answer the following research questions.

1. What is the relationship of CBM oral reading fluency and writing measures to reading and writing outcomes respectively?

2. Which curriculum-based measures of written expression will serve as the best predictors of the relevant criterion variables?

3. Can oral reading fluency be combined with existing or recently developed writing measures to yield a more accurate prediction of passing or failing a state standardized assessment?

4. Are published cut-scores for ORF, TWW, CWS, and WSC effective for predicting the passing and failing status of students above base-rate levels of performance?

METHOD

Participants

Data were collected from 202 third grade students. Due to the nature of the current study, it was necessary to only retain participants who had data available for all relevant variables Therefore, data from 124 third grade students was used for this investigation. Of those students whose data was not used (n = 78), half were missing two or more data points. In order to determine the appropriate sample size for this study, a power analysis was conducted utilizing methodology from Kraemer and Thiemann (1987). The necessary sample size for power of .80, to detect a medium effect size at alpha = .01 for a two-tailed test, ranges from 90 to 106 for the analyses conducted.

Students were from two schools within a predominantly rural school district located in the southern region of the United States. The school district serves approximately 22,000 students. In participating schools, 83% of students are Caucasian, 14% African American, 3% Hispanic, and .1% Asian. Sixty-four percent of the students receive free or reduced lunch. Male students make up 51% of the population, while 49% are females. The current investigation was composed of 43% male students (n = 53) and 50% female students (n = 62). Descriptive data was not available for 9 participants. Participants ranged in age from 7 (n = 7) to 10 (n = 2); however, the majority of students were aged 8 (n = 78) or 9 (n = 28). The study was conducted with an approval from the institutional review board and permission from the district. Data was coded by the primary investigator to maintain confidentiality of all participants.

Measures

integrated Louisiana Educational Assessment Program (*i*LEAP). The *i*LEAP is a criterion-referenced test administered to students in third, fifth, sixth, seventh, and ninth grade in the state of Louisiana. Development of the test items for the English Language Arts (ELA) and math tests of the *i*LEAP were from two sources. First, the *Iowa Tests of Basic Skills* (ITBS) was utilized for comparing its alignment against state content standards. Furthermore, in order to develop an assessment framework that would meet state performance standards, gaps in coverage were identified between the Iowa test items and Louisiana's grade-level expectations (GLEs). Additional (augmented) items were developed for GLEs not represented in the ITBS. Therefore, the norm-referenced test items developed from IOWA tests combine with the augmented (criterion-referenced) GLE-based items to form the criterion-referenced test (CRT) component of the *i*LEAP. The *i*LEAP assigns students into one of five categories based on their performance: unsatisfactory, approaching basic, basic, mastery, and advanced. The ELA component was the criterion measure of interest for the purpose of this study. The ELA test consists of four subtests, reading, language, writing, and using information resources. The subtests include 78 multiple-choice items and a writing prompt. The reliability of the third grade ELA test is adequate, with Cronbach's alpha and the Stratified alpha both revealing a .93 reliability coefficient.

Curriculum-Based Measures. DIBELS(University of Oregon, 2003) was used as the curriculum-based measure for oral reading fluency. Reading passages were developed to reflect the curriculum for a particular grade level. Each grade level has benchmark passages that are administered during the fall, winter, and spring. At each

benchmark, three separate passages are administered. Each passage yields an oral reading fluency score, which is the number of words read correctly in 1-minute. The median oral reading fluency score is recorded as the score for the benchmark period. A number of studies have established the reliability and validity of DIBELS as a measure of oral reading fluency (Good, Gruba, & Kaminski, 2001; Good, Simmons, et al., 2001; Kaminski & Good; 1998; University of Oregon, 2003). The test-retest reliability ranges from .92-.97 for the oral reading fluency measure. The alternate form reliability has a range from .89 to .94. Criterion-related validity ranges from .60 to .90 with comprehension tests (Wood, 2006).

A variety of dimensions were scored from three-minute writing samples. These measures were used to identify variables that have a strong relationship with the criterion measure. Reliability and validity of these measures was assessed to determine which measures have acceptable technical adequacy. The following measures fall into one of three types of methods of measurement: production-dependent indices, production-independent indices, or an accurate production indicator (Malecki et al., 2003).

- Production-dependent indices

1. Total words written (TWW). Total number of words written, including misspelled words, with in the 3-minutes was counted and recorded. This included any word-like string of letters with a space before and after (Shinn, 1989).
2. Words spelled correctly (WSC). English words spelled correctly within a low-inference judgment regarding appropriateness of context were counted as correctly spelled (Powell-Smith & Shinn, 2004).

3. Correct word sequences (CWS). Two correctly spelled words joined together to make a (1) mechanically, (2) semantically, and (3) syntactically correct sequence will count as a correct word sequence (Powell-Smith et al., 2004).
4. Correct punctuation marks. Punctuation marks that are correctly applied were counted and recorded. Each correctly used quotation mark counted as one (Gansle et al., 2006).
5. Complete sentences. Sentences were scored as complete if they had a capitol letter, a subject or understood subject, a verb, and an ending punctuation mark. Run-on sentences were not counted as a complete sentence (Gansle et al., 2006).
6. Words in complete sentence. Sentences scored as complete based on the previous definition were scored for number of words in the sentence (Gansle et al., 2002).

- Production-independent indices.

7. Percentage of words spelled correctly (%WSC). The number of words counted as spelled correctly (WSC) will be divided by the total number of words written (TWW) and multiplied by 100 (Malecki et al., 2003)..
8. Percentage of Correct word sequences (%CWS). The number of correct word sequences (CWS) was divided by the total possible word sequences and multiplied by 100 to calculate the percentage of correct word sequences.

- Accurate-production indicator.

9. Correct minus Incorrect word sequences (CMIWS). The number of incorrect word sequences will be subtracted from the number of correct word sequences (CWS).

Time Variables. Time data was collected for the length of time to score the measures of written expression. Fifty-percent of writing samples at each benchmarking period were randomly selected for timing. Reliability for timing of variables was conducted for thirty percent of these probes. Timing data were also collected from the secondary scorer to account for differences between scores on timing variables. The secondary scorer collected time data for thirty percent of the probes selected for timing. For timed samples, each dimension of written expression was timed. Scorers recorded the amount of time it took them to score each variable for each sample.

Score Reliability. Inter-rater reliability was calculated for all dimensions scored for the writing samples. Thirty percent of writing samples at each benchmark were randomly selected for calculating inter-rater reliability for each variable. Agreement was calculated using the procedure for rate-based measures outlined by Cooper, Heron, and Heward (1987). The smaller number was divided by the larger number and multiplied times 100.

Procedures

Probe Administration. The CBM data were collected in the two schools as part of an initiative by the local education agency to conduct universal screening data in a range of academic areas during the 2007 - 2008 school year. Data were collected three times throughout the school year for each CBM measure. Specifically, the bench marking samples were collected in the fall, winter, and spring. The winter and spring data collection for both oral reading fluency and written expression CBM measures were conducted with in one-week of each other. The fall data collection for oral reading fluency occurred approximately two months prior to the written expression data

collection. State standardized assessment data (*i*LEAP) were collected during the second week of March, approximately one month prior to the spring benchmark of oral reading fluency and written expression measures.

The CBM-R data was collected by teachers who were formally trained to collect DIBELS oral reading fluency. This measure was administered individually. Written expression data were collected by a team of district personnel who received formal training on collecting curriculum-based measurement data. Standardized procedures with written instructions were provided for data collection. A whole-class format was used to collect this data. A 3-minute writing sample was collected at each benchmark period. Participants were given a lined piece of paper with a story starter typed at the top. A different story starter was used for each benchmark. The fall story starter was, "On the way to school this morning…." The story starter used in the winter was, "When I went to the zoo….." The spring story starter was, "I was sitting by my friend at lunch when….." A timer was used for 1-minute and 3-minute timing periods. Participants were allowed 1-minute to think about the story starter and 3-minutes to write. During the think time, students were instructed to leave their pencil on their desk. At the end of the 3-minute writing period, students were instructed to place their pencil on their desk and the writing samples were collected.

Scorer Training. The primary investigator and an additional scorer for reliability analysis reviewed the definitions of the target variables for scoring the writing samples. They scored a sample of 10 probes together (sample probes utilized by Gansle et al., 2002). Definitions were clarified and 10 additional sample probes were scored independently. Two additional references were used to clarify definitions of variables,

the *AIMSWEB* training workbook for scoring written expression measures (Powell-Smith et al., 2004) and University of Minnesota procedures for scoring writing samples (http://www.progressmonitoring.net/pdf/RIPM_Writng_Scoring.pdf). The *AIMSWEB* training workbook was used as a primary reference for clarification, while the RIPM rules were used as a secondary source when further clarification was necessary. Scorers utilized *The Great American English Handbook* (1987) and the Merriam-Webster online dictionary (http://www.merriam-webster.com) to address questions related to grammar and spelling. After scoring the 10 practice probes, scorers compared agreements and disagreements. A final sample of 24 probes were independently scored and assessed for reliability. A minimum of 85% mean agreement was achieved for these additional probes for 8 out of the 9 indices. Reliability for correct minus incorrect sequences only achieved a level of 80% mean agreement. Additional probes were assessed independently targeting calculation of correct minus incorrect sequences. Thus, a mean reliability of 85% for correct minus incorrect sequences was achieved after assessment of 30 probes.

