Vocabulary matching: potential for a diagnostic performance indicator

Jodie Schraven
Louisiana State University and Agricultural and Mechanical College, jschra1@tigers.lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations
Part of the Education Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/2574

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
VOCABULARY MATCHING: POTENTIAL FOR A DIAGNOSTIC PERFORMANCE INDICATOR

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 School of Education

by

Jodie Schraven
B.S., State University of New York at Buffalo State College, 1999
MSW, State University of New York at University at Buffalo, 2002
Ed.S., Louisiana State University, 2010
May 2014
To all the students, clients, and children that I have helped and who have inspired me to become a special education teacher, social worker, or counselor. This dissertation is especially dedicated to Peter C. who, as a first grader in 2005 from Buffalo, New York, inspired me to seek a higher degree so that people would hopefully listen to what twice-exceptional children really need in public schools. I thank his father, Mr. C. for giving me permission to dedicate this to his son. Peter is doing very well in school today as a highly gifted student with Asperger’s Syndrome.

And to Autumn de Forest and her parents, Doug and Katherine de Forest, for giving me the opportunity to work with Autumn, I thank you. In just under two weeks, Autumn has enlightened me, made me continuously laugh every single day, and simply smile. Not only is Autumn an exceptional student and a child art prodigy, but she also has an empathy for the entire human race that is rarely seen even in adults. Thank you, Autumn, for telling me the second day we met that my high energy motivates you. You motivate all of us.
ACKNOWLEDGEMENTS

Thank you to Dr. Kenton Denny, my adviser, Dr. Paul Mooney, Dr. Kevin McCarter, Dr. Kristin Gansle, and Dr. Jennifer Jolly for being members of my dissertation committee. Thank you Dr. McCarter for your office hours when trying to help me understand multivariate statistics in my first semester at LSU. I was fortunate that you always told the Student Services staff I had testing accommodations and that I could take as much time as I needed to complete your exams. Thank you Dr. Gansle for always reminding me of all the updated APA changes, being funny, and for always being so understanding and diligent about helping me anytime that I was in town or home in Nevada, especially in the end. Thank you Dr. Jolly for helping me publish my first peer-reviewed journal publication as first author. You also influenced my more recent interests in teaching social studies. Because of you, I tier many lessons in my classroom.

I am especially grateful to you, Dr. Denny, for putting up with me for so long and for being so tactful in speaking to me when I did not grasp something and for understanding how I think and see things. PippsK, Pumpkin, and Paris Schraven send their best to you and Mrs. Denny. Dr. Mooney, I appreciate all of the opportunities that you have given me as a doctorate student. If not for you, I would only have one academic journal publication instead of five. To this day I train other teachers in my district and receive many compliments from district administrators and directors for my knowledge in assessment.

Dr. Matthew Leoni from UNR deserves recognition for checking in every few months to hoping that I would hurry and graduate. Thanks to you and your wife, Deb, for hosting me in Reno all those times just so I could focus away from home. I am glad we stayed in touch since my semester at UNR so long ago when I was your criminal justice student. I also appreciate
having Angela and Bob Elmore for pushing me for over half of my years in existence and demanding that I finally attend a graduation for the first time in my life.

The following should be recognized for their encouragement: members of WBR Pupil Appraisal Services, Ryan Reichart and Coach Joe Marciano from the Houston Texans, Susan Sernoe and Paul Feaker from CCSD SSSD, Trent Day and the entire administration from CHS and to administration from DHS, Cynthia McGriff, Lynn Duerr, Samantha Diaz, Maggie Abels, Mary Grover, Cheryle Gillmore, and members of the LVMC for not letting me bag peaks unless I ‘dissertated enough.’ Collin and Pippers Kamholz thank you for just being Collin and PippsK.

And, to my dear friend and former Disney World College Program roommate, Kasey Bains and the entire Bains family, because if not for them I would never have attended LSU in the first place.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2: REVIEW OF LITERATURE</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 3: METHODOLOGY</td>
<td>49</td>
</tr>
<tr>
<td>CHAPTER 4: RESULTS</td>
<td>67</td>
</tr>
<tr>
<td>CHAPTER 5: DISCUSSION</td>
<td>75</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>86</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>99</td>
</tr>
<tr>
<td>VITA</td>
<td>103</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

Table 1: 2 x 2 Standard Contingency Table.........................................................13

Table 2: Demographic Percentages for 6th Grade Students (n = 224)..........................50

Table 3: Descriptive Statistics for VM Scores and 6th Grade iLEAP Scores.....................68

Table 4: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores..................................................................................................................................................70

Table 5: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores..................................................................................................................................................70

Table 6: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores..................................................................................................................................................71

Table 7: Diagnostic Utility Statistics Calculation..................................................................................72

Table 8: Diagnostic Accuracy Across 6th Grade Students for Predicting Achievement.................74
LIST OF FIGURES

Figure 1: Vocabulary Matching Probe

Figure 2: Receiver Operating Characteristic Curves of the Model “Benchmark Probe Scores and Social Studies 2009 Pass”
ABSTRACT

Predictive validity was explored between vocabulary matching (VM) probe scores and the integrated Louisiana Education Assessment Program (iLEAP) social studies standardized subtest scores for 224 sixth-grade students across demographic categories from one rural southeastern Louisiana school district. Multiple Linear Regression (MLR) conducted in the Statistical Package for the Social Sciences (SPSS, 2007) produced Pearson correlations between .51 and .70 for five benchmark probes for the entire sample. Predictive cutscores created using Receiver Operating Characteristic (ROC) curve analyses and a diagnostic accuracy application software program illustrated that VM cutscores did not correctly predict which students would pass or fail with enough specificity (.48 and .69) or enough sensitivity (.61 and .73). Limitations and future implications are discussed.
CHAPTER 1: INTRODUCTION

In an era of significant education reform, a major policy initiative and two pieces of legislation have altered the context of schools by heavily influencing how to organize, plan, and evaluate the impact of our teaching efforts. As the emphasis and pressure on schools have migrated toward the outcomes of instructional endeavors, a shift from summative assessments to frequent formative assessments is receiving much needed attention. This attention has led to changes not only in how we organize schools for instruction, but also has highlighted areas of assessment in need of additional research.

In 1991, Congress revised the Elementary and Secondary Education Act (ESEA), renaming and greatly extending how states would be eligible for federal funding. Public Law 107-110, the No Child Left Behind Act of 2001 (NCLB), required that schools be held accountable for results in making Annual Yearly Progress (AYP), emphasized the use of empirically based instruction, mandated the use of highly qualified (HQ) teachers, and expanded options for parents who have children attending low-performing schools (Cusumano, 2007). The impact of NCLB was to encourage states to adopt sound instructional practices and to focus on assessment procedures to drive the instructional process. Schools often employed assessment procedures designed to reflect summative child performance that lacked the capability to inform instruction on a short-term basis. Many school districts began regularly assessing basic skills and encouraged teachers to monitor progress with shorter versions of the assessment that were comparable across time and that mirrored their high stakes summative tests (Dorn, 2010).

Following closely behind NCLB, the President’s Commission on Excellence in Special Education (PCESE, 2002) issued a report highlighting that the outcomes for students with disabilities were also unacceptable. Since the passage of the Education for all Handicapped
Children Act in 1975, the commission recognized that strides had been made in providing access to education for infants, toddlers, children, and youth with disabilities. However, the outcomes remained troubling. Students with disabilities were twice as likely to drop out of school, very unlikely to continue beyond mandated education age, likely to remain unemployed or underemployed, and were more prone to being under- or overrepresented by minority status (US Department of Education, 2002 - http://nces.ed.gov/pubs2002/2002205.pdf). The commission questioned not only the outcomes of those with disabilities, but also focused on evaluation procedures for identifying students as having disabilities.

Driven largely by increases in the number of students identified as disabled, the Commission focused specifically on the identification of students with specific learning disabilities (SLDs). As the largest single disability group, the process for identifying students as possessing SLDs has been historically problematic (Thurlow, Moen, Liu, Scullin, Hausmann, & Shyyan, 2009). Reliance on a discrepancy model, in which student achievement must fall significantly below tested ability for eligibility determination, may have resulted in waiting until students with academic performance issues displayed achievement that had deteriorated to levels that would support identification as having a disability. This “wait to fail” approach was criticized by the PCESE. The Commission recommended to the President that special education, as general education had been under NCLB, must reject the “wait to fail” model in evaluation and be held accountable for outcomes (PCESE, 2002).

The Individuals with Disabilities Education Improvement Act (IDEIA) of 2004 (P.L. 108-446) incorporated many of the PCESE recommendations in the re-authorization process. Two sections appear especially relevant to this discussion. First, under Early Intervening
Services the law allowed states (SEAs and LEAs) to set aside up to 15% of the funds under the act

“to develop and implement coordinated, early intervening services, which may include interagency financing structures, for students in kindergarten through grade 12 (with a particular emphasis on students in kindergarten through grade three) who are not currently identified as needing special education or related services, but who need additional academic and behavioral support to succeed in a general education environment.” [34 CFR 300.226(a)] [20 U.S.C. 1413(f)(1)]

The clear intent of Congress was to allow substantial funding to attempt to prevent or at least to lessen the impact of disabilities by financing and building structures to provide services to children as early as possible.

Second, IDEIA (2004) also provided the basis for alternative methods for identifying students suspected of having SLDs. The law required that states and districts could not compel the use of a severe discrepancy between intellectual ability and achievement for determining whether a child has a SLD (34 CFR 300.8(c)10). They allowed the use of identification procedures based upon a “child’s response to scientific, research based intervention” and “may permit the use of alternative research based procedures for determining whether a child has a specific learning disability” (34 CFR 300.8(c)10). To ensure that underachievement in a child suspected of having a SLD was not due to lack of appropriate instruction in reading or math, multi-disciplinary teams must have considered data that demonstrate that prior to, or as a part of, the referral process, the child was provided appropriate instruction in regular education settings, delivered by qualified personnel, and data-based documentation of repeated assessments of achievement at reasonable intervals, reflecting formal assessment of student progress during instruction (34 CFR 300.304 through 300.306).

Third, the legislative basis for assessment procedures that focus on documenting that students have been the recipients of competent and appropriate instruction and that assessment
must address student progress repeatedly under the context of this instruction became a basis for the Response to Intervention (RTI) model as a significant part of the identification process for SLDs (34 CFR 300.304 through 300.307). Concurrently, the requirement of NCLB for access to scientifically based instruction also helped to promote the model as a prevention of academic failure for students not necessarily suspected of being disabled.

Multiple tiers of intervention or support characterize RTI models. These tiers are intended to provide progressively intensive levels of instruction based on assessment data of student performance. Initially, a universal screening of student performance is used to verify that the majority of students have access to an effective curriculum. That is, the majority of students in the schools should be progressing at levels that reflect competent performance. For individuals who score below identified performance levels, the intensity of the instruction and the frequency of monitoring are increased. If a student does not make sufficient progress at one level, resources and expertise are increased at the next level with continued monitoring (Fuchs, Mock, Morgan, & Young, 2003). This process continues until a child responds to the interventions provided or until he or she is ultimately referred to special education. Cusumano (2007) states the benefits of this approach, in that many more students may respond and reach proficient levels of academic performance, should result in a decreasing number of students identified as having SLDs.

Although the IDEIA (2004) does not specifically identify an approved or recommended model for achieving RTI, the requirement for a scientific and/or research based process of identification led policy makers and school personnel to models created for improving the formative assessment process. Much of the recent empirical research had addressed models that were based on the actual curriculum being used in schools and procedures for the ongoing
monitoring of student progress within a curriculum. This curriculum based process relied heavily on the procedures of formative assessment (Fuchs & Fuchs, 2006; Fuchs, Fuchs, & Compton, 2012).

Another perplexing problem was the variation in curricula across districts and states. Policy makers and administrators continually scrutinized the outcomes of the public schools. The number of college remedial classes for incoming freshman is increasing as four out of ten entering students have to take remedial classes while students who take more rigorous high school courses are more likely to do well their freshman year (Beach, 2011). This raises the question as to whether or not high school students are mastering learning standards despite passing high stakes tests and earning high school course credit.

In response, the Common Core State Standards (CCSS) were created in 2010 so that public schools had a concise agenda for preparing children for post-secondary education and employment. Within these standards, it was proposed that consistent learning benchmarks across academic subjects be identified and aligned with college and work expectations. These standards were to be easily understood and rigorous in content and include the application of higher-order skills. It was suggested that these new curriculum standards would lead to secondary school graduates that not only were competitive across the various states, but also could compete on an international scale (www.corestandards.org). Coordinated through the National Governors Association Center for Best Practices and the Council of Chief State School Officers, 46 states plus the District of Columbia and outside territories have adopted CCSS with implementation led by each state rather than by educational mandates or policy makers not familiar with a state’s needs.
According to the CCSS website, educational standards aid teachers in ensuring that students have the skills and knowledge necessary to be successful by providing clear goals for student achievement. When students and parents know the standards, they can set clear and realistic goals for student learning (www.corestandards.org). The CCSS include components for the development and implementation of comprehensive assessment systems that measure student performance against the standards while replacing other widely used summative evaluation systems that are more costly, time-consuming, complicated, and are less able to address present student needs. CCSS may enable districts within each state to develop one common test by sharing costs and resources in developing valid assessments (www.corestandards.org).

It would appear from the previous sections that a confluence of legislative events, research findings, and policy decisions regarding content are driving changes in the educational system. Although not in the absence of detractors, there appears to be an increasing need to link content and instruction and to continually improve our assessment procedures to support these efforts.