Design

This study evaluated predictor-criterion relationships of curriculum-based measures and standardized group tests for the purpose of identifying useful measures of written expression. Curriculum-based measures included oral reading fluency and several measures of writing competence. The design was descriptive in nature.

Analyses

An examination of the relationship between student oral reading fluency (ORF), measures of written expression, and criterion variables was conducted utilizing the following analyses. First, student ORF and written expression data at each benchmarking

period (fall, winter, spring) served as predictors of the relevant criterion variables: student *i*LEAP total ELA score and ELA subtest percentage correct scores for reading, writing, and language. In order to assess the degree to which the CBM variables correlate with performance on the *i*LEAP, Pearson Correlations were calculated for each benchmark period. Correlations between two variables reflect the degree to which the variables are related. Pearson correlations reflect the degree of liner relationship. Correlations were calculated between the spring benchmark scores for CBM variables and the ELA scores to determine concurrent-criterion validities. Correlations were also calculated between the criterion variables and CBM variables from fall and winter data collections to assess the predictor-criterion relationship.

Due to the degree of intercorrelations between the variables, it was necessary to use a variable reduction strategy before conducting the regression analyses. Therefore, a factor analysis was conducted using principle components analysis (PCA) as the factor extraction method. This method forms linear combinations of the observed variables (Norušis, 2008). Clusters of highly correlated measures were identified. Principle components analysis was used to identify component variables and best exemplar variables necessary for conducting the multiple regressions. The best exemplars were chosen based on their correlations in the structure matrices. The best exemplar strategy was examined because of its relationship to practice in which a single measure is more likely to be preferred and used for repeated CBM administration.

To assess the concurrent and predictive relationship between independent variables and criterion test scores, multiple-regression analyses were conducted. The independent variables included the curriculum-based measures for writing from the fall, winter, and

spring data collections (only writing measures identified in the PCA). The criterion variables included the *i*LEAP language subtest, writing subtest, and total ELA score. Multiple regression is used when the interest is in predicting a criterion with a set of predictors. Data for the independent variables was entered into SPSS using blocks. This method is referred to as a hierarchical multiple regression. This type of regression forces the analysis to look at the first block of variables first; each additional block is evaluated after controlling for the impact of the previously considered blocks. This type of regression allowed for variables to enter the analysis in the sequence they were available through the school year, which would be important information when considering application of the findings. Fall variables were entered into block one, winter variables were entered into block 2, and spring variables were entered into block three. Each block was assessed using the stepwise, forward method. Stepwise regression was used because it is typically used in the exploratory phase of research. In a stepwise multiple regression analysis, statistical criteria are used to decide the number of predictors to be selected and the order of entry (Stevens, 2002). The stepwise method is similar to forward solution, in which predictor variables are entered one at a time contingent on the strength or their correlation with the criterion variable. Stepwise differs in that a variable previously selected may be deleted if it loses its effectiveness as a predictor when considered in combination with newly entered predictors (Hinkle, Wiersma, & Jurs, 1998). One problem sited by Thompson (1995) that should be considered when using stepwise method is it's capitalization on sampling error or chance. This problem may occur when variable selection is affected by sampling error. Specifically, the advantage of one variable over another may be due to sampling error and thus affect the remaining

conditional entry decisions. Thus, the order of predictors may not be entirely accurate which affects the generalizability of the results. This issue is most likely to be severe when a number of variables are highly similar in the strength of their relationship to the criterion. That is small dichotomous distinctions are being made.

Binary logistic regression analysis was used to discover the predictors of success or failure on the *i*LEAP using CBM measures of reading and writing. Logistic regression is a model used for the prediction of the probability of occurrence of an event. Correctly predicting the category of outcome for individual cases using the most parsimonious model is the goal of logistic regression. For this analysis, *i*LEAP data were dichotomized so that a score of 282 on the ELA section (scoring at the "basic" level) or higher was considered passing. This variable functioned as the dependent variable. The independent variables included CBM data for oral reading fluency and CBM measures of writing that were found to be related to the criterion measure through multiple regressions. Two regressions were run using best exemplar variables and component variables. Independent variables were entered using the blocks, as was done with the multiple regression. Block one contained the fall variables, winter variables in block two, and spring variables in block three.

A conditional probability analysis was used to asses the diagnostic accuracy of CBM measures for predicting the success and failure on the *i*LEAP. This analysis was conducted utilizing the CBM measures for which published cut-scores or normative data exist (ORF, TWW, CWS, and WSC). Four possible outcome proportions result from a diagnostic accuracy analysis: (a) sensitivity, (b) specificity, (c) positive predictive power (PPP), and (d) negative predictive power (NPP).

- Sensitivity describes the "true-positive rate" or the probability that when a diagnostic status is present on the criterion, the individual will be identified positively by the predictor (i.e. the probability that those who did not pass the ELA section of the *i*LEAP would have been predicted to fail on the basis of their CBM score).
- Specificity describes the "true-negative rate" or the probability that when a diagnostic status is absent on the criterion, the individual will not be identified by the predictor (i.e. the probability that those who did pass the ELA section of the *i*LEAP would have been predicted to pass on the basis of their CBM score).
- Positive predictive power (PPP) is the likelihood that an individual who scores lower than the cut score on the predictor measure will in fact have the condition of interest, based on the outcome of the criterion measure (i.e.the probability that those who were predicted to fail the *i*LEAP on the basis or their CBM score did in fact fail the *i*LEAP).
- Negative Predictive Power (NPP) is the likelihood that a score above the cut score on the predictor actually does not have the condition based on the criterion score (i.e. the probability that those who were predicted to pass the *i*LEAP based on their CBM score actually passed).

RESULTS

Descriptive Statistics

Means and standard deviations for all measures at each benchmark period are reported in Table 1. Based on mean results, the CBM production-dependent indices for written expression demonstrated growth from fall to spring. Total words written and words in complete sentences were the only measures that demonstrated successive growth across benchmark periods. Conversely, the means for production-independent and accurate production indices exhibited relative stability from fall to spring, with a slight decrease from fall to spring for percent of correct word sequences. Oral reading fluency demonstrated growth across benchmark periods. Mean performance on the *i*LEAP fell with in the *Basic* achievement level (282-337), which is the lowest level at which students are considered passing.

Reliability

Mean agreements for each variable are reported in Table 2. To ensure consistency of scoring throughout data collection, the scorers reviewed definitions of variables after scoring fall probes and winter probes. This is similar to procedures used by Weissenburger and Espin, whom allowed scorers to email questions throughout the scoring period to promote consistency of scoring. Also, Gansle et al. (2002) asked scorers to score specific variables more carefully when agreement on a set of probes fell below 80%.

Interrater agreement for the indices of written expression ranged between 72.5 and 99.5% on 30% of the CBM passages. Agreement for TWW, WSC, and CWS are consistent with previous studies whose coefficients ranged from .90 to .99 (Gansle et al.,

2002; Gansle et al., 2006; Marston et al., 1981; Marston et al., 1983; Tindal et al., 1991; Videen et al., 1982). Correct punctuation marks, complete sentences, and words in complete sentences also produced acceptable agreements, consistent with those achieved by Gansle and colleagues (2002 & 2006). Assessment of production-independent indices (percentage measures) also demonstrated adequate interobserver agreement

Table 1

Means and Standards Deviations for Curriculum-Based Measures for Written Expression, DIBELS-ORF, and *i*LEAP Variables by Benchmark Period

	Benchmark Period		
	Fall	Winter	Spring
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
CBM-WE			
Production-Dependent Indices			
Total Words Written	36.5 (13.6)	37.5 (12.7)	39.9 (13.8)
Words Spelled Correctly	32.9 (13.2)	32.7 (12.7)	35.5 (12.8)
Correct Word Sequences	28.2 (13.5)	26.9 (13.0)	29.8 (12.7)
Correct Punctuation Marks	2.4 (2.3)	3.3 (2.7)	3.3 (2.6)
Complete Sentences	1.8 (2.1)	1.8 (2.1)	2.2 (2.0)
Words in Complete Sentences	15.5 (17.2)	15.8 (17.4)	19.0 (16.6)
Production-Independent Indices			
Percent Words Spelled Correctly	89.3 (10.3)	86.0 (11.0)	89.1 (9.5)
Percent Correct Word Sequences	73.8 (16.0)	67.5 (18.1)	71.6 (17.4)
Accurate Production Index			
Correct Minus Incorrect Sequence	18.7 (15.3)	15.1 (15.6)	18.1 (16.9)
DIBELS			
Oral Reading Fluency	81.6 (28.2)	90.1 (30.6)	106.8 (28.7)
*i*LEAP			
Total Score ELA			302.9 (52.8)
Percent Correct for Reading			66.4 (19.1)
Percent Correct for Language			65.5 (18.5)
Percent Correct for Writing			59.6 (17.9)

.