**Academic Assessment**

As previously stated, both the President’s Commission (PCESE) and IDEIA (2004) were critical of current methods that relied upon a “wait to fail” model for identifying students with SLDs. Traditional methods of assessment of growth (i.e. grades, course completion) were not considered adequate for measuring student progress through the curriculum. Most schools traditionally use grading systems (i.e., report cards based on teacher-made tests and assignments) for determining students who pass or who do not master instructional content. Current letter and number grading systems have minimal validity for identifying at-risk students early enough or for predicting achievement on high stakes assessments mandated by NCLB (Bursuck, Munk, &
Olson, 1999; Munk & Bursuck, 2001; Silva, Munk, & Bursuck, 2005). These two needs: (1) to predict academic achievement, and (2) to appropriately identify students in need of intervention, highlighted the necessity of revised if not completely new systems of assessment.

Researchers and policy makers were then faced with the challenge of identifying measures that would allow screening for all students as they progressed through the curriculum, to provide different levels of intervention based on assessment data, to monitor student performance as those interventions were implemented, to adjust the level of intensity of instruction in response to student data (including referral for identification), and to deliver school based data for program evaluation purposes. A foundational component of this effort would be to determine the unit of measurement (the outcome) that would allow schools to develop and implement these procedures.

The identification of the measures that could be applied across multiple levels of curriculum within a subject area (e.g., reading and math) has proven to be challenging. Beginning reading has multiple facets (e.g., alphabetic principle, phonemic awareness, comprehension, etc.), and researchers have identified measures (e.g. Deno & Mirkin, 1977; Fuchs & Deno, 1991) that had wide utility, informed instruction, and predicted desired academic outcomes. The measures could be used to track student progress as they were provided instruction and often demonstrated the ability to predict future student performance (Shinn, 1989). For example, the National Reading Panel (National Institute of Child Health and Human Development, 2000) identified five critical areas of reading development that predicted reading abilities in later grades. One of the conclusions of the National Reading Panel, based on a comprehensive review of the empirical literature, was that fluency in these identified components would consistently predict later reading abilities (Cooper & Hedges, 1994; Snow,
Burns, & Griffin, 1998). Use of these repeated measures allows teachers to experiment with instructional pieces, sequences, or materials and examine how their instruction affects students’ performance in broader content. Procedures based on mastery measurement of curriculum (skills, skill sets, and units) can also help determine the content and sequence of instruction but have been criticized for lacking the sensitivity of fluency based measures (Fuchs & Deno, 1991). There remains an ongoing need to identify the critical areas of content and to develop measures of these critical skills (predictor variables) that allow the ongoing and valid measure of student progress and ultimately proficiency (criterion variables). To date, the most developed form of these measures is curriculum based measurement (CBM).

Curriculum Based Measurements

An extensive body of literature has discussed CBM for over four decades. Deno (1985) described CBM as simple procedures for measuring student growth within academic skill areas. Espin and colleagues (2000) describe the critical components of CBM as direct and frequent measurement of student growth using indicators of performance. Although indicators have long existed in other areas such as economics (e.g. employment figures; Gross National Product), indicators in education have often failed to meet criteria offered by Rockwell (1989) that allow them to function as indicators. These include the timeliness of reporting, a future orientation, potential to facilitate decision-making, collectability across time, accuracy, reliability, and validity (Rockwell, 1989). Curriculum based measurements incorporates three key components including test stimuli drawn from students’ curricula; repeated testings that occur across time; and that assessment information is used to formulate instructional decisions (Fuchs & Deno, 1991; Tucker, 1987). Curriculum based measurements provide valuable samples of student performance for teachers because they are valid, efficient, and easily understood inexpensive.
formative tests (Fuchs, 1985). Curriculum based measurement research clearly indicates a systematic process through which the development and implementation of CBM procedures have been established as acceptable indicators of the teaching and learning process. The identification of what should be measured, as an indicator, was one of the greatest challenges in the development of CBM (Deno & Fuchs, 1987).

In their review of 585 CBM studies, Espin and Wallace (2004) found that 305 were published in peer-reviewed journals, of which 141 included empirical studies. Curriculum based measurement’s design, purpose in progress monitoring procedures, its focus on the SLD classification, and its use within an RTI framework were discussed in depth. Curriculum based measurement research has been so positive in terms of progress monitoring that the Office of Special Education Programs (OSEP) funded the Research Institute on Progress Monitoring (RIPM) so that researchers could develop a “seamless and flexible system of progress monitoring” (p. 66). The intent was to create a system that could assess all students within different settings, accessing curricula in which valid tools of measurement could provide data useful to teachers (Wallace, Espin, McMaster, Deno, & Foegen, 2007).

Although it is recognized that the progress in establishing CBM as an indicator of educational performance has been substantial, there remains a need to extend the process of validation to incorporate a broader range of content areas across age groups and grade levels (Espin et al., 2013; Vannest, Parker, & Dyer, 2013). This expansion from basic skill assessment to content area knowledge assessment has posed additional challenges in identifying the general outcome measure (GOM) that would allow the CBM assessment and progress monitoring of more complex content. One such area of research has been the development and validation of
measures based on the language/vocabulary of content areas (Mooney, McCarter, Schraven, & Callicoatte, 2013).

**Vocabulary Matching**

Content-specific vocabulary words are the basic building blocks of a semantic network (Vannest, Parker, & Dyer, 2011) that students may need in order to fully comprehend written text. Following the same basic process for developing curriculum based measures, key vocabulary from the content area is identified and vocabulary probes are developed to reflect the language of the content area across an instructional period (e.g. school year). Probes are designed to be administered in just a few minutes, have demonstrated reliability and validity with other criterion measures, are easy to implement, and are efficient (Borsuk, 2010). Moderate to strong correlations between five-minute vocabulary matching (VM) probe scores in social studies and state standardized assessments have been produced (Mooney et al., 2010). Espin, Shinn, and Busch (2005), Borsuk (2010), and Mooney et al. (2013b) used hierarchal linear modeling (HLM) to demonstrate that a VM probe given over time illustrated student growth in middle school content courses. Vocabulary matching probes are measures that have potential for meeting all of CBM’s criteria for older students, even in terms of progress monitoring (Borsuk, 2010; Espin, Shin, & Busch, 2005).

Although progress has been made establishing VM as a valid measure of student performance, the need to identify levels of performance with cutscores to categorize which students need additional assistance to reach an acceptable proficiency level in terms of the standardized tests that describe their academic progress remains. This research is in the early stages and is directed toward the identification and validation of cutscores that would help
predict the progress toward criterion variables such as content mastery and/or performance on high stakes testing.

**Progress Monitoring and Cutscore Development**

A critical component of the CBM model is the process for monitoring performance over time. Several prerequisites must be present for this component to be valid and useful. First, the items within probes must reflect the curriculum. Second, these measures should be selected as samples of the entire length of instruction period (e.g. across the academic year or period of instruction). Third, there should be probes sufficient to allow for the ongoing measurement of performance without significant learning from testing occurring. Finally, the probes must be developed with equivalent difficulty levels. When these conditions are met, the frequent assessment of students should provide an ongoing illustration of student progress across content areas. This information may be graphically presented to allow for easy monitoring of learning across time and assists in monitoring not only ongoing student performance, but also allows for the comparison of performances within or across classrooms. Research has supported the use of these procedures in providing teachers with the information to make correct instructional decisions (Fuchs & Fuchs, 1986).

Following the development of an instrument to measure student progress across time, there is a need to calibrate the instrument to the specific curriculum. The process of calibration is quite simply determining at what point and to what degree the predictor variables (scores on probes) will actually predict the criterion variable (the identified outcomes of the instructional process) to an acceptable degree of accuracy. These levels or points are often referred to as “cutscores.” Silberglitt and Hintze (2005) define cutscores as “essential measures of proficiency that can be used to make data-driven determinations as to whether or not a student is on his or
her way to becoming a proficient reader” (p. 6). The precision of cutscore predictability that increases sensitivity and specificity to local student populations and state tests has been statistically developed with multiple regression, logistical regression, Excel diagnostic efficiency, and receiver operating characteristic (ROC) curve analyses (Hintze & Silberglitt, 2005; LeBlanc, Dufere, & McDougal, 2012; Mooney et al., 2008; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Silberglitt & Hintze, 2005). Curriculum based measures that have cutscores are a wise investment for predicting which students will fail or pass high-stakes assessments for schools that are expected to make AYP under NCLB. In terms of VM probes, they are efficient, easy-to-administer measures that are formative in nature, assess higher level content, and provide important predictive information about the probability that students will pass summative assessments for their grade levels.

**Receiver Operating Characteristic Curve Analyses**

Receiver Operating Characteristic statistical analyses were first used in Great Britain during WWII for radar detection, and later in signal detection theory, machine learning, mining research, eventually in psychology, and most often are used in medicine for disease prevention and in the evaluation of diagnostic tests (University of Georgia, n.d.). Quite simply, the ROC analysis attempts to identify the threshold at which accuracy is sufficient for prediction the future event or criterion variable. In theory, ROC performance reflects a curve to show how accuracy increases consistently across some measure toward a point where accuracy is either sufficient or maximized in its ability to predict a future event. A current example of the system could be the cellular phone system. When we are close to the source of a signal (cell) we would expect to have a clear signal, free of static and a small likelihood of dropped calls (better prediction). As we move away from the signal source (lose bars) the “noise” becomes greater and we are more
likely to experience “dropped” calls and poor call quality (lower level of accurate prediction and errors). The ROC attempts to quantify how we can move from a point (cutscore) toward a “maximized” quality of signal (aka: clear and accurate prediction). It does so by categorizing the type and accuracy of signals we receive (true/false) based on our initial point of prediction.

Table 1 illustrates the principle applied to the diagnosis of a disorder. We diagnose the future occurrence of a disorder based on some diagnostic measure(s). If our test indicates the presence of the disorder and the disorder actually exists, we would classify that as a true positive case. If however our test indicates that the disorder should be present and the disorder is actually not present we would classify that case as a false positive. In a third category, we indicate that the disorder should not be present and it actually is present, so we would categorize the case as a false negative. Finally, we could have cases where we diagnosed the absence of the disorder and the disorder is actually absent and we would classify these cases as true negatives. These findings are most often illustrated in a 2 X 2 standard contingency table (Table 1).

Table 1: 2 x 2 Standard Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Disorder Present</th>
<th>Disorder Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Positive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of cases</td>
<td># of cases</td>
<td># of cases</td>
<td></td>
</tr>
<tr>
<td>(True Positive)</td>
<td>(False Positive)</td>
<td>(False Positive)</td>
<td></td>
</tr>
<tr>
<td><strong>Test Negative</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of cases</td>
<td># of cases</td>
<td># of cases</td>
<td></td>
</tr>
<tr>
<td>(False Negative)</td>
<td>(True Negative)</td>
<td>(True Negative)</td>
<td></td>
</tr>
<tr>
<td><strong>Total disorder present</strong></td>
<td><strong>Total disorder absent</strong></td>
<td><strong>Total sample</strong></td>
<td></td>
</tr>
</tbody>
</table>
This form of analysis allows us to determine the specificity, sensitivity and accuracy of our measures from various initial positions (prediction points). This model will be discussed in greater detail in Chapter 2.

Receiver operating characteristic curve analysis has several advantages for creating achievement cutscores. There is no need for assumptions of normal distribution, multiple predictors can be evaluated simultaneously, results indicate interactions between a predictor and cut-points on those predictors, and ROC yields clinically relevant information. Disadvantages of ROC are that analyses require large sample sizes and capitalize on chance; a stringent stopping rule is needed to lessen the effects. The Area Under the Curve (AUC) produced by ROC considers only ranks of the scores (i.e. order) while ignoring specific scores (i.e. posterior probabilities of the positive class). When this information is ignored, suboptimal results, such as overfitting the test set when selecting classifiers with high AUCs, will produce suboptimal results (Majnik & Bosnić, 2011).

One critical component of successful RTI programs is standardized implementation of valid indicators of performance, or cutscores, that can predict performance. Cutscores that accurately predict students’ pass and fail rates with ORF and math computation probes have been created for a wide range of students throughout the country (Good, Simmons, & Kame’enui, 2001; Hasbrouck & Tindal, 1992; Hintze & Silberglitt, 2005; Mooney, McCarter, Schraven, Hintze, Mooney, Landry et al., 2008).

Most of the scholarly CBM research to date is centered on the ability to assess for basic skill development in both reading and in math for younger students. As students enter middle school, teachers remain concerned that their at-risk students may not be able to apply these basic elementary skills to content area courses. Such data can help teachers respond to older students
who are having difficulty prior to end-of-year accountability tests encouraged by special
education and general education mandates (Jiban & Deno, 2007).

The rationale for this study was to find a valid formative assessment for content area
teachers in middle school grades (predictor variable) that could predict achievement (criterion
variable). Specifically, the purpose of this study was to extend the existing literature base by
examining whether cutscores for VM could be developed with ROC that would predict
achievement on a statewide criterion referenced accountability instrument in social studies.
CHAPTER 2: REVIEW OF LITERATURE

In education, assessment is a process of collecting information for the purpose of making important decisions about students (Salvia, Ysseldyke, & Bolt, 2012). Evaluation is the systematic determination of merit, worth, and significance of a learning or training process by using criteria against a set of standards (Ysseldyke & Thurlow, 2012). In simple terms, assessment is the collection process for gaining valid information to inform evaluation decisions. Cronbach (1969) allegedly was the first to suggest the term “formative assessment” to refer to measures taking during the course of program development to determine the status of efforts to date and determine potential changes required to procedures or outcomes. Summative assessment refers to measures taken upon completion of a program to determine the “merit, worth, or significance” (Ysseldyke & Thurlow, 2012) of a program. From Cronbach (1969) to Salvia, Ysseldyke, and Bolt (2012), the extension of formative assessment and summative evaluation from program level to instructional level has progressed and matured as a science of assessment.

In the following sections, three areas of research will be addressed. First, key concepts related to formative and summative assessment and evaluation will be defined. Second, the research on the establishment of the measurement characteristics of CBM will be reviewed. Finally, the background on the extension of CBM to areas beyond basic reading, writing, and math to content area subjects will be presented as the basis for the current research. Specifically, the identification of language based measures as general outcome measures will be proposed and the procedures for establishing these procedures will be addressed.
**Formative and Summative Assessment and Evaluation**

Formative assessments are designed to identify student needs and make data-driven decisions (Salvia & Ysseldyke, 2007) for the purpose of providing valid modifications to improve learning. Formative assessments can guide instruction toward particular objectives and confirm student achievement (Tawney & Gast, 1984). Summative assessments are designed to document the effectiveness or outcome of an instructional unit (e.g. unit, semester or academic year) or more recently, the effectiveness of the individual providing the instruction (c.f. Marzano, 2011). It is generally recognized that summative assessment measures offer little in the ongoing modification of instruction to improve learning (Deno, 1985; Salvia, Ysseldyke, & Bolt, 2012; Tawney & Gast, 1984). The basis for the theory and practice of formative assessment has various roots. Although Cronbach may be credited with the terminology, the procedures for the ongoing evaluation of learning have their basis in the early behavioral sciences of the 1940s, 1950s and 1960s. Sidman (1960) chronicled not only the basis for behavioral learning theory but contributed the fundamental characteristics of measurements drawn from the emerging experimental analysis of behavior, and frequent and direct measurement of learning outcomes is a fundamental practice in behavior analysis (Skinner, 1953). Ideally, measurement should be directly focused on the outcome of the intervention (behavior) and be as low inference (direct) as possible. Measurement should occur as frequently as the intervention is applied to capture fluctuations as well as to study transition states (learning). Rate was the preferred dimension of recording behavior, providing not only the frequency but the time dimension of learning and allowing for cross analysis between experiments and enhanced sensitivity. The more direct and continuous the measure is, the lower the probability of error in decision making or need for extrapolation or interpretation. Sidman’s
research showed that learning could be measured over time, with sensitivity to small changes in the independent variable, and that graphic representation offers unique “pictures” of the learning process. Sidman’s contribution has remained one of the timeless classics of learning theory, measurement, and single case design remaining relevant for over five decades.

The contribution of the science of the experimental analysis of behavior to methods of teaching and measurement-based decision-making are most evident in two of the most empirically validated instructional programs in American education. Direct Instruction (Becker & Carnine, 1981; Bereiter & Engelmann, 1966) and Precision Teaching (Lindsley, 1971) had been in development for several years prior to being submitted as components in the largest educational research study in the history of American education. Project Follow Through (Bock, Stebbins & Proper, 1977) was a multi-year study focusing on the impact of various educational models of the achievement and development of at-risk children and youth. Direct Instruction was compared with 12 other models of instruction in research sponsored by the U.S. Department of Education and conducted by the Stanford Research Institute / Abt Associates (Bock, Stebbins, & Proper, 1977; Watkins, 1997). Project Follow Through ran from 1967 to 1975 with program funding continuing until 1995. In its early years, 75,000 children per year in 120 communities participated. The other models included the Behavior Analysis Model, the Florida Parent Education Model, and several constructivist/progressivist models that were language-oriented, "student-centered," and cognitive-developmental—including the High/Scope cognitive curriculum, the Bank Street College Model, Open Education, Responsive Education, and the Tucson Early Education Model. Scores on the Metropolitan Achievement Test, the Coopersmith Self-Esteem Inventory, and the Intellectual Achievement Responsibility Scale, showed that Direct Instruction was superior both to control schools and to every other model in fostering
basic reading and math skills, higher-order cognitive-conceptual skills, and even self-esteem (Adams & Engelmann, 1996; Becker & Carnine, 1981; Bock, Stebbins & Proper, 1977).

Ogden Lindsley developed precision teaching with his associates, who have kept records on over a half a million precision teaching programs from studies conducted through the years (Binder, Haughton, & Van Eyk, 1990; Haughton, 1980; Johnson & Layng, 1996; Lindsley, 1993; Potts, Eshleman, & Cooper, 1993). Lindsley based Precision Teaching on Skinner's suggestion that the rate of a behavior (the number of occurrences per unit of time) is a dimension of the behavior, and not just a measure of that behavior. This implies that fluent (automatic, effortless, fast, and accurate) behavior differs fundamentally from behavior that is not fluent.

Direct Instruction and Precision Teaching tightly link instruction with ongoing assessment. Assessment occurs during instruction as the basis for determining how students respond to the instructional procedures. Students who do not respond as expected provide information as to how instruction should be modified. This link of performance to instruction is based on decision rules. For example, in Precision Teaching, students who fail to maintain a predicted rate of performance for three consecutive periods should have their programs of instruction altered. Low overall rates of performance (both corrects and errors) would suggest a change in reinforcement. More erratic patterns of correct and error performance would indicate a faulty algorithm, rule, or strategy for responding. The instructional program would then be altered to reteach focusing on the explicit teaching of the correct approach. Both Precision Teaching and Direct Instruction connect movement through a curriculum (content) to demonstrated mastery (fluent and automatic performance) of the information and/or skills being instructed. Lindsley (1993) opined

“If a student is progressing according to plan, the program is good for the student. If not, the program is flawed, and needs to be changed; therefore, there is no
failure by the student as a product of the student but rather as a product of the teaching.”

Given the conscious linking of assessment and instruction, these approaches are some of the earliest demonstrations of the power of formative assessment with data driven decision-making in education (c.f. (Darch, Gersten, & Taylor, 1987; Gersten, Keating & Becker, 1988; Meyer, 1984; Meyer, Gersten, & Gutkin, 1983). The components of what would become known as curriculum-based measurement (Deno & Mirkin, 1977; Deno, 1985) are deeply steeped in the foundations of these two instructional approaches.

It is not surprising that many education researchers, given the data on the impact of frequent assessment on academic achievement, are critical of current and historical assessment practices in education (Bursuck et al., 1996). Most notable are the criticisms of the use of “report card” grades as indicators of academic performance and progress. Test scores, end-of-unit tests, and report cards have a long history in schools. However, they present many problems in their use as indicators of progress or learning. Bursuck and colleagues criticized the use of unit tests or chapter quizzes since they are only rough estimates of student learning, having neither the reliability nor validity to accurately assess growth. Numeric homework grades may reflect student motivation, the ability to use a textbook and take notes, teacher impressions of student behavior or learning, or actual learning of content. Report card grades based on either alphabetic or numeric scales have similar weaknesses.

Many forms of current grading rely on the use of teacher made tests, quizzes, and assignments that are intended to be criterion referenced. These grades, however, lack empirical evidence for determining student mastery (Fuchs & Deno, 1991). Correlations between course grades and the Iowa Test of Basic Skills (ITBS) scores were only .57, which was not surprising considering course grades may include factors such as on-task completion and behavior (Espin et
Assessments based on teacher judgment are even more ambiguous and unreliable in terms of predicting future performance. Furthermore, the impact of various types of grading systems on students in special education is unknown (Munk & Bursuck, 2001). Confusion exists for establishing grading criteria and assigning grades between general and special education teachers (Bursuck et al., 1996). Regular and special educators are often at odds as to what degree content expectations may be altered and still are able to balance the (limited) validity of a grading system with the mandate to adapt and modify instruction to meet the needs of the individual student. Clearly, many assessment practices that are intended to be formative are problematic and are in need of substantial revision or replacement in light of research on more useful procedures to support formative assessment.

Vast amounts of money have been poured into the development of summative standardized achievement tests in order to qualify students for special education (Idol, 1986; 1996). Many earlier assessments were expensive, did not pertain to subjects being taught in the classroom, and were criticized for being used to label groups of students (Cohen & Spruill, 1990; Idol, 1986; 1996). These tests lacked both the ability to impact educational programming and adequate psychometric properties (Salvia & Ysseldyke, 1985; Ysseldyke, 1979). Fuchs and Fuchs (1986) criticized that summative assessments fail to demonstrate learned objectives and proposed formative measures as the desired alternative for driving instructional decisions.

Benefits for conducting formative assessments can be seen from multiple avenues. From an educational perspective, students with disabilities whose programming is ‘monitored systematically and developed formatively over time achieve, on average, .7 standard deviation units higher’ (p. 205) than those students who are not monitored (Fuchs & Fuchs, 1986). Teachers who are trained in formative assessment activities are more likely to make frequent
instructional changes as graphic feedback enhances correct decision-making (Fuchs, Deno & Mirkin, 1984; Fuchs & Fuchs, 1984; 1986; 1987). From a political standpoint, the President’s Commission on Excellence in Special Education expresses that answers to increasing learning stem from data that is put to immediate use by teachers. Teachers would be wise to create objective and measurable IEP goals not only to monitor student learning, but also to meet the requirements of IEP design (see Cobb County School District v. Kristen B., 1997).

With increases in accountability nationally and locally, the pressure on schools to use required state-developed summative assessments to determine student proficiency has increased.

“Given the high-stakes nature of state tests for schools in terms of meeting NCLB standards, and for students who are required to pass reading tests to graduate, the tests are an important outcome for students and schools at the secondary-level.” (Espin et al., 2010; p. 61)

In response, valid formative assessments that can be administered frequently, are efficient, and are instructionally pertinent to continuous data collection in combination with summative tests have been developed (Jiban & Deno, 2007; Mooney et al., 2008). However, formative assessment’s ability to predict false-positives and false-negatives has been questioned in the research due to the high costs of providing tiered levels of service during different RTI screening procedures (Fuchs, Fuchs, & Compton, 2012). Teachers need such information to determine how students are progressing as they infer learning based on perceived changes in student behavior. More reliable than present day grading practices, formative alternatives like curriculum based measures that evaluate student learning over time can provide the primary component of an RTI program (Fuchs, Fuchs, & Compton, 2012).

Formative assessment models have primarily addressed students’ needs in basic reading and math skill acquisition. At the middle and high school levels, RTI’s focus must not only address basic skills, but the application of those skills within content courses, as the bulk of
students should have mastered the basic skills and instruction has shifted focus to content. In order for formative assessment to be effectively used in this model, the assessment must address content, an area that is receiving increasing attention in areas such as math/algebra (Foegen & Deno, 2001), written expression (Amato & Watkins, 2011; Lopez & Thompson, 2011; McMasters & Espin, 2007), and social sciences (Mooney, et al., 2013).

**Curriculum Based Measures**

Curriculum based measures were originally designed to be part of a problem-solving approach to address academic weaknesses by focusing on specific academic performance variables to describe student growth or lack of growth in a numeric format that would lend itself to intervention rather than simply focusing on the child or characteristics of the child as the primary contributing negative factor to achievement. Curriculum based measures are closely related to present instructional level and are sensitive to growth, making for useful formative assessment or progress-monitoring tools to determine if students are benefiting from instruction (Wayman et al., 2007). Curriculum based measurement is a set of standardized steps of measurement that can quantify student performance in basic skill areas. It is drawn from the curriculum in which students are currently instructed. Although fluency is the primary metric of interest, additional areas such as accuracy may also be measured.

Curriculum based measurement data provide general and special education teachers with common metrics that allow for determination of performance differences both between students and within students over time (Ardoin et al., 2004). A benchmark CBM can be typically given three times a year, but a strategic CBM is given to at-risk students multiple times a month for measuring response to intervention. Teachers who use a simple set of procedures to monitor progress and who critique and modify their instructional procedures based thereon positively
impact achievement compared to teachers who do not strategically evaluate their teaching (Espin & Foegen, 1996; Fuchs, Deno, & Mirkin, 1984). Curriculum based measurement is generally used to assess progress in reading, mathematics, and written language, with standardized implementation and scoring procedures that have a large body of research attesting to their validity and reliability (Marston, 1989; Reschly et al., 2009; Wayman et al., 2007), especially for the purposes of screening and benchmarking (Ardoin et al., 2013). Using CBM, schools can potentially meet the demands of current education policies centered on accountability as well as monitor students’ attainment of CCSS.

Curriculum based measurement for basic skills is commonly administered across reading, mathematics, and written language, especially in the elementary grades. However, an ongoing need to assess critical areas of content and develop measures of these skills for more advanced students is clear. These measures should allow the ongoing and valid measurement of student progress and ultimately proficiency in those content areas. Given that reading is a critical skill for gaining knowledge for middle and high school students, the literature on CBM in different areas with various groups of students will be reviewed.

**CBM in Reading**

At the elementary level, CBM has been used as a component of effective classroom interventions tailored to the specific needs of students (Fuchs et al., 1997). Curriculum based measurement is flexible enough to provide a basis for establishing a progress monitoring system that can be used across diverse groups of students at different ages and reading abilities (Wayman et al., 2007). In terms of accountability, CBM has allowed data to be aggregated across entire districts to make district-level evaluation decisions (Marston & Magnusson, 1998). Student learning can be assessed with measures across various curricula without negatively
impacting technical adequacy. The three most common measures used to track reading progress are reading aloud, maze (cloze) tasks, and word identification (Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

**Oral reading fluency.** Initially, the most widely used classroom-based reading assessment used by teachers was Informal Reading Inventories (IRI; Parker, Hasbrouck, & Tindal, 1992). Students would read aloud within graded passages allowing the teacher to judge a range of reading skills and overall performance. Similarly, tests or probes of oral reading fluency (ORF) were those in which students read from graded passages and teachers recorded the fluency at which the student read the passages (correct words per minute; CWPM). There is support for the reliability and validity of ORF as a measure of reading progress (Wayman et al., 2007; Marston, 1989) and to monitor reading progress over time (e.g., Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Deno, & Mirkin, 1984; Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005). While extensively applied in elementary settings, ORF has less research support for addressing academic progress in the middle and secondary grades (Silberglitt, et al., 2006).