Interrater reliability for correct minus incorrect word sequences was considerably lower than other measures (72.5%). This percentage provides an estimate of the reliability achieved in this study. Due to the nature of the measure, some scores resulted in negative numbers. In instances when both scorers found negative numbers, the absolute value of those scores was taken and agreement was calculated (small/large). However, in situations where one scorer calculated a positive number and the other scorer calculated a negative number, a score of 0% reliability was indicated. This allowed for a conservative estimate of reliability. Reliability of CMIWS has not been as thoroughly investigated as other measures. Espin et al. (2000) calculated agreement based on 20 probes from both story and descriptive writing samples for middle school students. Thus, 9% of both story and descriptive writing samples of seventh and eighth graders was assessed for reliability. Interscorrer agreement ranged from 88.3 to 92.5% for 3- and 5-minute story writing and descriptive writing samples. Jewell et al. (2005) only calculated reliability during training for the scoring indices of TWW, WSC, and CWS. They did not calculate reliability coefficients for correct minus incorrect word sequences or percentage measures because of their imbedded relationship with the other indices. Weissenburger et al. (2005) calculated agreement for approximately 14% of writing samples (20 from each of 3 grade levels). They found agreement for correct minus incorrect word sequences for fourth grade students to be relatively lower (74% to 84%) and similar to results in this study. The authors attributed this lower agreement to the high number of errors in the passages of the fourth grade samples and the difficulty reported by scorers in consistently applying rules for this measure. The lower agreement in this study may have been due a combination of factors: high number of errors in third

grade students' writing, low number of total word sequences in some cases (ranged from 7 to 81), poor handwriting of students, and inconsistent application of rules by scorers.

Table 2

Interrater Reliability on Curriculum-Based Measures for Written Expression Indices

Measure	Mean Percent Agreement
Production-Dependent Indices	
Total Words Written	99.5%
Words Spelled Correctly	97.9%
Correct Word Sequences	91.7%
Correct Punctuation Marks	91.4%
Complete Sentences	88.5%
Words in Complete Sentences	87.6%
Production-Independent Indices	
Percent Words Spelled Correctly	98.0%
Percent Correct word Sequences	91.9%
Accurate Production Index	
Correct Minus Incorrect Sequences	72.5%

Time Variables

Efficiency to score is an important characteristic of curriculum-based measures. Thus, time data was collected by the primary investigator (rater 1) on 50% of the samples at each benchmark period (n = 186). Table 3 provides the means for timing of variables. Data were collected by the primary investigator (rater 1) and a second scorer (rater 2). Data from two raters were collected in order to provide information on differences between scorers that may represent actual differences which may occur when educators utilize these measures. Reliability for timing was also collected on 30% of probes timed by Rater 1.

Table 3

Means for Time to Score Probe Variables in Minutes and Seconds by Rater

Measure	Rater 1 (*r*) [a]	Rater 2 [b]
Production-Dependent Indices		
Total Words Written	0:22 (98.5%)	0:51
Words Spelled Correctly	0:55 (99.2%)	1:36
Correct Word Sequences	2:03 (98.6%)	2:38
Correct Punctuation Marks	0:21 (98.0%)	0:22
Complete Sentences	0:30 (97.2%)	0:31
Words in Complete Sentences	0:40 (98.0%)	0:42
Production-Independent Indices		
Percent Words Spelled Correctly	1:03 (99.0%)	1:47
Percent Correct Word Sequences	2:32 (98.4%)	4:37
Accurate Production Index		
Correct Minus Incorrect Sequences	2:26 (98.5%)	4:34

[a] n = 186, *r* = reliability for timing of 30% of timed probes
[b] n = 57 of the 186 probes assessed by rater 1

Criterion-Related Validity

Pearson correlations were calculated between scores from curriculum-based measures and the relevant criterion measures (subscale and total ELA score). Correlations were calculated for each benchmark period. Due to the exploratory nature of these analyses, α for each group of comparisons was set at .10, and was divided by the number of comparisons (10) to arrive at the corrected α of .01. (Gansle et al. 2002). The probability of committing Type I errors is reduced by using this Bonferroni correction. (see tables 4, 5, and 6 for these correlations).

Inspection of Table 4 reveals that CWS, %WSC, CMIWS, and oral reading fluency had significant correlation with all criterion measures for the fall. Reading fluency demonstrated the highest correlation coefficient for the reading subtest, language subtest, and ELA total score. Words spelled correctly demonstrated the highest

coefficient for the writing subtest. All measures of writing demonstrated significant coefficients with the language subtest and the ELA total score. Only TWW, WSC, CWS, %WSC, and CMIWS were significantly related to the writing subtest.

Table 4

Correlations between Fall Curriculum-Based Measures and Criterion Variables

CBM Variable	*i*LEAP Subscale and Total Scores			
	Reading	Language	Writing	ELA Total
Total Words Written	.11	.30*	.42*	.30*
Words Spelled Correctly	.18	.40*	.46*	.38*
Correct Word Sequences	.24*	.50*	.45*	.44*
Correct Punctuation Marks	.17	.46*	.21	.41*
Complete Sentences	.18	.35*	.14	.30*
Words in Complete Sentences	.21	.41*	.18	.35*
% Words Spelled Correctly	.34*	.43*	.25*	.38*
% Correct Word Sequences	.30*	.47*	.18	.38*
Correct – Incorrect Sequences	.28*	.54*	.39*	.46*
DIBELS- Oral Reading Fluency	.51*	.61*	.41*	.64*

* Correlation is significant at the .01 level

Reading fluency demonstrated the highest correlation coefficients with all criterion measures. Consistent with data from the fall measures, all curriculum-based measures demonstrated significant coefficients with the language subtest and the ELA total score. All measures of writing, except TWW and %WSC, had significant coefficients with the writing subtest.

Table 5

Correlations between Winter Curriculum-Based Measures and Criterion Variables

CBM Variable	*i*LEAP Subscale and Total Scores			
	Reading	Language	Writing	ELA Total
Total Words Written	.09	.31*	.21	.23*
Words Spelled Correctly	.18	.41*	.26*	.32*
Correct Word Sequences	.24*	.49*	.32*	.41*
Correct Punctuation Marks	.12	.35*	.28*	.34*
Complete Sentences	.16	.39*	.29*	.36*
Words in Complete Sentences	.17	.40*	.25*	.40*
% Words Spelled Correctly	.34*	.49*	.22	.42*
% Correct Word Sequences	.35*	.51*	.27*	.45*
Correct – Incorrect Sequences	.31*	.54*	.34*	.47*
DIBELS- Oral Reading Fluency	.51*	.62*	.41*	.63*

* Correlation is significant at the .01 level

Table 6 shows that spring scores for correct punctuation marks, %CWS, CMIWS, and ORF demonstrated significant correlation coefficients with all criterion measures. Reading fluency demonstrated the highest coefficients with all criterion measures. Total words written demonstrated no significant correlations with criterion measures; however, the remaining writing measures displayed significant coefficients with the language subtest. Words spelled correctly, CWS, correct punctuation, %CWS, and CMIWS were significantly related to the writing subtest.

Table 6

Correlations between Spring Curriculum-Based Measures and Criterion Variables

CBM Variable	*i*LEAP Subscale and Total Scores			
	Reading	Language	Writing	ELA Total
Total Words Written	-.08	.15	.21	.09
Words Spelled Correctly	.04	.29*	.27*	.22
Correct Word Sequences	.19	.46*	.30*	.38*
Correct Punctuation Marks	.24*	.42*	.24*	.36*
Complete Sentences	.21	.31*	.10	.26*
Words in Complete Sentences	.21	.33*	.13	.27*
% Words Spelled Correctly	.38*	.49*	.22	.42*
% Correct Word Sequences	.40*	.56*	.24*	.50*
Correct – Incorrect Sequences	.33*	.55*	.27*	.47*
DIBELS- Oral Reading Fluency	.52*	.63*	.39*	.65*

* Correlation is significant at the .01 level

Data Reduction

The curriculum-based measures of written expression were moderately intercorrelated. A data-reduction strategy was utilized to identify a small number of factors that summarized the observed correlations among variables; along with, identifying the best exemplar for each factor. Principle components analysis (PCA) was utilized as the data reduction technique. In order to determine the "correct" number of components to retain for the PCA, a parallel analysis was necessary to compare the observed eigenvalues to the 95th percentile eigenvalues derived from random data (Cota,

Longman, Holden, Fekken & Xinaris, 1993). Results of parallel analysis are used to identify the observed eigenvalues that are greater than that which could be expected from an equivalent random data set. The number of observed eigenvalues which are greater represents the number of components that should be extracted from the PCA. Linear interpolation has been shown to be an accurate method of implementing parallel analysis. Linear interpolation was conducted using published tables of 95th percentile eigenvalues (Cota et al., 1993). Results of this analysis (comparison of eigenvalues) revealed that three components should be kept for the principle components analysis.

As displayed in Tables 7, 8, and 9, the PCA revealed three components at each benchmark period. While the amount of variance explained by each component changed between benchmark periods, the same solution for components was consistent. For each component, the highest loading items produced in the pattern matrix were identified as the solution. Three production-dependent variables consistently loaded highly together: TWW, WSC, and CWS. This component accounted for most of the variance for the fall benchmark (58%) but functioned as the second component for winter (19%) and spring (23%). The second component for the fall benchmark period accounted for 21% of the variance and included correct punctuation, complete sentences, and words in complete sentences. This component loaded as the third component for winter (13%) and spring (17%). Two Production-independent measures and the accurate production measure loaded for component 3 in the fall (15%) but as the first component for winter (61%) and spring (53%).

The structure matrix for each benchmark period is also presented. The correlations between variables and factors are demonstrated in the structure matrices.