**Maze.** One way in which comprehension has been assessed is by using short fill-in-the-blank reading tasks. Cloze tasks were validated for determining readability of a passage, assessing general reading achievement, diagnosing reading problems, and in measuring listening skills (Evans & Balance, 2001). The first cloze procedure required a student to silently read a passage that contained blanks in sentences. When the child came to a blank, he/she was expected to write in a word that made the sentence make the most sense (Ardoin et al., 2004). Between the early 1950s and the turn of the century, cloze was well supported by its correlations with other reading assessments (Greene, 2001). However, predicting achievement depended on the students’ levels of sentence comprehension and writing skill, scoring was time-consuming,
and lower-level readers found the cloze frustrating (Parker et al., 1992). In the 1970s, the initial untimed maze tasks were developed to improve upon the classic cloze procedure for students who were economically disadvantaged, ELLs, or who were demonstrating significant reading problems (Parker et al., 1992).

Curriculum based measurement read-aloud and a CBM-maze task have produced high correlations between .70 and .80 with 8th grade high-stakes tests (Espin, Wallace, Lembek, Campbell, & Long, 2010; Muyskens & Marston, 2006). Silberglitt, Burns, Madyn, and Lail (2006) report moderate to strong correlations between this CBM and a state standardized reading test for 7th and 8th grade students. The first reported content-area CBM tested for comprehension and for the application of critical thinking skills (Nolet & Tindal, 1994; Tindal & Nolet, 1995). For measuring the development of abstract conceptual ideas, concept-mazes were used to assess middle school students’ abilities to generalize comprehension and vocabulary knowledge skills (Twyman & Tindal, 2007). Later, VM probes were used to measure vocabulary attainment related to course content for older students.

The maze task was easier to design and administer, less time-consuming to score, and more sensitive to growth than oral and written retell of stories and cloze passages for measuring comprehension (Fuchs & Fuchs, 1992). Similar to a cloze, the maze task required students to circle the correct choice to make the sentence grammatically and synthetically correct (Guthrie, 1973). The number of items per maze and the required time for scoring depended on passage length and word deletion ratios, most ranging from 125 to 400 words. The maze was constructed by deleting every 5th or 7th word and replacing it with a blank while offering students three possible choices to pick the correct word that made the sentence make the most sense. Researchers later incorporated speed as a factor by setting stringent time limits (1.5 to 2.5
minutes) on 60-item maze tasks that improved the reliability in discriminating among students (Fuchs et al., 1990; Jenkins, Pious, & Jewell, 1990). Current assessment conventions dictate that students are given one and a half to three minutes to complete a 60-word maze task (Shinn, Deno, & Espin, 2000).

Researchers have demonstrated that fluency was part of comprehension ability as rate-based scores produced higher correlations with other criterion reading tests. The validity of the maze measure has been supported with a Gates-MacGinitie Reading Test (GMRT) standardized reading vocabulary test and a reading comprehension test (.85; .82 respectively; Guthrie, 1973).

**Reading comprehension.** In addressing the assumption that standardized reading tests and measures of ORF do not provide accurate pictures of students’ true reading comprehension abilities, which are especially important as students advance to middle and high school, Dupuis (1980) had 212 10th grade students read two short passages and one novel using a pretest cloze procedure and a post-test multiple-choice comprehension test. Dupuis (1980) suggested that cloze was one means for matching student reading levels to appropriate literary selections for passages but that the maze provided teachers with more information because it took less time, feedback was quickly provided, strategy intervention could have been provided thereafter, was easy for teachers to administer, and more accurately portrayed the student’s processing of text. At a middle school level, maze tasks appear to have met the components of a CBM.

Espin et al. (2010) investigated the technical adequacy of a reading aloud and a three-minute maze task as performance indicators on a state-standardized reading test. The reading aloud and the maze measures were reliable and valid (.93 to .96; .79 to .96 respectively), similar to findings in elementary school research. When they investigated for adequacy as progress-monitoring measures, only the maze task resulted in substantial growth over time and the growth
rates for the maze were significantly related to performance on the state test (Espin et al., 2010). Up to this point, middle school CBM research still only focused on basic comprehension skills and not necessarily on the comprehension or application of content knowledge.

Addressing the content necessary to succeed in middle and high school was a logical next step for CBM. Curriculum based measurement for that content can be developed after reviewing the standards and curriculum, organizing that content into critical knowledge protocols that focus on declarative, procedural, or conditional knowledge, and designing content assessments that directly measure what is being taught (Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006). An analysis of the content for relevant and adequate representation needs to be encompassed (Ketterlin-Geller et al., 2006) since most high-level learning is based on the application of basic skills for understanding concepts.

Tindal and Nolet (1995) define concepts as: “part of a taxonomy of increasingly complex knowledge forms that consist of facts, concepts, and principles” (p. 5). Ketterlin-Geller et al., (2006) defined domain-specific concepts as:

“facts, or one-to-one relationships between names, objects, places, or events, are the simplest forms of knowledge; one to two word abstractions that share a common set of defining characteristics or attributes to which factual examples may be applied. Facts are critical for building in-depth conceptual knowledge in that they provide the example set of any given concept” (p. 43).

Students focus their reading on examples of attributes, ignoring those that do not belong while applying critical thinking skills to complex material (Ketterlin-Geller et al., 2006).

**Concept maze.** Researchers have examined the validity of a “concept maze” for predicting student performance and monitoring progress over time for middle school students (Ketterlin-Geller et al., 2006; Twyman & Tindal, 2007). The concept maze is similar to a traditional maze measure but its text is taken directly from the content course materials (e.g.
social studies, science). The first concept-maze was created by rewriting 6th grade topic-related textbook sections into a series of maze passages with exemplary words and phrases and 8-12 missing words were replaced with four possible choices. Six concept mazes were administered over a four-week period after the teacher taught the corresponding lesson but no conclusive data were drawn (Ketterlin et al., 2006). In a second study where missing words were replaced with attribute word choices, the concept maze had weaker consistency in readability with low alternate-form reliability (.38, .48, and .58 respectively; Twyman & Tindal, 2007). Johnson, Semmelroth, Allison, and Fritsch (2013) administered science content maze passages to 367 middle school students across three states to determine whether maze passages had sufficient reliability and validity to serve as benchmarking tools. Using multiple maze passages and state level tests in science, they found alternate form reliability ranging from .56 to .80 and concurrent predictive correlations ranging from .63 to .67 (Johnson et al., 2013).

**Content Area Assessment through General Outcome Measures**

A General Outcome Measure (GOM) is a formative assessment that is proficient on a global outcome from which teachers direct their entire curriculum (Hintze, Christ, & Methe, 2006). General outcome measures are simple sets of procedures that can assist teachers in planning, adapting, individualizing, and evaluating instructional programs for students (Deno, 1985; Deno & Fuchs, 1987; Fuchs & Deno, 1991; & Shinn, 1989). Instead of being summative in nature, GOMs are direct repeated measurements of student progress toward long-range instructional goals. Compared to earlier subskill mastery measurements (like ORF), GOMs assess proficiency across hierarchies within a curriculum (Espin & Deno, 1994-95; Fuchs & Deno, 1994; Hintze et al., 2006). They are cost and time-effective because they can be repeatedly group or individually-administered in just a few minutes and cost significantly less
than standardized achievement tests that generally require administration by a trained professional (e.g. psychologist; Espin & Foegen, 1996). General outcome measures have stronger predictive validity for long-term learning than summative assessments because they are designed to immediately provide valid feedback to teachers so that at-risk students can be identified before failing end-of-the-year tests in order to meet demands of education mandates and social expectations.

Research has focused primarily on elementary students’ learning using GOMs. It has only been studied as a performance measure and not as a growth measure in the secondary grades (Borsuk, 2010; Espin et al., 2010; Mooney et al., 2013a). The validity and reliability of a GOM may differ across subjects, teachers, and grades depending on practitioners’ uses (Espin et al., 2010). What might be an appropriate progress-monitoring tool for some students may not be sensitive enough for students who learn at different rates. For example, 30% of low performing students had science maze probe scores that showed inadequate sensitivity and 25% showed no improvement over five and six weeks (Vannest, Parker, & Dyer, 2011). Such data may not provide teachers with the best information to make accurate decisions regarding instructional changes. Although the research on measuring progress for older students is increasing, predicting and monitoring achievement in subjects outside of reading and writing is only a recent phenomenon.

“For an assessment tool to be considered a valid indicator of performance, evidence must demonstrate that performance on the measure relates to performance in the academic domain more broadly.” (Espin et al., 2010; p. 61) Since GOMs are standard tasks that can be used as indicators of student proficiency, they are applicable to content courses often taught in secondary classrooms (Espin & Foegen, 1996). Content-area measures, such as VM require students to
read fluently, know the meaning of key vocabulary words, and be able to hold that content knowledge in their memories (Busch & Espin, 2003). A content area CBM like VM may be more suitable for older students than curriculum skills mastery assessments that only assess a section of content or only basic skills.

**Vocabulary matching.** Valid screening and progress-monitoring practices are an essential component of successful RTI models (Mooney et al., 2010). The Research Institute on Progress Monitoring (RIPM) reports success for using five-minute VM probes to measure student learning in content area subjects. Importantly, VM probes are valid measures of both student performance and progress monitoring over time (Bursuck, 2010). While literature is limited, previous findings support using a vocabulary based CBM to measure progress towards general outcome measures in content area classrooms.

Vocabulary matching probes are created in paper-pencil format or as an online quiz and are individually or whole group-administered (Mooney, Benner, Nelson, Lane, & Beckers, 2008). Probes contain 20 randomly selected terms listed alphabetically on the left side of the page with 22 randomly ordered corresponding definitions, including two distracters, listed on the right side of the page preceded by letters (e.g. a, b, c, etc.; Mooney et al., 2010). Terms and definitions are taken from a student textbook glossary, teacher’s notes, or other sources like state standardized tests (Espin, Busch, Shin & Kruschwitz, 2001; Mooney et al., 2010). Definitions are typically rewritten word-for-word unless glossary definitions are longer than 15 words in length or contain the corresponding vocabulary term. Definitions are rewritten without the extraneous words (e.g. the, and, etc.) and the vocabulary term is replaced with a synonym. Since speed (in addition to accuracy) must be accounted for when correctly calculating growth to
predict success on state test outcomes (Wiley & Deno, 2005) five minutes are allotted for probe completion.

Espin and Deno (1993; 1994-1995) investigated the criterion-related validity of student-read study activities and student vocabulary knowledge to determine the stronger indicator of learning. Tenth grade students (n = 121) from a midwestern rural high school completed tasks in science and in English that included a VM probe, a pre-study read aloud, a content-area study task, and a post-test reading aloud drill. Results showed moderate correlations (.56 to .64) for all predictors, but VM had stronger achievement prediction than did the reading aloud task (Espin & Deno, 1993; 1994-95).

Extending the previous study, Espin and Foegen (1996) used a 20-term probe that was made up of timed readings with 10 terms taken from reading passages and dictionaries. Results represented strong correlations with three other instructional tools of .62 to .65 in favor of VM. Espin et al. (2001) compared 7th grade students’ scores from a student-read and an administrator-read VM probe with student overall grades, the Iowa Test of Basic Skills (ITBS) social studies subtest scores, and a knowledge pre- and post-test. Differences in sample correlations between probe scores and statewide tests across demographics were statistically non-significant at .56 to .64. Results included moderate to moderately strong correlations between the predictor and criterion variables, except with student grades. Given the shortcomings of grades as general outcome measures (c.f. Bursuck et al., 1999), this finding is not surprising. Vocabulary matching was the best predictor of future performance when compared to a reading aloud task and the maze for 10th grade students (Ketterlin-Geller et al., 2006). Higher correlations (.63 to .76) between students’ three VM probe scores and the 6th grade integrated Louisiana Educational Achievement Program (iLEAP) social studies subtest were found (Mooney et al., 2010).
Others have demonstrated VM’s usefulness in measuring progress over time (Busch & Espin, 2003; Espin & Deno, 1993; 1994-1995; Espin & Foegen, 1996; Mooney et al., 2013a; Mooney et al., 2013b). Growth rates per week have varied between .02 and .65 correct matches (Borsuk, 2010; Espin et al., 2005; Mooney et al., 2013b). Twyman and Tindal (2007) suggest using attribute maze procedures alone may not be adequate to measure older students’ mastering of content, but they do appear to have adequate progress monitoring capabilities in identifying students who are not progressing. Unfortunately, the numbers produced by statistical procedures may not be seen by classroom teachers as that impressive in terms of vocabulary knowledge growth.

**Vocabulary matching limitations.** The few VM studies do present with limitations. Generalizability is an issue when it comes to extending validity across geographic locations and student subgroups because only a small amount of research exists. Also, teachers were not always consistent in probe administrations possibly affecting drawn conclusions. Lastly, there are only a couple of growth studies and none have examined the use of cutscores for predicting future achievement on standardized state content assessments.

Generalizability is limiting based on the history of small sample sizes. Espin and Deno (1993) had a sample size of 121 that only included students from a rural community with few identified as having content-area struggles. Espin and Foegen (1996) included 184 urban students from three different grade levels, but only seven percent were classified with disabilities. Espin et al. (2001) included 58 students from one classroom with just eight percent classified with disabilities. Mooney et al. (2010; 2013b) included only 146 and 153 6th grade students of which eight percent and nine percent of students had disabilities and only eight percent were classified as gifted or talented in their 2010 study. When subgroups are small in
number it is hard to generalize research findings across broad ranges of abilities. Increasing sample sizes potentially would include more students with exceptionalities. Students from more than one school and across multiple grades are warranted (Brown-Chidsey et al., 2005; Espin et al., 2005). As sample sizes increase the need for increased attention to the reliability of measurements must be considered. One threat to scoring reliability within large sample sizes is that interobserver agreement is not always practical, even as an estimate (Silberglitt et al., 2006).