Table 7

Rotated Three Factor Model for Principal Components Analysis of Nine Direct Measures of Writing from Fall Data

		Matrix variance accounted for by eigenvalues	
Factor	Eigenvalue	Percentage	Cumulative Percentage
I	5.18	57.59	57.59
II	1.87	20.79	78.38
III	1.41	15.64	94.03

Pattern Matrix following Promax rotation with Kaiser Normalization

	Component		
Variable	1	2	3
TWW	1.121		
WSC	1.037		
CWS	.830		
%WSC			1.056
%CWS			.951
CMIWS	.478		.551
Complete Sentence		1.029	
Correct Punctuation		.986	
Words in Complete Sentence		.929	

Structure Matrix

	Component		
Variable	1	2	3
TWW	.961		
WSC	.993	.334	.450
CWS	.971	.556	.640
%WSC	.309		.911
%CWS	.413	.468	.947
CMIWS	.814	.608	.856
Complete Sentence	.325	.971	.360
Correct Punctuation	.330	.932	.335
Words in Complete Sentence	.424	.953	.456

Table 8

Rotated Three Factor Model for Principal Components Analysis of Nine Direct Measures of Writing from Winter Data

		Matrix variance accounted for by eigenvalues	
Factor	Eigenvalue	Percentage	Cumulative Percentage
I	5.49	60.97	60.97
II	1.75	19.41	80.38
III	1.18	13.07	93.44

Pattern Matrix following Promax rotation with Kaiser Normalization

	Component		
Variable	1	2	3
TWW		1.133	
WSC		.982	
CWS	.314	.706	
%WSC	1.053		
%CWS	1.062		
CMIWS	.730		
Complete Sentence			.968
Correct Punctuation			.931
Words in Complete Sentence			.931

Structure Matrix

	Component		
Variable	1	2	3
TWW	.409	.973	.388
WSC	.625	.994	.419
CWS	.780	.949	.583
%WSC	.958	.494	
%CWS	.979	.949	.391
CMIWS	.932	.740	.553
Complete Sentence	.393	.435	.961
Correct Punctuation		.317	.865
Words in Complete Sentence	.403	.466	.946

Table 9

Rotated Three Factor Model for Principal Components Analysis of Nine Direct Measures of Writing from Winter Data

		Matrix variance accounted for by eigenvalues	
Factor	Eigenvalue	Percentage	Cumulative Percentage
I	4.77	53.03	53.03
II	2.10	23.28	76.32
III	1.57	17.39	93.71

Pattern Matrix following Promax rotation with Kaiser Normalization

	Component		
Variable	1	2	3
TWW	-.304	1.078	
WSC		.992	
CWS	.379	.739	
%WSC	1.051		
%CWS	1.00		
CMIWS	.882		
Complete Sentence			1.01
Correct Punctuation			.868
Words in Complete Sentence			.965

Structure Matrix

	Component		
Variable	1	2	3
TWW		.950	
WSC	.406	.994	
CWS	.711	.913	.467
%WSC	.936		
%CWS	.967		.445
CMIWS	.951	.583	.506
Complete Sentence	.363		.963
Correct Punctuation	.436		.893
Words in Complete Sentence	.382		.946

This matrix was used to identify the "best exemplar" for each component, at each benchmark. These "best exemplars" were then used to run three regression analyses (one for each criterion variable). The best exemplars for fall, winter, and spring included WSC, %CWS, and complete sentences. Component scores were also used to run three additional regressions with the criterion variables. Component scores were created by summing of the variables that defined each component.

Relationship between Writing Probe Variables and Criterion Test Scores

A series of stepwise, forward hierarchical multiple regression analyses were conducted to explore the relationship of writing variables and criterion tests scores. These analyses are presented in Tables 10, 11, 12, and 13. The independent variables included best exemplar and component scores of writing variables. The component variables were identified as the production-dependent (TWW, CWS, WSC), production-independent (%CWS, %WSC, CMIWS), and punctuation / sentences (correct punctuation marks, complete sentences, words in complete sentence). While CMIWS is an accurate production measure, it loaded with the production-independent measures in the PCA, thus it was categorized under the "production-independent" term for simplification. The dependent variables, or criterion variables, for these analyses included the language subtest, writing subtest, and total ELA score of *i*LEAP ELA portion. A total of six regressions were conducted three with best exemplars and three with the component variables. For these regressions, the variable that accounted for the most variance entered the equation at each step. For a variable to enter the equation, the probability was set at 5%, the probability to a remove a variable was set at 10% (Gansle et al., 2002; Norušis, 2008). The following statistics are represented in the tables: *SE B*

(standard error of the unstandardized regression coefficient), β (standardized regression coefficient), R^2 (proportion of variance of the dependent variable that is shared with the predictors), and ΔR^2 (the change in the proportion of variance shared with the predictors that is made with the addition of another variable to the regression equation; Gansle et al., 2002; Norušis, 2008).

Language Subtest. When best exemplars for fall, winter, and spring variables were assessed in the regression analysis, five variables entered the regression equation (Table 11). These variables included the percent CWS from the fall, WSC in the fall, percent CWS from the winter, winter complete sentences, and percent CWS from the spring benchmark. These variables together accounted for 45% of the variability (R^2) for the language subtest. The change in R^2 at each variable entry was significant. However, as new variables entered the model, the significance of β values changed for the variables previously entered. Thus, once percent CWS from the winter screening entered the model, fall percent CWS had a β value no longer significantly different from 0. Subsequently, when spring percent CWS entered the model, neither fall nor winter CWS β values were significantly different from 0. Therefore, in model 5, the variables making a significant contribution to the model included fall WSC, $t(118) = 3.24$, $p < .01$; winter complete sentences, $t(118) = 2.20$, $p < .05$; and spring percent CWS, $t(118) = 4.10$, $p < .01$.

An additional analysis was run with the fall, winter, and spring component variables. This model included four predictors. The fall production-independent component, fall punctuation/sentences, winter production-independent, and spring production-independent components entered the equation (Table 11). The proportion of

variability in the dependent variable accounted for in this regression equation was 41%. This is slightly less than that found with the best exemplar variables. The addition of each component to the equation resulted in a significant addition of variance to the equation ($p < .05$). Each variable was positively related to the criterion. However, the significance of the β values changed for each variable at the addition of new variables. In model 4, only fall punctuation/sentences, , $t(119) = 2.42$, $p < .05$, and spring production-independent, , $t(119) = 3.08$, $p < .01$, were significant predictors of performance on the language subtest.

Writing Subtest. Two best exemplar variables entered the regression equation for the writing subtest, fall words spelled correctly and winter complete sentences (Table 11). Each variable was positively related to the criterion. The variables together accounted for 25% of the variability in the dependent variable (R^2). The addition of complete sentences was significant at $p < .05$.

Only one component variable entered the regression equation for the writing subtest (Table 11). The fall production-dependent component accounted for 21% of the variance in the dependent variable. This was less than the amount attributed to the regression equation with the best exemplar variables. This variable was positively related to the criterion.

Total English Language Arts (ELA) Score. Six variables entered the regression model for the total ELA score. These variables included fall WSC, fall percent CWS, winter percent CWS, winter complete sentences, spring percent CWS, and spring WSC. Excluding spring WSC, these are the same variables which entered the model for predicting performance on the language subtest. These variables accounted for 39% of

the variability in the dependent variable (R^2). The addition of each variable was significant at $p < .05$. However, the final model only contained four variables with significant β values: fall WSC, $t(117) = 3.87$, $p < .01$; winter complete sentences, $t(117) = 2.27$, $p < .05$; spring percent CWS, $t(117) = 4.07$, $p < .01$; and spring WSC, $t(117) = -2.07$, $p < .05$. Three variables were positively related, while spring WSC had a negative relationship.

The model summary for the component variables was similar to that found for the language subtest. The following components entered the equation for total ELA: fall production-independent, fall punctuation/sentences, winter production-independent, and spring production-independent (Table 13). The four component variables together accounted for 31% of the variability in the dependent variable, which is less than that accounted for by the best exemplar variables. Variables entered the equation at $p< .05$ significance level. The final model resulted in two variables with significant β values, these included fall punctuation/sentences, $t(119) = .185$, $p < .05$, and spring production-independent, $t(119) = 2.53$, $p < .05$. Both variables were positively related to the total ELA score.

Modeling the Probability of Performance on the ELA Portion of the *i*LEAP

Results from the prediction models for the ELA test are presented in Tables 14 and 15. Two forward, stepwise logistic regression analyses were conducted. While stepwise procedures run the risk of modeling noise in the data, they are considered useful for exploratory purposes (http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm). Data were entered into blocks (hierarchical regression) as they were available during the school year (block 1 = fall data, block 2 = winter, block 3 = spring).

Table 10

Summary of Stepwise, Forward Multiple Regression Analyses of Writing Variables Variables Related to Language Subtest of the *i*LEAP

Independent Variable		*SE B*	β	R^2	ΔR^2
Best Exemplar	Model 1:				
	Fall Percent CWS	.093	.473[a]	.224	.224[a]
	Model 2:				
	Fall Percent CWS	.096	.379[a]	.284	.060[a]
	Fall Words Spelled Correctly	.116	.262[a]		
	Model 3:				
	Fall Percent CWS	.119	.171	.342	.058[a]
	Fall Words Spelled Correctly	.112	.238[a]		
	Winter Percent CWS	.102	.325[a]		
	Model 4:				
	Fall Percent CWS	.118	.141	.368	.027[b]
	Fall Words Spelled Correctly	.111	.215[a]		
	Winter Percent CWS	.102	.285[a]		
	Winter Complete Sentences	.717	.180[b]		
	Model 5:				
	Fall Percent CWS	.116	.013	.447	.079[a]
	Fall Words Spelled Correctly	.105	.241[a]		
	Winter Percent CWS	.103	.137		
	Winter Complete Sentences	.674	.166[b]		
	Spring Percent CWS	.097	.376[a]		
Component Variable	Model 1:				
	Fall Production-Independent	.037	.533[a]	.284	.284[a]
	Model 2:				
	Fall Production-Independent	.042	.434[a]	.314	.030[b]
	Fall Punctuation/Sentences	.076	.199[b]		
	Model 3:				
	Fall Production-Independent	.055	.193	.367	.053[a]
	Fall Punctuation/Sentences	.074	.212[b]		
	Winter Production-Independent	.045	.329[a]		
	Model 4:				
	Fall Production-Independent	.054	.122	.414	.047[a]
	Fall Punctuation/Sentences	.071	.196[b]		
	Winter Production-Independent	.048	.178		
	Spring Production-Independent	.044	.304[a]		

[a] $p < .01$ [b] $p < .05$

Table 11

Summary of Stepwise, Forward Multiple Regression Analyses of Writing Variables Related to Writing Subtest of *i*LEAP

Independent Variable		*SE B*	β	R^2	ΔR^2
Best Exemplar	Model 1:				
	Fall Words Spelled Correctly	.109	.464[a]	.216	.216[a]
	Model 2:				
	Fall Words Spelled Correctly	.111	.419[a]	.246	.030[b]
	Winter Complete Sentences	.707	.179[b]		
Component Variable	Model 1:				
	Fall Production-Dependent	.037	.455[a]	.207	.207[a]

[a] $p < .01$ [b] $p < .05$

Table 12

Summary of Stepwise, Forward Multiple Regression Analyses of Best Exemplar Writing Variables Related to Total ELA Score on the Third Grade *i*LEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Best Exemplar	Model 1:				
	Fall WSC	.335	.378[a]	.143	.143[a]
	Model 2:				
	Fall WSC	.346	.280[a]	.209	.066[a]
	Fall Percent CWS	.286	.275[a]		
	Model 3:				
	Fall WSC	.335	.255[a]	.271	.062[a]
	Fall Percent CWS	.355	.058		
	Winter Percent CWS	.306	.337[a]		
	Model 4:				
	Fall WSC	.333	.232[a]	.298	.027[b]
	Fall Percent CWS	.353	.028		
	Winter Percent CWS	.307	.297[a]		
	Winter Complete Sentences	2.151	.181[b]		
	Model 5:				
	Fall WSC	.317	.257[a]	.372	.074[a]
	Fall Percent CWS	.353	-.096		
	Winter Percent CWS	.312	.153		
	Winter Complete Sentences	2.044	.167[b]		
	Spring Percent CWS	.295	.365[a]		
	Model 6:				
	Fall WSC	.390	.377[a]	.394	.022[b]
	Fall Percent CWS	.354	-.136		
	Winter Percent CWS	.311	.184		
	Winter Complete Sentences	2.023	.180[b]		
	Spring Percent CWS	.295	.397[a]		
	Spring WSC	.398	-.199[b]		

[a] $p < .01$ [b] $p < .05$

Table 13

Summary of Stepwise, Forward Multiple Regression Analyses of Component Writing Variables Related to Total ELA Score on the Third Grade *i*LEAP

Component Variable					
	Model 1:				
	Fall Production-Independent	.113	.446	.199	.199[a]
	Model 2				
	Fall Production-Independent	.128	.353[a]	.225	.026[b]
	Fall Punctuation/Sentences	.231	.187[b]		
	Model 3:				
	Fall Production-Independent	.168	.120	.275	.049[a]
	Fall Punctuation/Sentences	.225	.199[b]		
	Winter Production-Independent	.137	.318[a]		
	Model 4:				
	Fall Production-Independent	.168	.057	.312	.037[b]
	Fall Punctuation/Sentences	.220	.185[b]		
	Winter Production-Independent	.149	.183		
	Spring Production-Independent	.135	.271[b]		

[a] $p < .01$ [b] $p < .05$

Beta weights, standard errors, Wald statistic, and *p* values are presented for each predictor. The Wald Statistic and the p-value are used to test the significance of individual logistic regression coefficients for each independent variable.

To compare the changes in the classification trend with the addition of variables at each step, the number of true negatives, false negatives, true positives, false positives, and hit rate is presented. The hit rate is simply the number of correct classifications divided by the sample size. This rate can be compared to the baseline hit rate provided by SPSS, which uses the most numerous categories to classify all cases. The most numerous category was the number of students passing, which resulted in a baseline comparison of 71%. Thus, if one guessed for all cases that the test would be passed, one would be correct 71% of the time.

DIBELS oral reading fluency (ORF) for fall, winter, and spring were included in both analyses, along with the relevant written expression variables that were found to be significantly related to the total ELA score in the multiple regression analyses. The first analysis was conducted with the best exemplars: fall WSC and percent CWS; winter percent CWS and complete sentences; and spring percent CWS and WSC. The second analysis was conducted using the component variables: fall production-independent, fall punctuation/sentences, winter production-independent, and spring production-independent.

Best Exemplar. In a forward, stepwise logistic regression analysis, DIBELS ORF fall score entered the equation first, in block 1 (Table 14). The addition of fall WSC had no improvement in the correct classification of case (82.3% to 82.3%). When winter complete sentences entered in block 2, correct classification improved to 85.5%. No spring variables for writing entered the equation. DIBELS ORF scores for winter and spring did not enter the equation. The correct classification rate for the last step (85.5%) is considered "good" compared to the baseline classification rate of 71%.

Component Variables. In a forward, stepwise logistic regression analysis, DIBELS ORF fall score entered the equation first (Table 15). The addition of the spring production-independent component decreased the correct classification hit rate from 82.3 to 80.6. Norušis (2008) noted that using the percentage of correct classification depends heavily on the types of cases included in the sample. Moreover, the percentage of correct classification does not necessarily depend on how well a model fits. It ignores actual probability values and replaces them with a cutoff value (by default, it looks at if the estimated probability is greater or less than 0.5 to identify its classification). Thus, a

highly significant variable may be added to the model and result in a decrease in the correct classification rate, which may explain why the hit rate decreased when the spring production-independent variable was added. The last step classification rate (80.6%) can be considered "good" when comparing it to the baseline classification rate (71%).

Table 14

Logistic Regression Analysis for Predicting Pass / Fail Status of the English Language Arts Section of the *i*LEAP from Best Exemplar Variables and DIBELS ORF

Variable	β	*SE*	Wald	*p*	TN	FN	TP	FP	Hit Rate
BLOCK 1									
Step 1					20	16	82	6	82.3
Fall ORF	-.074	.015	24.745	.000					
Step 2					21	15	81	7	82.3
Fall ORF	-.064	.015	17.374	.000					
Fall WSC	-.047	.024	3.913	.048					
BLOCK 2									
Step 1					25	11	81	7	85.5
Fall ORF	-.062	.016	15.722	.000					
Fall WSC	-.055	.026	4.573	.032					
Winter Complete Sentences	-.660	.224	8.677	.003					

Diagnostic Efficiency Statistics

Diagnostic efficiency statistics are reported in tables 16, 17, 18, and 19. These tables provide information regarding accuracy of predicting the passing or failing status of students on the English Language Arts (ELA) section of the *i*LEAP based on whether

Table 15

Logistic Regression Analysis for Predicting Pass / Fail Status of the English Language Arts Section of the *i*LEAP from Component Variables and DIBELS ORF

Variable	β	*SE*	Wald	*p*	TN	FN	TP	FN	Hit Rate
BLOCK 1									
Step 1					20	16	82	6	82.3
Fall ORF	-.074	.015	24.745	.000					
BLOCK 3									
Step 1					20	16	80	8	80.6
Fall ORF	-.059	.015	15.007	.000					
Spring PI	-.016	.007	5.265	.022					

a student scores at or above a cut off point on the selected CBM measures at each benchmark period (fall, winter, spring). The selected CBM measures including oral reading fluency (ORF), total words written (TWW), correct spelled words (CSW), and correct word sequences (CWS). The cut scores used for ORF were the scores published for DIBLES assessment (https://dibels.uoregon.edu/benchmark.php#3grade3). In regards to the scores used for the written expression data, the aggregate norm data collect by *AIMSWEB* since the 1999 – 2000 school year through the 2007 - 2008 school year (www.aimsweb.com) was used and the scores at the 50th percentile at each benchmark period were selected as the cut scores to be assessed for diagnostic efficiency. The base rate of the population assessed for passing the ELA portion of the *i*LEAP was 71% (n = 88), while the rate for failing was 29% (n = 36).

Diagnostic accuracy was represented using the following descriptive statistics (Shapiro, Keller, Lutz, Santoror, & Hintze, 2006): (a) *sensitivity* refers to the probability that the CBM score will accurately identify those students who were not successful at passing the ELA section of the *i*LEAP; (b) *specificity* refers to the probability that the CBM score will accurately identify those students who have been successful on the ELA section of the *i*LEAP; (c) *positive predictive power* refers to the probability that those students identified as failing on the CBM measure will be correctly identified as failing on the ELA section *i*LEAP; and (d) *negative predictive power* refers to the probability that students identified as successful on the CBM measure will also be identified as successful on the ELA portion of the *i*LEAP.

Of the CBM for writing, correct word sequences demonstrated higher percentages for almost all of the statistics. The positive predictive power, or the probability of the measure correctly predicting who would fail the ELA section, for CWS was .74, .51, and ,45 for each benchmark period. This suggests that 74%, 51%, and 45% would be correctly identified as failing based on the 50th percentile score used as the cut score, which is considerable above the 29% base rate. The negative predictive power, or the probability of correctly predicting who would pass the ELA section based on the cut scores, was above base rate levels for passing (71%) at each benchmark. For the population sampled, the negative predictive power was .84, .87, and .87.