Vocabulary matching probes created by multiple parties and administered by multiple content area teachers may be able to address concerns related to subjectivity and predictability. Researchers suggest predictive power and strength in the relationship with state tests may have been reduced due to active teacher involvement (Mooney et al. 2008; Silberglitt & Hintze, 2005). Researchers were actively engaged throughout probe development, the scoring process, and administration, which may have contributed to bias findings (Mooney et al., 2010). Crawford et al. (2001) stated that their study was limiting because participating teachers, despite training, administered only some measures and there were no formal reliability checks. Espin et al. (2001) mentioned that since they only had one teacher administering probes, the teacher’s instructional style might have influenced probe formatting. Perhaps if more teachers from multiple settings participated in probe development stability across administrators might improve (Mooney et al., 2010).

Further research is needed to determine if VM probes are valid progress monitoring tools (Byers, Lembke, & Curs, 2013). Probes used to measure progress have only been applied within short time frames (Espin & Deno, 1994-95; Espin & Foegen, 1996; Espin et al., 2001) or after students have taken end-of-year tests. Byers, Lembke, and Curs (2013) suggest that VM probes be examined for predictive ability with standardized tests. “Sensitivity to change is a desirable
characteristic of assessment tools that will be administered repeatedly to assess student skill development.” (p. 29; Gansle et al., 2004). Using cutscores for predicting achievement on state tests can potentially help teachers set short-term instructional goals. Receiver operating characteristic curve analysis provides total control of the diagnostic accuracy levels desired by researchers lending themselves to subjectivity into the choosing of cutscores (Keller-Margulis et al., 2008).

There may be opportunity to use CBM to assess the benefit of specified modifications and accommodations that are often listed in students’ individual education programs (IEPs). Silberglitt and Hintze (2007) suggest that schools more frequently measure progress (e.g. weekly) in order to really determine if a prescribed intervention is indeed benefitting an at-risk student. Students’ IEPs may become more helpful to teachers in terms of appropriate interventions when such conclusions can be drawn. No studies have provided teachers with interventions to address students’ weaknesses (Espin & Deno, 1993-95). To ultimately be useful to practitioners, vocabulary-based assessment must not only provide diagnostic information, but also provide some level of prescription for intervention. The use of data to guide instructional decisions remains an underdeveloped area of vocabulary based assessments. Under an RTI model, screening for failure is only the first step in prevention. In order to “sell” the idea to teachers and administrators, researchers and administrators must provide evidence based interventions to “win over classroom educators” who have limited instruction time.

**Growth and Instruction**

One of the critical components to RTI is that educators have empirically supported measures such as CBM, as well as a reliable and valid method for measuring growth following instruction to determining whether that growth is adequate. Statistical procedures such as HLM
and multiple regressions have illustrated the rates at which students respond to instruction. Recommendations have been for schools to conduct universal screenings multiple times throughout the year (Ardoin et al., 2004). Despite the growing empirical studies illustrating CBM’s ability to predict achievement, this area is still limiting in content courses, especially for VM.

**Progress monitoring and measuring growth.** Although it is critical to monitor progress and adjust instruction (Crawford, Tindal, & Stieber, 2001), the increasing emphasis on accountability in schools makes it important for educators to be able to use the data they collect to determine whether students are on track for success or on the path of failure. This need for diagnostic accuracy, or deciding whether students are on track to pass their summative assessments based on their formative assessments, is increasing over time (Silberglitt, Burns, Madyun, & Lail, 2006). Elliot and Fuchs (1997) recommend that five essential components be considered when developing a valid progress monitoring tool in that it: (a) be quick to administer, (b) has reliability and validity, (c) be representative of the content that the student is learning, (d) influences the development of appropriate interventions, and (e) is sensitive to students’ gains so that intervention can be matched appropriately. Most progress monitoring and predictive studies have examined CBM’s relationships between researcher-developed cutscores and high-stakes tests or state assessments.

Shin, Deno, and Espin (2000) conducted a study that included 43 second grade students whose reading performance was measured monthly over one school year. Pearson correlations produced medium effect sizes, but multiple regression analyses showed that a single ORF probe predicated future performance better for administrations given in the fall, winter, and spring than did ORF probes taken all year. One-minute ORF probes were highly correlated (.84) between
2nd and 3rd grade administrations, but scores on the criterion-referenced multiple-choice reading test were only moderately correlated with test performance and not at all with math achievement (Crawford, Tindal, & Stieber, 2001). Students who had a reading rate of 72 words per minute all passed the state reading assessment and 82% of those who read 54 or fewer words per minute in 2nd grade failed the 3rd grade math state test. Despite lower gains for lower-achieving students, the CBM-ORF probe still provided enough data for teachers to at least identify students’ growth rates (Crawford et al., 2001).

A maze had alternate-form reliability with a mean coefficient of .81 that was positively related to achievement on a standardized reading test (Shin et al., 2000). Maze correlations were statistically significant for correct choices and for correct minus incorrect choices (.50; .55 respectively) when three months of maze and reading aloud data from 31 9th grade students were collected to determine predictive ability with a Basic Standards Test (MBST) high stakes reading assessment (Espin et al., 2010).

Stage and Jacobson (2001) suggest that HLM is a strong growth measurement tool as it allows examiners to measure growth slopes for individual students and for one to test for change in the student’s slopes at different points in time. Growth curve analyses and HLM statistical procedures have made it possible to study individual student growth patterns, enabling researchers to identify target ORF cutscores and confidence intervals (CI; Stage & Jacobson, 2001). Espin et al. (2005) used HLM to test VM probes’ sensitivity to improvement in student performance, the sensitivity to inter-individual differences in growth rates, and the validity of the growth rates produced by the measures with respect to the criterion tests. Probes illustrated significant group growth rates although only the student-read VM probe produced growth lines that were both valid and reliable predictors of future performance (Espin et al., 2005). For 7,544
students in 2nd through 6th grades, HLM demonstrated that growth was less for the weakest and highest students while growth became less apparent in 4th through 6th grades (Silberglitt & Hintze, 2007). Espin et al. (2013) reported reliability coefficients using HLM that accounted for 57% of the growth rate variance with growth rates of .63 matches per week. Mooney et al. (2013b) found statistically significant growth rates using linear mixed model analyses for two different semesters but these varied between subgroups of students. For weeks 2-14, the growth rate was .25 compared to .11 matches per week for probes 15-25 for families who paid full price for lunch compared to those receiving free/reduced lunch growth rates that were only .10 and .02. To date, there have only been two studies reviewing VM’s potential as a growth measure (Borsuk, 2010; Mooney et al. 2013b).

Development of more advanced statistical analyses, like HLM, allow for the examination of the characteristics of CBM measures as progress monitoring tools and performance measures (Espin et al., 2010; Shinn, et al., 2004). Curriculum based measurement can be used to measure student growth over short periods of time because they can be given in practical ways (Shin, Deno, & Espin, 2000). But as Deno et al. (2001) point out, the growth standards that practitioners use in schools must be higher than merely normative data if teachers are to respond in the most effective ways. What was emerging became known as “cutscores” that could be produced based on individual school’s data relevant to their own state specific tests.

**Diagnostic accuracy and establishing cutscores.** Cut scores or “decision thresholds” (Christ et al., 2013) are a fundamental component of curriculum based measurement. The ability of a point in time measure of performance (time series data) to correlate with other measures such as achievement, measured by summative measures, are the basis of the validity and reliability estimates previously discussed as supporting CBM (c.f. Reschly et al., 2009; Wayman...
et al., 2007). In practice, these cutscores are used to screen students for various levels of intervention within an RTI model. The pragmatic impact of these scores are that students identified as performing under or over these set performance levels (predictor) will reach or fail to reach a future performance on some criterion measure (e.g. state measures of achievement).

Diagnostic accuracy has been defined as the ability of an instrument or procedures to distinguish between two diagnostic options (e.g. pass and fail) and to select the one that is correct (Swets, Dawes, & Monnahan, 2000). Hintze and Silberglitt (2005) identify four possible categories of proportions that result from examining the diagnostic accuracy of predictive measure (e.g. cutscore) and the criterion measure (e.g. achievement measure at a future time). The first category of diagnostic accuracy is the sensitivity (a true positive rate) where the individual that failed on the criterion measure was predicted to fail by the cutscore. The second category was specificity (a true negative rate) where the individual that passed on the criterion measure was predicted to pass based on the cutscore. The third category was the positive predictive power (PPP) where individuals predicted to fail based on the cutscore actually failed on the criterion measure. The final category was the negative predictive power (NPP) where the individual that was predicted to pass on the predictive score actually passed on the criterion measure.

According to Silberglitt and Hintze (2005),

“By using a systematic method for establishing cut-scores at all grades and benchmark periods, educators can easily apply the concepts of formative assessment to evaluate the progress of an individual or group of students...a consistent set of cut-scores has powerful implications for the classroom by allowing for regular, frequent, and valid measurement to a common outcome...providing benchmarks on which to base the student’s responsiveness to intervention” (p. 322).
Researchers (Good et al., 2001; Hasbrouck & Tindal, 1992) have demonstrated the usefulness for CBM-ORF cutscores in predicting reading achievement. Discriminative analysis, logistic regression, and ROC curve analyses produced consistent diagnostic accuracy results of a R-CBM on a high-stakes 3rd grade reading test for 1,766 students (Hintze & Silberglitt, 2005). Researchers found that using a R-CBM to determine cutscores in a consecutive format from one target benchmarking period to the next across grades was more accurate and efficient than using a high-stakes test as the criterion despite grade level (Hintze & Silberglitt, 2005). Silberglitt and Hintze (2005) report that ROC curve analyses was the best method for increasing precision of cutscores compared to logistical regression, the equipercentile method, or the discriminant analysis for 2,191 3rd grade students from five districts.

Receiver operating characteristic statistical procedures was used to determine the predictive validity of a science content maze with the Idaho Student Achievement Test, Science (ISAT). While exploratory, authors stated that the science concept maze yielded helpful results through ROC curve analyses for labeling at-risk students, but the ROC produced cutscores influenced an overidentification of at-risk students with a 69% classification accuracy rate (Johnson et al., 2013). The optimized cutscore was 26 and resulted in missing 31 of 47 students who were not successful on the ISAT. But, when cutscores were set to 90%, the cutscore of 37 correct items on the maze overidentified 76 of 135 as being at-risk for failing the ISAT. “Optimizing specificity and sensitivity at .70 for each resulted in a 69% classification accuracy rate.” (p. 220, Johnson et al., 2013).

Researchers have systematically extended the application of cutscores by expanding the size and diversity of their sample sizes. Silberglitt et al. (2006) increased their sample to 5,472 students in grades three, five, seven, and eight and found that R-CBM had a wider range of
correlations (.51 to .71) than did the maze (.49 to .54). Keller-Margulis, Shapiro, and Hintze (2008) included 1,461 students in grades one through five and 1,477 students in grades two through five and reported moderate and significant correlations between ORF, math computation, and basic skills math probes with the TerraNova Achievement Test and the Pennsylvania state reading and math achievement tests. Approximately 84% were correctly identified based on who scored above or below the ROC cutscore (Keller-Margulis et al. 2008). Mooney et al. (2008) increased the generalizability of the above studies because not only was their population located in the southeastern United States, but also 60% of their sample received free or reduced lunch and 48% of the students were of minority background. Christ, Zopluogly, Long, Monaghen, and Van Norman (2013) included 1,517 2nd graders and 1,561 3rd graders. Even though the use of ROC procedures is increasing, growth and predictive studies still present limitations.

Curriculum based measurement has the ability to predict achievement, but the underdeveloped research areas require attention. One, small sample sizes, often limited by demographic characteristics such as classification, grade, ethnicity, and geographical location were found. Stage and Jacobson (2001) included only 173 4th grade students whose average performance outperformed the average student taking the WASL reading assessment limiting their results across academic abilities. Crawford et al. (2001) had a sample size of 51 3rd grade students increasing their standard error of measurement (SEM) and limiting their generalizability for a one-minute ORF probe to predict reading and math test scores. Silberglitt et al. (2006) had one year of 7th grade test data and Ticha, Espin, and Wayman (2009) only included 8th grade students. Keller-Margulis et al. (2008), Mooney et al. (2008; 2010; 2013a), and Espin et al. (2013) included students from only one school limiting generalizability within even a single
school district. Vannest, Parker, and Dyer (2011) only included 117 5th grade students (38% identified with disabilities and 62% were identified as having dyslexia) from 10 different schools in only Texas. The sample demographics, otherwise, were diverse including 26% African American and 37% Hispanic students and 53% of their total sample received free or reduced lunch (Vannest, Parker, & Dyer, 2011). Mooney et al. (2013a) included 106 5th grade students from one university lab school but focused on different content area than in their previous studies. Johnson et al. (2013) included 422 7th grade students from three schools across three states increasing generalizability across regions. Espin et al. (2013) included 198 7th grade students from 10 science classrooms but had just 14 weeks of progress monitoring data, whereas Mooney et al. (2013b) included 26 weeks worth of VM growth data.

Other researchers have addressed limited sample sizes in some aspects, but not in others. Silberglitt and Hintze (2005) increased previous sample sizes but their sample was predominantly Caucasian (95.3%). Silberglitt et al. (2006) had fewer than six percent minority students despite increasing their sample size to include students from multiple grades. The demographic breakdowns for the Johnston et al. (2013) study were still limiting with 77% identified as Caucasian, seven percent having disabilities, and none were ELLs. In Mooney et al. (2013a) no students had disabilities or were identified as gifted/talented, all were receiving high passing grades, all had scored above basic on a state science exam, and only 12% were minorities. Due to research studies conducted in very small geographical areas, Mooney et al. (2008; 2010; 2013a; 2013b) were unable to report results for ELLs and had zero to 10% of students receiving special education or gifted/talented services. Further research in VM with more heterogeneous samples across content areas over longer periods of time is necessary.
Two, student attrition rates were noted as threats to external validity. Mooney et al. (2008) report that making test predictions from 1st grade ORF probes to the 3rd grade iLEAP was limited because 35% of their 1st grade population was unavailable for comparison. Attrition of subjects, specifically in one area of reading, may have led to consistently higher means compared to benchmarks from a similar study (Keller-Margulis et al., 2008).

Three, content area predictive validity has rarely been discussed in the peer-reviewed literature, even though the need for extensive research has been recommended (Espin & Foegen, 1996; Espin et al., 2006; Mooney et al., 2010; 2013a). Only one study discussed the concurrent and predictive correlation coefficients for a science maze with other measures and ranges were from .63 to .67 (Johnson et al., 2013). The literature has left researchers wondering what exactly secondary content CBMs measure in middle schools. So far, according to the previously discussed summary, middle school content CBM research has only included reading and vocabulary knowledge assessments. In all abovementioned growth or predictive studies, VM probes or concept maze CBM has only been studied in either social studies or in science. Espin et al. (2013) recommend that studies include more science areas and other content subjects.

Other researchers caution teachers in using R-CBM for predicting achievement in middle school grades (Keller-Margulis et al. 2008; Silberglitt et al., 2006). Silberglitt and others (Silberglitt et al., 2006) examined the relationship of CBMs for reading, grades, and state achievement test scores. The sample size included 5,472 students in grades three, five, seven, and eight. Authors report that as grades increased (e.g. from 3rd to 8th) the correlations between the R-CBM and the state accountability tests decreased significantly from .71 for 3rd grade students to .51 for 8th grade students (Silberglitt et al., 2006). For 7th graders who took a maze, correlations were .54 and for 8th graders, .49. The mean number of words read increased about
30 WPM between 3rd and 5th grade, 20 WPM between 5th and 7th, but two WCPM between 7th and 8th grades (Silberglitt et al., 2006). The lessening ability of R-CBM to predict success on high stakes testing as grade levels increased was identified as problematic and supports the need for additional research into methods and procedures for improving CBM procedures in upper grades.

Maybe teachers do not value middle school CBMs, even though criterion assessments include important student information, because they do not have the means in which to apply the data. Perhaps this small amount of middle school CBM research is the result of apprehension met by teachers and administrators because of intense, lengthy, or inefficient scoring procedures. Or, since basic reading and calculation skills are not typically taught in middle and high school grades, teachers are seeking interventions that can be used with their older students to help them with content application. Research is needed that includes growth and predictive studies across states, examiners, demographic groups, grades, classifications, and different test and CBM instruments to extend generalizability to more representative groups (Silberglitt & Hintze, 2007; Stage, Hintze, & Silberglitt, 2005). Additional research is warranted if educators are going to ‘buy into’ consistent VM use within the classroom because until teachers see ‘how’ a brief outcome measure can benefit their instruction, CBM is less likely to achieve social validity. Predictability power for the purposes of meeting accountability at school- and district-wide levels is favorable or else school administrators may be less inclined to implement middle school content-area CBM. Because, so far, what researchers have not been able to describe is how middle school CBM, like VM, are affecting school performance measured primarily by summative assessment.
Study Rationale

With over 40 years of school based CBM research, what has emerged is that achievement can be improved by testing students (1) using standard valid tests, (2) that measure something important (3) on tasks of about equal difficulty tied to general curriculum and (4) that are conducted over time. Curriculum based measurement addressed components one, two, and four. General outcome measures further addressed component three (Powell-Smith & Shinn, 2004) as their broad dimension procedures could be applied (Fuchs & Deno, 1991) to achievement goals in content courses. The significant lack of research at the middle school level and the examination of how teachers can better meet their students’ needs all year is perhaps preventing solid RTI models from “taking off” in the upper grades.

Even though teachers have recognized the struggle in reading that middle school students have, current practices do not generally include content area literacy instruction. In a review of RTI implementation, 100% of participating middle schools reported using benchmark procedures to identify students at-risk for reading but none used measures in the content areas. Teachers will need to address this concern with appropriate reading interventions in order for students to comprehend textbooks (Johnson, et al., 2013). Another option for middle school teachers may be to use a content-specific CBM that is appropriate to their courses for measuring content acquisition in conjunction with RC and writing CBMs. While VM probe scores are correlating with state content subject tests, there remains concern as to whether or not they in fact indicate that content learning is taking place, if teachers are even using data to make instructional decisions, and if probes have predictive ability. Kratochwill, Clements, and Kalymon (2007) and Vaughn and Fuchs (2003) remind researchers that RTI criticisms center on the limited availability and validity evidence of screening protocol.
Prior research has addressed criterion-related validity in middle school content areas in a few published peer reviewed studies (Espin et al., 2001; Mooney et al., 2010; Mooney et al., 2013b). Seventh grade students’ VM scores had moderate to moderately strong correlations in relation to a statewide achievement test, students’ grades in a social studies class, and a researcher-created knowledge test (Espin et al., 2001). Criterion-related validity was again measured in southeast Louisiana by comparing 6th grade students’ VM scores with the iLEAP (Mooney et al., 2010; Mooney et al., 2013b). Additional exploratory information included investigating the differences between categories of students with the following: all 6th grade students, 6th grade students not identified with exceptionalities, 6th grade students identified with disabilities, and 6th grade students identified as gifted or as talented. Authors were able to compare patterns of academic performance across variables related to gender, ethnicity, socioeconomic status (SES), exceptionality, and across one entire school district. It extended the literature by using a more comprehensive procedure for generating course content to develop the probes. Together, these previous findings support the technical features of the reputable scores and present Stage 1 (Fuchs, 2004) evidence for the justification of VM.

In a theoretical sense, it was predicted that middle school performance and advanced assessment in middle school schools can be done in the same way that current researchers and school staff are able to predict reading and math achievement. The present study replicated previous research and included the development of an applied VM probe created with content drawn from course-based materials in 6th grade world history textbooks and key terms typically found on the 6th grade iLEAP social studies subtest.

The purpose was to extend the Mooney et al. (2013b) study by investigating researcher-developed cutscores to predict future performance by:
1. Examining the predictive validity of weekly world history VM probe scores with future achievement on a 6th grade standardized social studies subtest.

2. Determine if demographic variables are linked in 6th grade students who are at risk for school failure.

The predictive variables (independent variables; IVs) were defined as: (a) the number of correct matches (probe raw scores) on five five-minute VM probes. The criterion variables (dependent variables; DV) were the standardized scores from the LA iLEAP social studies assessment. To complete the analysis, differences in demographic characteristics across five of 22 VM probes using multiple linear regression (MLR) analyses were conducted to ascertain that predictors (cutscores) determined strength of the relationship between variables. Finally, the ROC curve statistical analysis was used to determine if VM probes given throughout the school year could produce cutscores that predicted student achievement on the SS iLEAP subtest with enough sensitivity and specificity. This process was established in previous research utilizing ROC to examine math and reading CBM. Receiver operating characteristic curve procedures allowed researchers flexibility in creating cutscores for identifying adequate levels of both sensitivity and specificity of .75, as recommended by Swets (1996).

Data produced from the statistical analysis could potentially help researchers to improve the validity of VM probes so that schools can develop their own VM probe norms for purposes of decision-making (LeBlanc, Dufore, & McDougal, 2012). Creating CBM cutscores that are associated with a high probability of passing the state content test may provide more valid individual student data that assists teachers in determining which students are more likely to benefit from tier two or three interventions. For example, ROC analyses could illustrate for students who correctly match seven vocabulary terms with their definitions have a 75% chance
of passing the social studies iLEAP test. A table of probable success (see Espin et al., 2008; 2010) could be developed for 6<sup>th</sup> grade social studies teachers in this school district to identify annual goals of 15 vocabulary terms, for instance, for a student who only knows five at the beginning of the year.

The following research questions were addressed:

1. Can a CBM-VM probe cutscore predict the likelihood of meeting ‘basic’ requirements for passing the 6<sup>th</sup> grade iLEAP social studies criterion assessment?

2. Will there be different correlations for five probe scores between student demographic subgroups and the iLEAP social studies criterion assessment?

**Research Hypotheses**

It was hypothesized that there would be statistically significant predictive relationships between the VM probe scores with the 6<sup>th</sup> grade iLEAP social studies subtest:

**HO1.** CBM-VM probe scores will not predict students’ likelihood of passing or failing the iLEAP social studies criterion subtest.

**HA1.** CBM-VM probe scores will predict students’ likelihood of passing or failing the iLEAP social studies criterion subtest.

**HO2.** CBM-VM probe scores will not have statistically different correlations with the iLEAP social studies test based on students’ demographics.

**HA2.** CBM-VM probe scores will have statistically different correlations with the iLEAP social studies test based on students’ demographics.
CHAPTER 3: METHODOLOGY

Participants and Setting

The Louisiana State University (LSU) Institutional Review Board (IRB) approval was obtained in order to conduct this study. Students from 6th grade social studies classrooms in three schools were chosen because social studies is a middle school content area in which CBM could potentially assess content learning and predict future performance. Students and teachers from these classrooms were already familiar with probe administration. Vocabulary matching probes should include the most important terms developed by university researchers and content-experienced teachers may enhance VM validity research (Mooney et al., 2010). In order to increase the generalizability of the VM probe utility, researchers should implement the probes across schools given by multiple teachers alongside of additional standardized measures (Mooney et al., 2010). This sample set is identical to that of Mooney et al. (2013b) as the researcher was involved in both studies but ran additional statistical analyses in answering the present research questions that were not yet explored.

Sampling procedures

Demographics. Of 232 6th grade students, 224 students who returned signed parental consents and child assent forms were included in the study. Eight students were excluded because their iLEAP data were unavailable. Student and school demographics are reported below (Table 2). For the gifted/talented and disability categories numbers are small. The most students attended school A and school A had the only 6th graders in the district classified as gifted/talented. More students from school A paid reduced/full price for lunch compared to schools B and C.
Table 2: Demographic Percentages for 6th Grade Students (n = 224)

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>(percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>112</td>
<td>(50%)</td>
</tr>
<tr>
<td>B</td>
<td>39</td>
<td>(17%)</td>
</tr>
<tr>
<td>C</td>
<td>73</td>
<td>(33%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>103</td>
<td>(46%)</td>
</tr>
<tr>
<td>Female</td>
<td>121</td>
<td>(54%)</td>
</tr>
<tr>
<td>Age (mean)</td>
<td>12.52</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>111</td>
<td>(49.5%)</td>
</tr>
<tr>
<td>African American</td>
<td>111</td>
<td>(49.5%)</td>
</tr>
<tr>
<td>Latino</td>
<td>2</td>
<td>(.05%)</td>
</tr>
<tr>
<td>Total Minority</td>
<td>113</td>
<td>(50%)</td>
</tr>
<tr>
<td>Socioeconomic Status (Lunch Status)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free</td>
<td>135</td>
<td>(60%)</td>
</tr>
<tr>
<td>Reduced Paying</td>
<td>16</td>
<td>(7%)</td>
</tr>
<tr>
<td>Paying</td>
<td>73</td>
<td>(32.6%)</td>
</tr>
<tr>
<td>Free + Reduced Paying</td>
<td>151</td>
<td>(67.4%)</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Students Classified as Gifted</td>
<td>4</td>
<td>(1.8%)</td>
</tr>
<tr>
<td>Students Classified as Talented</td>
<td>4</td>
<td>(1.8%)</td>
</tr>
<tr>
<td>Total students Classified as either Gifted or Talented</td>
<td>8</td>
<td>(4%)</td>
</tr>
<tr>
<td>No Classification or general education (students with 504 plans)</td>
<td>191</td>
<td>(85%)</td>
</tr>
<tr>
<td>Students Identified with Specific Learning Disability</td>
<td>13</td>
<td>(5.8)</td>
</tr>
<tr>
<td>Students Identified with Speech Impairment</td>
<td>8</td>
<td>(3.6)</td>
</tr>
<tr>
<td>Students Identified with Autism Spectrum Disorder</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Students Identified with Emotional Disturbance</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Students Identified with Health Impairment</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total Students with Disabilities</td>
<td>25</td>
<td>(11.1%)</td>
</tr>
</tbody>
</table>

Achieved Pass Criteria on 2009 iLEAP Social Studies Test

| (BASIC or higher) | 141 | (63%) |
| Achieved Advanced Level on iLEAP Social Studies Test | 20 | (9.8%) |
| Achieved Mastery Level on iLEAP Social Studies Test | 22 | (9.9%) |
| Achieved Basic Level on iLEAP Social Studies Test | 109 | (49%) |
| Achieved Approaching Basic Level on iLEAP Social Studies Test | 54 | (24%) |
| Achieved Unsatisfactory Level on iLEAP Social Studies Test | 19 | (8.4%) |

Note: iLEAP = integrated Louisiana Educational Assessment Program.
Students’ own 6th grade social studies teachers taught them throughout the year and administered the probes. School A’s female teacher was in her eighth year of teaching and had earned a bachelor’s degree and an elementary education certification. The teacher from school A attended a week-long summer institute in August 2008 where the VM probe development was reviewed in detail and she had previously administered three VM probes to her previous year’s 6th grade students as part of an earlier study (see Mooney et al., 2008). School B’s male teacher was in his sixth year of teaching, had earned a bachelor’s degree, and was certified as an elementary education teacher. School C’s female teacher was in her second year of teaching and received a bachelor’s degree in social studies and elementary education. These rural middle schools were three of ten school buildings in one southeast Louisiana parish. Kindergarten through 12th grade school district enrollment for the 2008-2009 academic year was 3,810 students.