Positive predictive power for the DIBELS ORF cut scores was also consistently above base rate level for failing (29%), with statistics ranging from .43 to .56. The negative predictive power for the ORF cut scores produced probabilities above the base rate level of passing (79%), ranging from .89 to .94.

Table 16

Diagnostic Efficiency Statistics for Total Words Written

	Fall	Winter	Spring
Total Words Written 50th percentile	26 tww	32 tww	37 tww
Sensitivity	36%	61%	47%
Specificity	85%	73%	58%
Positive Predictive Power	50%	48%	32%
Negative Predictive Power	77%	82%	73%

Table 17

Diagnostic Efficiency Statistics for Words Spelled Correctly

	Fall	Winter	Spring
Words Spelled Correctly 50th percentile	23 wsc	28 wsc	33 wsc
Sensitivity	44%	67%	58%
Specificity	85%	75%	58%
Positive Predictive Power	62%	52%	36%
Negative Predictive Power	77%	85%	77%

Table 18

Diagnostic Efficiency Statistics for Correct Word Sequences

	Fall	Winter	Spring
Correct Word Sequences 50th percentile	18 cws	24 cws	28 cws
Sensitivity	55%	82%	85%
Specificity	92%	70%	61%
Positive Predictive Power	74%	51%	45%
Negative Predictive Power	84%	87%	87%

Table 19

Diagnostic Efficiency Statistics for DIBELS Oral Reading Fluency

	Fall	Winter	Spring
Oral reading fluency cutoff	77 wpm	92 wpm	110 wpm
Sensitivity	78%	92%	86%
Specificity	75%	57%	55%
Positive Predictive Power	56%	46%	43%
Negative Predictive Power	89%	94%	90%

DISCUSSION

This study aimed to further clarify and identify measures of written expression that demonstrate sufficient technical adequacy for educational decision making. Such measures are needed to identify student progress toward writing proficiency, detect struggling students, and aid in instructional decisions to improve writing ability (McMaster et al., 2007). Successive analyses were used to identify the measures significantly related to educational outcomes on a state standardized assessment. Results from this study revealed a consistent three component solution at each benchmark period, which provides evidence of possible unique dimensions of writing (Tindal et al., 1991). Fall words spelled correctly, winter complete sentences, and spring percent correct word sequences were the variables which made consistent significant contributions to the regression models for predicting performance on the criterion variables. In regards to the language subtest and total ELA score criterion variables, percent correct word sequences was the only variable which entered the models for fall, winter, and spring scores. However, only fall words spelled correctly, winter complete sentences, and fall DIBELS ORF contributed to the prediction of passing or failing the ELA portion of the *i*LEAP above that predicted by the baseline. Percent correct word sequences did not combine with DIBELS ORF to predict performance on the criterion variable.

This study addressed four main research questions. The first research question inquired about the relationship of the curriculum-based measures with reading and writing outcomes, as measured by performance on subtests and the total score for the English Language Arts section. Results of the Pearson correlation analyses revealed consistent moderate correlations for ORF across benchmark periods. The strongest

correlation for ORF was with the total ELA score and the weakest occurring with the writing subtest. Correlations with the total ELA score across benchmark periods (r = .64, .63, and .65, respectively) were similar to those found by Shapiro et al. (2006) for third graders (r = .65, .66, .67, respectively). The correlation for the winter score was slightly lower than that found by Wood et al. (2006) for third graders (r = .70); and slightly higher than that found by Crawford et al. (2001) (r = .60). Interestingly, it's correlation with the language subtest was higher than with the reading subtest across benchmark periods.

In regards to the relationship of writing measures with student performance, as expected, all correlations with the reading subtest were weak. Total words written demonstrated inconsistent, weak correlations across benchmark periods. These weak correlations with the criterion variables are consistent with results from Gansle et al., (2002). However, Espin et al. (2000) found eighth graders' scores for total words written for story writing in the winter correlated with the state writing assessment at a similar level to that found in this study for the fall correlation with the writing subtest (r = .46 versus r = .42). Espin found a similar correlation for descriptive writing (r = .43). Weissenburger et al. (2005) found correlations which varied depending on grade level. Correlations with the state ELA test for fourth graders (r = .45) was higher than that found for eighth graders (r = .26). The present study found that the correlations declined across benchmark periods with the ELA total score (r = .30, .23, .09, respectively). These findings reveal the inconsistent relationship of total words written with criterion measures assessed.

Correct word sequences and percent correct word sequences had moderate correlations with the language subtest across benchmark assessments (r = .47 - .56). These correlations were higher than those found for third graders by Gansle et al. (2002) for words in correct sequence (r = .36). Espin et al. (2000) found eighth graders' scores with the district writing test were stable for story and descriptive writing (r = .61). Correlations for CWS from this study were not as strong and declined across benchmark periods. Moreover, the correlations with the total ELA score for CWS declined across benchmark periods, whereas an increase across periods was observed for percent correct sequences. Correlations with the total score (r = .38 - .44) were similar to that found by Gansle et al. (2002) for correct sequences (r = .43). Weissenburger et al. (2004) found moderate correlations with the total language arts test and writing assessment for fourth and eighth graders (r = .50 - .62) but a weak correlation for tenth graders with the total language arts score (r = .18).

Correct minus incorrect word sequences exhibited moderate correlation coefficients with the language and total ELA score across benchmark assessments (r = .46 - .55). Weissenburger et al. (2004) also found slightly higher, moderate correlations with the writing test and total language arts scores for fourth and eighth graders (r = .56 - .68), but a lower correlation for tenth graders on the total language arts test (r = .29). Espin et al. (2000) demonstrated the highest correlation with the district writing test for eighth graders (r = .69). Correct minus incorrect word sequences demonstrated consistently higher correlations than other measure.

Multiple regression analyses were conducted to identify the best predictors of performance on the criterion variables. However, before this analysis could be

conducted, the number variables needed to be reduced in order to combat multicollinearity. Principle component analysis was used as the variable reduction strategy to summarize the observed correlations among CBM writing variables. The three component solution was consistent across benchmark periods. The best exemplar variable for each component was also consistent across benchmark periods. When component scores and best exemplar variables were regressed on the criterion variables, the best exemplar equations consistently accounted for a higher amount of variability for criterion variables than that accounted for by component variables. Fall words spelled correctly and winter complete sentences entered the regression equation for all criterion variables and maintained significant contributions in the final models. Fall, winter, and spring percent correct word sequences entered the equations for the language subtest and the total ELA score. However, only the spring percent correct word sequences variable was a significant predictor in the final model, which is consistent with the highest correlation between this measure and the criterion variable occurring at the spring benchmark assessment. Gansle et al. (2002) observed a different regression model resulting for the language usage and language total scales. Essentially, the model for language usage included words in correct sequence, verbs, and correct punctuation; while the total language model included correct punctuation, words in correct sequence, total words written, and long words. The differential results between these studies should be interpreted with due caution. A variable reduction strategy was used in the present study to reduce the number of variables used for the regressions, whereas, all variables assessed in Gansle et al. (2002) had an opportunity to enter the equations. Also, Gansle et al. (2002) had greater diversity in the variables assessed than the present study.

The final model for the total ELA score was similar to the model produced for the language subtest. However, an additional variable (spring words spelled correctly) entered at the last step for the total ELA score. The spring words spelled correctly variable had a negative relationship with the total ELA score. This variable may have been functioning as a suppressor variable (Stevens, 2002). A suppressor variable is essentially variable which may have a low correlation with the criterion but has high correlations with other predictors. The variance of the predictors related to the suppressor variable but not to the criterion variable is termed irrelevant variance. When a suppressor variable enters the equation, the irrelevant variance on the other predictors is partialled out (or suppressed). This results in an increase in the predictive power of the other variables because the remaining variance is more strongly related to the criterion (Stevens, 2002). This may have been the function of the spring WSC variable. It's correlation with the criterion was .22, while its correlation with fall WSC was .63, with winter complete sentences was .25, and with spring percent CWS was .27. When spring WSC entered the equation, the contribution of each of these variables increased, while the overall predictive power also increased by approximately 2%.

Logistic regressions were analyzed to answer the question of whether any variables of written expression would add to the predictive utility of ORF for identifying the passing / failing status of students. The model produced by the exemplar variables resulted in a higher hit rate than that resulting from the component variables. Fall WSC and winter complete sentences contributed to the fall ORF score for predicting overall performance on the ELA portion of the *i*LEAP. In regards to the component variables, the spring production-independent component variable entered the model with fall ORF

but a corresponding increase in the hit rate was not observed. The winter and spring ORF scores did not enter the regression analyses presumably because they added no incremental predictive validity beyond what the fall score contributed. Results from this investigation (r = .63 -.65) and scores for third graders in Hintze et al. (2005) (r = .66 - .69) demonstrated relative stability of the correlations between ORF and state assessments throughout the school year. Hintze et al. (2005) also found diagnostic accuracy statistics for the logistic regression analysis to be consistent across fall, winter, and spring for the third grade scores. Thus, it appears that the fall score for ORF may predict as well as that produced in the spring, which is relevant for identifying struggling readers at the beginning of the year.