Participants were taught social studies for 52 minutes a day in a departmentalized setting at their individual schools. School A was made up of six classes, School B of two classes, and School C of four classes. Sixth-grade social studies content and instruction focused on world history from the beginning of mankind through the 15th century (see Mooney et al., 2008). Teachers used the Holt Social Studies World History, 2006 textbook by Stanley Mayer Burstein and Richard Shek. Teachers reported that regular instruction included lecture led by the teacher, individual class work, or small group instruction and that students classified with disabilities received their special education instruction as guided by their IEPs. Instruction was mandated by the state to be aligned with the Louisiana Department of Education Comprehensive Curriculum (2008a), which aligns state content standards and GLEs by laying out material in units.
Research Design

This correlation research design was used to analyze quantitative data using parametric methods to study researcher-developed CBM-VM probes. The study was conducted between September 2008 and May of 2009 with students who were enrolled in 6th grade in the fall of 2008. I made initial campus contact, regarding this particular study, that fall. Students completed the social studies section of the iLEAP on one of five state mandated administered days in April of 2009. Data on VM probe scores, iLEAP scores, and demographics were collected from August 2007 through May 2009. The purpose of the design was to correlate the scores from selected VM probes and student demographic variables with the iLEAP, as well as measure the predictive relationship of researcher created cutscores.

The primary hypothesis, CBM-VM probe scores will predict students’ likelihood of passing the iLEAP social studies criterion subtest with enough sensitivity and specificity was tested by using ROC curve analyses. The crosstabs feature in SPSS was used to sort the data before testing scores’ sensitivity and specificity using a diagnostic accuracy application program. In addressing the second research question, multiple linear regression (MLR) analyses were used to model the relationship between the iLEAP score with variables such as ‘classification’ and ‘gender.’

Measures

Two types of measures were used to examine the relationship between a vocabulary-CBM score and the subject-based statewide assessment instrument: (a) 22 of 40 researcher-developed VM probes in 6th grade world history content (see Mooney et al., 2013b) and (b) the criterion-referenced social studies portion of the 2009 sixth-grade iLEAP. Predictors were VM-probe scores from five probes (7, 11, 15, 19, and 22). The probes for each ‘cutscore’ were
determined to be a fair representation of each other as each was administered in sequential order every four weeks prior to iLEAP administration (with the exception of 22).

**Predictor variables**

**Demographic variables.** Student demographic variables were coded in SPSS as follows: gender (0 = female, 1 = male), ethnicity (0 = white, 1 = black, 2 = Latino), school (0 = school A, 1 = school B, and 2 = school C), age at time of first probe administration was a continuous variable, repeated 6th grade (0 = did not repeat 6th grade, 1 = was repeating 6th grade), SES (0 = free lunch, 1 = reduced price lunch, 2 = paid full price lunch), and classification (0 = general education, 1 = classified under IDEA with a disability and had an IEP, 2 = classified as gifted or talented). Additional continuous variables included 2008 iLEAP social studies scores and 2009 iLEAP math, English, science, and reading standardized scores.

**Vocabulary matching probe scores.** VM probes were designed based on an adaptation of the procedures described in the review of literature (e.g. Espin et al., 2005; Mooney et al., 2008; 2010; 2013b) by a special education professor, an experimental statistics professor, 10 teachers (from two parishes who were either special education or social studies teachers), and me. Teachers administered probes during the 2008-2009 school year as part of the regular curriculum.

An Excel database included all glossary terms from the five state-approved world history textbooks and was created by me in the summer of 2008. The Excel file of over 900 terms was sent to 10 teachers from two school districts who were asked to rank terms of importance with a number “1” being the most important to number “3” being the least important. Rankings were averaged together for each term and an average number (1, 2, 3) was assigned to the term. The spreadsheet of terms was given to an employee of the Louisiana State Department of Education
(LSDE) who was a state assessment committee member and who taught social studies instruction at the university level. She deleted terms that appeared more than once and added some terms that were not in the glossaries but that were important vocabulary typically found on the iLEAP (e.g. latitude, longitude). This revised list was given to teachers to rank which semesters (first, 1 or second, 2) they typically taught the terms. These rankings were averaged and that mean number was assigned to the term (1 or 2). When only terms ranked with “1s” or “2s” were included as being the most important, the database decreased to 245 terms.

Next, the statistics professor, with an extensive background in statistics, used a randomization function in SAS to create 40 probes having 20 terms listed in alphabetical order on the left and 22 definitions that included two distracter definitions listed in random order on the right. This researcher recreated probes 15-27 in the same way with the Excel randomization function but ensuring that ten words ranked with a “1” and ten words Ranked as a “2” were included on each of the “revised” second semester probes; probe #26 was identical to probe #3 and probe #15 was the same as probe #27 (for purposes of the other ongoing study; see Mooney et al., 2013b). For this reason reliability and validity may have been impacted for probe 15, which was part of this study. One probe was administered each week that school was in session across the three schools (Figure 1).

Reliability and technical data for VM probes were included in earlier research. Alternate-form reliability for single probes ranged from .58 to .87 and from .70 to .85 for adjacent measures (Espin, Shinn, & Busch, 2001). Researchers report criterion-related validity between student-read VM probes and an experimenter-developed knowledge assessment that ranged
### Probe 7  
Social Studies 2008-09

**Name:** _______________________________  
**Date:** _____________

Directions: Place the letter associated with the response on the right on the line next to the term that the response best describes.

<table>
<thead>
<tr>
<th>No.</th>
<th>Term</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Appeasement</td>
<td>a. An object made by people long ago</td>
</tr>
<tr>
<td>2.</td>
<td>Arid</td>
<td>b. The worship of many gods</td>
</tr>
<tr>
<td>3.</td>
<td>Artifact</td>
<td>c. All the living and nonliving things that affect life in an area</td>
</tr>
<tr>
<td>4.</td>
<td>Astronomer</td>
<td>d. Philosopher and teacher whose beliefs influenced Chinese life</td>
</tr>
<tr>
<td>5.</td>
<td>Bilingual</td>
<td>e. To receive from a family member who has died</td>
</tr>
<tr>
<td>6.</td>
<td>Civilization</td>
<td>f. dry</td>
</tr>
<tr>
<td>7.</td>
<td>Code</td>
<td>g. Person who studies stars, planets, and other heavenly bodies</td>
</tr>
<tr>
<td>8.</td>
<td>Confucius</td>
<td>h. Complete freedom</td>
</tr>
<tr>
<td>9.</td>
<td>Currency</td>
<td>i. Giving into an aggressor so to keep the peace</td>
</tr>
<tr>
<td>10.</td>
<td>Environment</td>
<td>j. Having two official languages, as Canada has</td>
</tr>
<tr>
<td>11.</td>
<td>Independence</td>
<td>k. The New Stone Age; people learned to make fire and tools</td>
</tr>
<tr>
<td>12.</td>
<td>Levee</td>
<td>l. Region where southern Mexico, Belize, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, and Panama are</td>
</tr>
<tr>
<td>13.</td>
<td>Massacre</td>
<td>m. A city area</td>
</tr>
<tr>
<td>14.</td>
<td>Middle America</td>
<td>n. The practice of one person owning another person</td>
</tr>
<tr>
<td>15.</td>
<td>Polytheism</td>
<td>o. The kind of money used by a group or a nation</td>
</tr>
<tr>
<td>16.</td>
<td>Scholasticism</td>
<td>p. Society with cities, central government, and workers specializing in certain jobs, leading to social classes</td>
</tr>
<tr>
<td>17.</td>
<td>Slavery</td>
<td>q. Medieval way of thinking that tried to bring together reason and faith in religion studies</td>
</tr>
<tr>
<td>18.</td>
<td>Urban</td>
<td>r. The killing of many helpless people</td>
</tr>
<tr>
<td>19.</td>
<td>Valley</td>
<td>s. A dike used to control flooding</td>
</tr>
<tr>
<td>20.</td>
<td>Vandal</td>
<td>t. An organized list of laws or rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>u. Low stretch of land between mountains or hills that is drained by a river</td>
</tr>
<tr>
<td></td>
<td></td>
<td>v. A person who destroys property</td>
</tr>
</tbody>
</table>

Figure 1: Vocabulary Matching Probe

from .60 to .81 (Espin, Shinn, & Busch 2001). Mooney et al. (2010) had alternate-form reliability ranging from .76 to .82. Alternate-form reliabilities for 14 probes in 7th grade science
ranged from .64 to .84 with a mean reliability of .77 (Espin, Shinn, & Busch 2013). Probe scores were calculated based on the total number of correct choices.

**Criterion Variables**

*iLEAP standardized scores.* The LDE uses an expanded scale norm-reference testing model to comply with federal NCLB mandates. The program assists educators in measuring student proficiency of state standards and achievement compared to a national norm sample. The *iLEAP* is Louisiana’s statewide assessment in ELA, math, science, and social studies in grades three, five through seven, and nine (Mooney et al., 2010). The *iLEAP* tests were aligned with the Iowa Tests and items were developed in ELA and mathematics in grades two, five through seven, and nine to fully align with LA standards and grade level expectations. Students receive a “0” for an incorrect response and a “1” for a correct response on both subtests. Students scoring at or above ‘basic’ level are considered to have passed the test. Students who have scored ‘basic’ have demonstrated “only the fundamental knowledge and skills needed for the next level of schooling (LDE 2008a, p. 3). The tests are flexible in meeting legislative changes and are supported by coordinated program management (LDE, 2008).

In-state committees, consisting of LA educators and assessment consultants, designed test content using state curriculum and measured it in terms of content validity. After test content standards circulated across the state, a test blueprint was created so that it aligned with those standards. The final blueprint was field-tested to ensure that all items were functioning correctly according to content standards. A .90 correlation coefficient is recommended for tests in which educational decisions are made for students (Salvia et al., 2007). *iLEAP* reliability was reported in terms of internal consistency with a Chronbach’s alpha of .82 for the social studies subtest. The *iLEAP* technical summary reported no validity coefficients (LDE, 2008).
**iLEAP Grade 6 social studies criterion-referenced test.** Students have 40 multiple-choice items to complete under no time constraint on one day. Multiple-choice questions focus on world history and geography that is taken from the Louisiana’s social studies content standards, benchmarks, and GLEs (LDE, 2005). Three-fourths of the test is based on world history (p. 4-1) and one-fourth focuses on geographical information including cultures and human activities (p. 4-2). Achievement levels and associated scaled scores in 2008-2009 were as follows: (a) unsatisfactory, 100-260; (b) approaching basic, 261-291; (c) basic, 292-337; (d) mastery, 338-363; and (e) advanced, 364-500 (LDE, 2008). iLEAP standardized scores were entered into SPSS as continuous variables.

**Reliability and Validity**

Reliability is a major consideration in evaluating any assessment procedure because invested stakeholders would like to see that an assessment’s utility can be generalized to other populations, settings, and times (Salvia, Ysseldyke, & Bolt, 2007). Inter-rater reliability was examined through correlation approaches and by calculating percentage of agreement. Point-to-point agreement, for purposes of determining scorer reliability, was investigated in this study. Salvia et al. (2007) recommends “calculating point-to-point agreement” because it is a more precise way of computing percentage of agreement when considering agreement for each data point. The correlation approach, as described by Salvia et al. (2007), entails two people independently scoring a set of tests followed with running correlation statistical procedures. The resulting correlation coefficient is a reliability coefficient for scorers that are an estimate of interscorer reliability or agreement (Salvia et al., 2007). Another teacher, who was also a graduate student and employee of the school district, and I independently scored 27 weekly probes. I then entered the scores into a self-created Excel database.
The percentage of point-to-point agreement equaled the number of agreements on occurrence and nonoccurrence multiplied by 100 and then divided by the total number of observations for the VM-CBM. Cohen established Kappa that corrects:

“the proportion of agreement by removing the proportion of agreement that would occur by chance. Kappa values range from -1.00 (total disagreement) to +1.00 (total agreement); a value of 0 indicates chance agreement. A positive index of agreement indicates agreement above what test givers would expect to find by chance.”

(Salvia, Ysseldyke, & Bolt, 2007; p.129) Criterion-related validity was measured as the extent to which a students’ performance on a criterion measure (iLEAP) could be estimated from their performance on the VM probes (probe scores).

**Data Collection Procedures and Procedural Integrity**

After IRB approval was obtained, permission was secured from three middle school principals for teachers to administer weekly probes as part of their regular instruction programming. All three middle schools participated in the study. Teachers were trained in probe administration in the beginning of the year. This researcher was readily available to all three teachers daily. This researcher monitored weekly probe administrations five times throughout the year and teachers were permitted to contact university staff with questions. Teachers correctly followed protocol during the five visits. Procedural integrity, however, could only be assumed to be 100% during the remaining weeks.

Teachers provided students with pencils and probes were left upside down on student desks. Students were permitted to turn them over after all students appeared to understand the teachers’ instructions and had no remaining questions. Total time for instructions, administration, and data collection was approximately ten minutes each week. The iLEAP social studies subtest was administered as required by the state education department in April 2009.
Students had as much time as the needed to finish the test. If students were absent probes were not retaken and schools had ‘make-up’ time built into ‘i/LEAP week’ if students needed to make up the state test.