An analysis of diagnostic accuracy was conducted to further assess how well CBM measures predict outcomes. The diagnostic utility of ORF, TWW, CWS, and WSC was assessed for predicting ELA performance on the *i*LEAP. Of the written expression measures, CWS, on average, had the highest sensitivity, specificity, PPP, and NPP. Use of the 50th percentile cut point resulted positive predictive power (correct prediction that if a student was below the criterion on CWS, they were also below criterion on the ELA test of the *i*LEAP) of around 45% to 74%, and negative predictive power (correct prediction that a student above the criterion on the CBM was also above the criterion for the ELA score) of around 84% to 87%. Sensitivity (ranged from .55 to .85) and specificity (ranged from .61 to .92) were close to acceptable levels for screening purposes (.6 or higher; Shapiro et al., 2006). This diagnostic accuracy may be improved by looking at values other than the 50th percentile. Oral reading fluency demonstrated higher negative predictive power (89% to 94%) than CWS, but lower positive predictive

power (43% to 56%). Sensitivity was also better for ORF (.78 to .92) but specificity was lower (.55 to .75). These percentages are highly sensitive to the cut score used for calculation, thus use of a different cut score may improve these statistics. Shapiro et al. (2006) used receiver operating characteristics (ROC) curves to identify scores at which sensitivity was maintained with little change in specificity. They achieved sensitivity (range from .69 to .86) and specificity (range from .67 to .83) with similar ranges. However, their negative predictive power (43% to 68%) was lower than their positive predictive power (.83 to .94), which is the converse to what this study found.

In addition to addressing the validity and utility of these measures, the reliability was also assessed to ensure the precision, accuracy, and consistency of the measurement procedures (Thorndike, 2005). Interrater reliability for scoring the measures of written expression was consistent with previous research (Gansle et al., 2002; Gansle et al., 2006; Marston et al., 1981; Marston et al., 1983; Tindal et al., 1991; Videen et al., 1982). Reliability for correct minus incorrect word sequences was relatively poor. While no previous study has as extensively assessed the reliability for this measure, previous results of reliability were stronger but also reported similar difficulties with this measure (Espin et al., 2000; Weissenburger et al., 2005). The scorers assessed the reasons for the weak reliability coefficient. Aside from scorer error in application of rules, additional difficulties occurred due to poor handwriting of students which resulted in differential interpretations of writing samples. Also, due to the nature of the measure, it was sensitive to small differences. For example, on a sample with 30 word sequences, when the raters were off by 2 correct word sequences (rater 1 = 18 cws, rater 2 = 16 cws) this would result in a difference of 4, or a doubling of the CWS difference, when calculating

the correct minus incorrect measure (rater 1: 18 – 12 = 6; rater 2: 16 – 14 = 2). Thus, reliability for CMIWS in this example would be 33%, when the reliability for the CWS measure was 89%. This measure may be more precisely assessed when students' writing samples are more accurate and legible.

Time variables were also reported in this study to provide evidence of efficiency for scoring the various indices of written expression. Gansle et al. (2002) also collected time data for scoring of variables. They found sentence fragments and correct punctuation took the least amount of time to score (8.5 and 10.7 seconds, respectively). Results from the current investigation also found correct punctuation marks (21 seconds) to be the least time consuming measure followed by total words and complete sentences (22 and 30 seconds, respectively). Gansle et al. (2002) found words in correct sequence to be the most time consuming measure (57.3 seconds). Similar results were found for this study with all the correct sequence variables taking the most time. Correct word sequences took, on average, two minutes and three seconds, while the accuracy measure (CMIWS) took two minutes and twenty-six seconds. The measure which took the longest was the percent measure at two minutes and thirty-two seconds. Rater two's average time for scoring the percent and accuracy measures were four minutes, thirty-seven seconds and four minutes, thirty-four seconds, respectively. This time difference between raters may be attributed to the amount of times the scorers rechecked their writing samples before writing the scores down. Also, the discrepancy between time means from Gansle et al. (2002) and the current investigation may be due to the writing quality of the samples. As stated previously, the handwriting quality may have affected the difficulty of scoring some passages. Regardless, the efficiency of scoring CBM is

crucial to its intended use. When measures are to be administered system-wide and frequently, these measures must have efficiency in order to be used informatively. This should be considered in future investigations to ensure that the measures we are assessing truly adhere to the characteristics of a curriculum-based measure as brief or time efficient (Jenkins et al. 1979; Marston, 1989).

Limitations

This study had a number of limitations that restrict interpretation of the findings. First, the study was conducted in only two schools within one school district in Louisiana. The percentage of students who passed the ELA section of the *i*LEAP, in the district studied, was 76%, in comparison, the percentage who passed statewide was approximately 64%. The percentage who passed in this study (79%) is similar to the district's percentage but not to the state percentage. Thus, it is indeed possible that the outcomes of these schools would not fully represent the whole district, much less, the state.

Secondly, only students whose full data sets were available (fall, winter, and spring ORF and written expression CBM, and *i*LEAP data) were utilized for this study. Attrition of student data collected throughout the school year is inevitable; however, such attrition may result in a final analysis sample that is somewhat different from the original sample. By comparing the performance of the sample versus the performance for each school, the effects of this can be taken into perspective. For example, one school's passing performance as reported by the state was approximately 79%, however the sample of students from that school had a passing rate of approximately 87%. Thus, the students who did not pass the ELA portion of the *i*LEAP were underrepresented for this

school. In regards to the other school, it's overall passing rate was approximately 49%, while 50% of the sample selected passed. These discrepancies affect the interpretability of the data.

Other important limitations limit the generalizability of the results to other populations or individuals. Model validation was not used to assess the generalizability of the prediction equations obtained from the regression analyses (Stevens, 2002). A model almost always fits the sample from which it was derived better than it would fit a sample from the same population (Norušis, 2008). Also, considering that best exemplar variables were chosen based on the correlations that resulted from the population sampled, different correlation values may have possibly resulted in different best exemplars being chosen. Ideally, the variables chosen best represented their component. However, regressions were not run with the other variables; therefore, the performance of variables left out of these analyses is unknown. Thus, the problem of generalization of these findings may also be affected by the best exemplars chosen. This limitation should be addressed in future investigations by continuing to identify unique measures of writing through component analyses and utilizing a form of model validation to check for generalizability of the findings.

Implications for Practice and Future Research

Based on results from this investigation and previous investigations, the correct word sequence measures (correct word sequence, percent correct word sequence, correct minus incorrect word sequence) have demonstrated promising potential as a CBM of written expression (Espin et al., 2000; Gansle et al., 2002; Weissenburger et al., 2005). Words spelled correctly in the fall and complete sentences from the winter were also

promising measures. However, more investigation is needed on how these measures perform as progress monitoring instruments.

While the principle components analysis was not the primary analysis of focus for this investigation, the results were interesting. Specifically, the PCA resulted in the same three factor solution at each benchmark period, with variables consistently loading under the same component. The component solution for the production-dependent variables is consistent with that found by Tindal et al. (1991). Tindal et al. found number of words written, correctly spelled words, correctly sequenced words, and total number of sequenced words to account for 81% of the variance. They also found percent of correctly sequenced words and subjective judgment of conventions-mechanics accounted for 26% of the variance. Despite Tindal and Parker's recommendation that more studies involve factor analysis of written expression, extensive replication has not occurred. This type of research may provide information on the different dimensions of written expression being studied, which is considered important when attempting to rate students' writing on unique or independent dimensions (Tindal et al., 1991). This is supported by the results for CMIWS. While CMIWS is referred to in previous studies as an accuracy measure (Jewel et al., 2005), it loaded on the component with the production-independent measures. Future research should continue to investigate which measures provide unique information regarding writing proficiency.

When conducting further investigations of CBM for written expression, researchers should continue to collect reliability and time information from teachers. Gansle et al. (2002) noted the frustration teachers currently experience with using total words written as an indicator of writing proficiency. CBM is becoming more extensively

used due to the changes in NCLB (2001) and IDEA (2004). The combination of teacher frustration and increased use of CBM makes for an opportune moment to increase teacher involvement in validation studies. Deno et al. (1982) noted that while technical characteristics of data for writing indices seem ideal for use in a routine, systematic formative evaluation, teacher application of this information to improve students' writing has yet to be evaluated. Gansle et al. (2002) reiterated this need for treatment utility. Thus, future research should provide evidence for teachers' reliable, efficient measurement of these promising CBM indices while using them to guide instructional decisions.

This study sought to identify quantitative aspects of writing that would provide the most useful information for identifying student success on a state assessment. This information is important for identifying these students early in the school year and providing remedial instruction before high stakes testing occurs. Results from this study were relatively consistent with other studies in terms of correlations. However, future investigations need to continue consideration of qualitative measures in combination with quantitative measures (McMaster et al., 2007; Tindal & Hasbrouck, 1991). Technically sound qualitative measures may serve as unique dimensions of writing which are important for predicting success and modifying instruction. The importance of these measures demonstrating growth should also be considered, as this is an informative element of a curriculum-based measure (Shinn, 1995). Future research should continue to assess the function of writing indices across grade levels and benchmark assessments. This information would help identify the unique dimensions of writing, assess whether

these dimensions change as student's progress through school, and whether the importance of these dimensions changes over time as writing skills develop.

REFERENCES

Brown-Chidsey, R., & Steege, M. W. (2005). *Response to intervention: Principles and strategies for effective practice.* New York: Guilford Press.

Children's Educational Services. (1987). *Test of Reading Fluency.* Minneapolis, MN: Author.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cooper, J., Heron, T., & Heward, W. (1987). *Applied Behavior Analysis.* Upper Saddle River, New Jersey: Prentice-Hall.

Cota, A. A., Longman, R. S., Holden, R. H., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurment, 53,* 585-596.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7,* 303-323.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52,* 219-232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37,* 184-192.

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship, 48,* 368-371.

Deno, S. L, Marston, D., Mirkin, P. K., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (Research Report No. 87). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities (ERIC Document Reproduction Service No. ED 227 129).

Deno, S. L., Marston, D., Shinn, M. R., & Tindal, G. (1983). Oral reading fluency: A simple datum for scaling reading disability. *Topics in Learning and Learning Disability, 2,* 53-59.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49,* 36-45.

Deno, S. L., Mirkin, P., & Marston, D. (1980). Relationships among simple measures of written expression and performance on standardized achievement tests (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody Individual Achievement Test: Manual.* Circle Pine, MN: American Guidance Service.

Espin, C., Shin, J., Deno, S., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*, 140-153.

Fewser, S., & Macmillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement or reading and writing. *Remedial and Special Education, 23*, 149-156.

Fuchs, L. S., & Deno, S.L. (1981). *The relationship between curriculum-based mastery measures and standardized tests in reading.* (Research Reprot No. 57). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities (ERIC Document Reproduction Service No. ED 212 662).

Fuchs, L., & Deno, S. (1982). *Developing goals and objectives for educational programs* [Teaching Guide]. Minneapolis, MN: Institute for Research in Learning Disabilities, University of Minnesota.

Fuchs, L., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30*, 1-16.

Fuchs, L. S., Fuchs, D., & Maxwell, S. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20-28.

Fuchs, L. S., Deno, S. L., & Marston, D. (1982). Use of aggregation to improve the reliability of simple direct measures of academic performance (Vol. IRLD-RR-94). University of Minnesota, Institute for Research on Learning Disabilities.

Fuchs, L. S., Tindal, G., Fuchs, D., Shinn, M. R., Deno, S. L., & Germann, G. (1983). *Technical adequacy of basal readers' mastery tests: Holt basic series* (Research Report No. 130). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Fuchs, L. S., Tindal, G., Shinn, M. R., Fuchs, D., Deno, S. L., & Germann, G. (1983). *Technical adequacy of basal readers' mastery tests: Ginn 720 series* (Technical Report No. 122). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Gansle, K. A., Vanderheyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*, 435- 450.

Gansle, K. A., Noell, G. H., VanDerheyden, A. M., Naquin, G. M., & slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477-497.

Good, R. H., Gruba, J. & Kaminski, R. A. (2001). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes driven model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV*(pp. 679-700). Washington, DC: National Association of School Psychologists.

Good, R., & Jefferson, G. (1998). Contemporary perspectives on Curriculum-Based Measurement Validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61-88). New York: Guilford Press.

Good, R. H., Simmons, D. C., & Kameenui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high stakes outcomes. *Scientific Studies of Reading, 5*, 257-288.

Hammill, D. D., & Larsen, S. C. (1978). *Test of Written Language.* Austin, TX: Pro-Ed.

Hammill, D. D, & Larsen, S. C. (1983). *Test of Written Language.* Austin, TX: Pro-Ed.

Harcourt Brace Educational Measurement. (1996). Stanford Achievement Test, Ninth Edition. San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997). Stanford Achievement Test, Ninth Edition. San Antonio, TX: Harcourt Brace Educational Measurement.

Hasbrouck, J. E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children, 24*, 41-44.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools, 43*, 45-56.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372-386.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding word reading and comprehension: Do the relations change with the grade? *School Psychology Review, 34*, 9-26.

Howell, K. W., & Nolet, V. (2000). *Curriculum-based evaluation: Teaching and decision making* (3rd. ed.). Belmont, CA: Wadsworth.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. 1400 *et. seq.*

Jenkins, J. R., Deno, S. L., Mirkin, P. K. (1979). Measuring pupil progress toward the least restrictive environment. *Learning Disability Quarterly, 2*, 81-92.

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34*, 27-44.

Kaminski, R. A., & Good, R. H. (1998). Assessing early literacy skills in a problem-solving model: Dynamic Indicators of Basic Early Literacy Skills. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 113-142). New York: Guilford.

Karlsen, B., Madden, R., & Gardner, E. F. (1976). Stanford Diagnostic Reading Tests. New York: Harcourt Brace Jovanovich.

Lee, L., & Canter, S. M. (1971). Developmental sentence scoring. *Journal of Speech and Hearing Disorders, 36,* 335-340.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1978). *Stanford Achievement Test.* New York: Harcourt Brace Jovanovich.

Malecki, C, & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40,* 379-390.

Markell, M. A., & Deno, S. L. (1997). Effects of increasing oral reading: Generalization across reading tasks. *Journal of Special Education, 31,* 233-250.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it. In M. Shinn (Ed.), *Curriculum – based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.

Marston, D., & Deno, S. L. (1982). *Implementation of direct and repeated measurement in the school setting* (Research Report No. 106). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Marston, D., Fuchs, L. S., & Deno, S. L. (1986). Measuring pupil progress: A comparison of standardized achievement tests and curriculum-related measures. *Diagnostique, 11*, 77-90.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33,* 193-203.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *Journal of Special Education, 41*, 68-84.

Mehrens, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools, 30,* 241-254.

Merriam-Webster Online Dictionary. Retrieved January 2009, from http://www.merriam-webster.com

National Association of State Directors of Special Education. (2005). *Response to intervention: Policy consideration and implementation.* Alexandria, VA: Author.

No Child Left Behind Act. (2004). Public Law 107-15.N

Northwest Regional Educational Laboratory. (2000). *6 + 1 Trait® writing: A model that works.* Greensboro, NC: Carson-Dellosa Publishing.

Norušis, M. J. (2008). *SPSS 16.0 statistical procedures companion.* Upper Saddle River, NJ: Prentice Hall.

Oregon Department of Education. (2000). Statewide assessment results 2000. Retrieved from the World Wide Web: http://www.ode.state.or.us/.

Parker, R., Tindal, G., & Hasbrouck, J. (1991a). Countable indicies of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2,* 1-17.

Parker, R,. Tindal, G., & Hasbrouck, J. (1991b.). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children, 58,* 61-73.

Powell-Smith, K. A., & Shinn, M. R. (2004). *Administration and scoring of written expression curriculum-based measurement (WE-CBM)for use in general outcome measurement.* Retrieved January 2009, from www.aimsweb.com

Rafoth, B. S., & Rubin, D. L. (1984). The impact of content and mechanics on judgements of writing quality. *Written Communications, 1*, 446-458.

Shapiro, E. S., Keller, M. A., Lutz, G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessments and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*, 19-35.

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children.* New York: Guilford Press.

Shinn, M. T. (1995). Identifying and defining academic problems: CBM screening and eligibility procedures. In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 90-129). York, NY: Guilford Press.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., and Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

Shinn, M. R., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *Remedial and Special Education, 6*, 31-45.

Shinn, M. R., Ysseldyke, J., Deno, S. L., & Tindal, J. (1982). A comparison of psychometric and functional differences between students labeled learning disabled and low achieving (Vol. IRLD-RR-71). University of Minnesota, Institute for Research on Learning Disabilities.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

The Great American English Handbook. (1987). Jacksonville, IL: Perma-Bound Books.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Tindal, G. (1989). Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In M. Shinn (Ed.), *Curriculum – based measurement: Assessing special children* (pp. 202-238). New York: Guilford Press.

Tindal, G., Fuchs, L., Fuchs, D., Shinn, M., Deno, S., & Germann, G. (1983). *The technical adequacy of a basal reading series mastery test: The Scott-Foresman reading program* (Technical Report No. 128). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tindal, G., Germann, G., & Deno, S. (1983). Descriptive research on the Pine Country Norms: A compilation of findings (Research Report No. 132). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Tindal, G., & Hasbrouck, J. (1991). Analyzing student writing to develop instructional strategies. *Learning Disabilities Research and Practice, 6*, 237-245.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Report No. 109). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education, 23*, 169-183.

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice, 6*, 211-218.

Tindal, G., Shinn, M., Fuchs, L. Fuchs, D., Deno, S., & Germann, G. (1983). *The technical adequacy of basal reading series mastery test* (Research Report No. 113). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

University of Oregon. (2003). *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS). Retrieved February 16, 2008, from http://dibels.uoregon.edu

University of Minnesota. (2005). *Procedures for scoring writing samples.* Retrieved January 2009, from http://www.progressmonitoring.net/pdf/RIPM_Writng_Scoring.pdf

Videen, J., Deno, S. L., & Marston, D. (1982). Correct word sequences: A valid Indicator of proficiency in written expression (Vol. IRLD-RR-84). Unviersity of Minnesota, Institute for Research on Learning Disabilities.

Yell, M. L., Deno, S. L., & Marsten, D. B. (1992). Barriers to implementing curriculum-based measurement. *Diagnostique, 18*, 99-112.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2207). Literature Synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*, 85-120.

Weissenburger, J. & Espin, C. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*, 153-169.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*, 85-104.

Woodcock, R. (1973). *Woodcock reading mastery tests manual* Circle Pines, MN: American Guidance Service.

VITA

Marie Henderson is currently a graduate student in the school psychology program at Louisiana State University. She received her Bachelor of Science degree in psychology form the Louisiana State University in May 2002 and her Master of Arts degree in school psychology from Louisiana State University in May 2005. Marie Henderson is a candidate for the degree of Doctor of Philosophy in school psychology at Louisiana State University.