**Scorer training and inter-rater scoring procedures.** All scorers participated in a training session led by me prior to probe scoring. Mooney et al. (2010) scored the probes independently by two separate trained individuals. They reported an inter-rater reliability of .998, .994, and .995 consecutively for three probe sets. Across total scores, scoring agreement existed 88.1% of the time demonstrating strong inter-rater reliability. Espin et al. (2013) reported an accuracy of scoring 7th grade VM probes of 94%. An inter-rater reliability of 93.5% was reported in Mooney et al. (2013b).

I gave copies of the answer keys and weekly probes to the second scorer, who was an employee of the school district, and to a university professor. I scored the original data set and scoring procedures followed those described in earlier studies (see Espin et al., 2008; Mooney et al., 2010; 2013b). All VM probes were collected weekly and photocopied prior to any data scoring. The photocopied probes were given to the second scorer. Both sets of probes were scored and kept them in a secure cabinet in a locked room. The i/LEAP test was scored at the state level. The LDE distributed results to schools as per state regulations and district staff later shared those with me.

**Data Entry**

Interscorer agreement statistical analyses involved: (a) the correlation of two separate scorings per probe for all probes and (b) a percentage of agreement approach across scorers for all probes. The two sets of scored VM data were compared and if there was a discrepancy between raters, a third score was obtained and entered to measure for inter-rater reliability for all
of the VM scores. To ensure accurate data entry, the researcher visually checked accuracy of data entries. The data in the third column was used for all descriptive and inferential statistics. Once all data were entered, student names were replaced with student identification numbers in order to protect confidentiality.

SPSS software (2010) was chosen for all statistical analyses because of its prior use in similar studies (Hintze & Silberglitt; 2005; Shapiro et al., 2008; Silberglitt & Hintze, 2005). The Excel spreadsheets were transferred into a SPSS data set for statistical analyses. Before conducting data analyses, data were examined for accuracy of data entry, missing values, outliers, distributional properties, and parametric assumptions. Descriptive statistics, Pearson correlations, and MLR were computed to determine if preliminary correlations illustrated any relationship between the predictor variables. Receiver operating characteristic curve analysis was conducted to derive VM cutscores. Crosstabs was used to form tables for entry into a diagnostic accuracy software application program to determine sensitivity and specificity of cutscores.

**Data Analyses**

**Multiple linear regressions.** Multiple linear regression models helped to determine if values of the criterion variable (iLEAP score) could be predicted from the values of the predictors, which variables were linearly related to the iLEAP score, and if a subset of predictor variables could be identified that would potentially support predicating the iLEAP score. Assumptions of regression generally included if the residuals were normally distributed around the criterion (iLEAP score) and had straight-line relationships with those scores. The predictors (criterion variable, iLEAP social studies score) = slope * (VM probe score + intercept). Standard (simultaneous) stepwise multiple regressions in SPSS were used to determine the best predictors
of the relevant criterion variables: the VM total raw correct scores and the iLEAP SS standard scores.

Tabachnik and Fidell (2007) state, “Regression techniques can be applied to a data set in which the independent variables are correlated with one another and with the dependent variable to varying degrees.” (p. 117). The best multiple linear regression model for predicting the criterion variable was $Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_n X_n$ where $Y'$ was the predicted value and $X_1$ is the ith IV (the value of Y when all the X values are zero). For predicting iLEAP scores for schools, based on classifications (e.g. disability, general education, or gifted/talented), ethnicity, and gender, the model looked like this: Predicted score = $B_0 + B_1 (\text{classification}) + B_2 (\text{ethnicity}) + B_3 (\text{gender})$. The B’s were the coefficients assigned to each of the predictor variables during regression (see Tabachmick & Fidell, 2007).

**Multiple linear regression model assumptions.** Linearity and absence of outliers were determined by examining a graph that showed residual plots plotted against the X- and Y- axes in order to best assess the relationship between variables (as recommended by Tabachnick & Fidell, 2007). To test for normality, data plots, skew, kurtosis, and P-P plots were visually inspected. To test for homoscedasticity, SPSS scatterplots were examined. To test for independence that level-1 residuals and level-2 residuals were uncorrelated, cases would have been entered in order or by plotting residuals against the sequence of cases in conjunction with the Durbin-Watson statistic in SPSS (as recommended by Tabachnick & Fidell, 2007).

Regression coefficients were produced in SPSS and interpreted as the effects of the predictor variables on the criterion (iLEAP standardized score). When these assumptions were met, only one run of multiple regressions was necessary for accurate data analysis. Because all coefficients were reflective of normally distributed data, a parametric correlation procedure,
specifically the Pearson's product-moment correlation coefficient, was calculated. The Pearson correlation coefficient is a measure of the linear association between two variables and ranges in value from -1 to +1. “The absolute value of the Pearson correlation coefficient tells one the strength of the linear relationship.” (p. 196)

**Receiver Operating Characteristic Curves (ROC)**

Swets et al. (1996) recommend that cutscores maximize sensitivity and specificity ranging from .75 to .80 for predicting achievement. Swets, Dawes, and Monahan (2000) describe the ability of an instrument to distinguish between two diagnostic alternatives and selecting the correct one as ‘diagnostic accuracy.’ Diagnostic accuracy considers (1) sensitivity, (2) specificity, (3) positive predictive power (PPP), and (4) negative predictive power (NPP). Power = sensitivity = 1 – β. As described by Christ et al. (2013), interpreted as a correlation coefficient for categorical variables, the Phi coefficient is the measure of association between two binary decisions made based on true slopes and ordinary least squares (OLS) estimated slopes (= TP * TN – FP * FN / √(TP * FN * N * P * F)). Within this study, ‘sensitivity’ was defined as the true positive rate (TPR) or probability that VM cutscores correctly identified students who would pass the iLEAP SS test and actually passed. Specificity was defined as the true negative rate (TNR) that students who did not earn the cutscore indeed did not pass. Positive predictive power (PPP) was defined as the probability of those students whose VM cutscores indicated that they would pass the iLEAP subtest and had. Negative predictive power (NPP) referred to the likelihood that those students who did not earn the ROC determined cutscore still passed the iLEAP. The ROC function in SPSS was used to create five cutscores and a Diagnostic Utility Statistics application was used to determine if those five cutscores indeed met components 1-4 above.
Receiver operating characteristic curve analyses were conducted using achievement levels of ‘fail’ (included scores of ‘unsatisfactory’ and ‘approaching basic’) and ‘pass’ (included all iLEAP scores of ‘basic, mastery,’ and ‘advanced’) on the iLEAP 6th grade SS assessment as the criteria for all cutscore analyses and included only the probes taken in weeks 7, 11, 15, 19, and 22. Probe one was only administered in School A. Probes two through five were administered so early in the year and students may still have been getting accustomed to probe administration. Probe six was completed in the wrong week at School B, probes 24 - 27 were given after the iLEAP, and probe 23 was given so close to iLEAP administration; none of these were included.

Receiver operating characteristic curve analysis is often used in education research as it strongly differentiates between false positive and false negative errors (VanDerHeyden & Burns, 2010). The ROC area is a measure of the accuracy of the model. An area of 1.0 illustrates a perfect model whereas an area of .5 shows a useless model (Geerts et al., 2006). An ROC curve illustrates the tradeoff between sensitivity and specificity for all predicted outcomes. The higher the curve and the closer it follows the vertical axis, the higher the accuracy. The area under the curve (AUC) was reported because it is sufficient as an effect size indicator (Swets, 1996). When sensitivity and specificity are both at least .70 it is sufficient to use as a diagnostic indicator (Swets, Dawes, & Monahan, 2000) and is used as a measure of predictive power (Christ et al., 2013). Christ et al. (2013) describe the AUC in the following way:

“It is obtained by calculating the sensitivity and specificity for all possible cut-off points on observed slope estimates by fixing a cut-off point on true slopes, and plotting 1-specificity (or TPP) against sensitivity (or TNP). Area under the curve is expected to be 0.50 if predictions of which students are at risk were performed randomly without any information (e.g. by chance). Area under the curve is expected to be 1 for a perfect diagnostic method to identify students at risk…..an AUC of at least 0.80 is required for routine low-stakes decisions and 0.90 for high stake decisions.” (p. 32)
The ROC curve analysis function in SPSS was computed in the following manner for each of the five probes. After one ROC test was conducted for one probe, the same procedure was again repeated for the remaining probes (Figure 2). The list of cutscores was used for the researcher determined target points establishing the slope data given in terms of any long-term relationship between the VM scores and an acceptable iLEAP score for at least earning ‘basic.’

Probe scores were entered as the ‘test variable’ in SPSS since I wanted to ‘test’ if the probe scores could be valid future predictors; an iLEAP score of at least a 292 was the ‘state variable’ since 292 was the minimum score needed to pass the iLEAP-‘BASIC.’ A new variable, ‘State_Exam_SS_09’ was computed and coded either as a ‘1’ (iLEAP score was at least a 292) or as a ‘0’ (iLEAP score was below a 292). This variable was calculated against the researcher chosen probe scores.

Figure 2: Receiver Operating Characteristic Curves of the Model “Benchmark Probe Scores and Social Studies 2009 Pass”
Coordinates of the ROC curve determined cutscore criteria for iLEAP scores. Levels of specificity and sensitivity were derived for each raw probe score. The combination of sensitivity and specificity is best at one point for each curve. I chose cutscores in the same way cutscores were chosen in Burns et al. (2011), Mooney et al. (2008), and in Stage and Jacobson (2001) with a sensitivity and specificity of at least .70. The following guide was provided by Geerts et al. (2006) for determining accuracy of the ROC area: .90 - 1.0, excellent; .80 - .90, good; .70 - .80, fair; .60 - .70, poor; and 0.50 - .60, fail. In this study, sensitivity was read at .75 but specificity at .25 because in SPSS ROC produces the specificity as a component of -1. Next, the probe score that met the above criteria was chosen as the cutscore. In cases where scores had a .5 decimal they were rounded up to the next whole number since students’ probe scores are based on raw whole number scores for the number of correct choices.

A new variable was computed for each cutscore and labeled as, for example, ‘CBM_CutScore_7, ‘CBM_CutScore_11,’ and so on. Cutscores were then recoded as separate variables as a ‘1’ if a student earned or exceeded the cutscore (received that many correct matches) on the probe and coded as a ‘0’ if the number of correct matches fell below the cutscore.

The SPSS cross tabulations feature was used to illustrate the number of students who had the cutscore and passed, the number of students who had it but failed, those who did not have the cutscore but who passed, and those who did not earn it and failed the exam. VanDerHeyden and Burns (2010) cautioned that while ROC curve analyses might have produced more specific and sensitive cutscores for improving CBM predictability; they should be checked for diagnostic accuracy. So, my crosstabs data were next entered into the diagnostic accuracy program.
Diagnostic utility statistics displayed the frequency data for the number of students who had and did not have the VM cutscore and failed and for those who had and did not have the cutscore but who passed the iLEAP. Since the diagnostic accuracy program is based on a disease prevention model in modern day medicine, it predicts who is more likely to catch a disease based on a predictive factor (e.g. smoking is the test variable and lung cancer is the state variable or disease). In this study the cutscore was used as the predictive factor and the iLEAP ‘pass’ or ‘fail’ was used as the ‘disease.’ These data were entered into a diagnostic accuracy application program and ran for each newly created VM ‘cutscore’ to get their levels of 1) sensitivity, (2) specificity, (3) false positive rate ($\alpha = type\ I\ error = 1 - specificity = \frac{FP}{FP+TN}$), (4) false negative rate ($\beta = type\ II\ error = 1 - sensitivity = 1 - sensitivity = \frac{FN}{TP+FN}$), (5) positive predictive power (PPP), (6) negative predictive power; (7) Hit Rate (observed agreement), (8) Kappa, and (9) Phi coefficient. iLEAP (disorder) with ‘fail’ (disorder absent or ‘condition negative’) and ‘pass’ (disorder present or ‘condition positive’) against the VM-CBM cutscore (test for disorder) as either ‘fail’ (test outcome positive; did not have the cutscore) or ‘pass’ (test outcome negative-had the cutscore; Fail Prediction-Failed, Pass Prediction-Passed, Fail Prediction-Passed, or Pass Prediction-Passed) were calculated.
CHAPTER FOUR: RESULTS

Fidelity of Probe Administration

Fidelity checks on teacher administration on five occasions across the study revealed that teachers administered the probes correctly. Teachers reported no particular problems with probe administration. Although not indicated by the fidelity results obtained, one should still consider the possibility of administration error as a potential explanation for the pattern of results.

Interscorer Agreement

For agreement on all probes, individual participant tests scores were compared across scorers, with the number of agreements across scorers for the total score of each test divided by the number of test administrations. Interscorer agreement correlations averaged .996 across all 27 probes. Scoring agreement was as follows: VM (correct responses), 93.5 percent. SPSS inter-rater statistical analysis (Kappa) indicated inter-rater reliability ranging from .88 to .99 across all 27 probes averaging .93. Probe 10 had the lowest and probe 27 had the only score falling below .90. For probes 4, 14, and 23, Pearson correlations (.994, .993, .999 consecutively) were substituted for Kappa given that analysis could not be computed because the first scores on each were not identical.

Data Accuracy

The data were screened for outliers and no cases were removed as demographic predictor variables demonstrated appropriately degrees of skewness and kurtosis at less than +/-1.96 (as recommended by Madansky, 1988). The final statistical analyses ran were based on the scores that included all blank cells (no scores reported) in SPSS.
Test Descriptives

To check the statistical significance and relative importance of each predictive variable, the R squared was used to examine the relationships between the various predictor variables and the iLEAP scores. Pearson correlations were calculated among the 12 predictive variables (seven demographic variables and five VM probe scores) and the iLEAP standardized score. Confidence interval (CI; 95%) estimates of the correlations between the iLEAP, demographic variables, and each of the benchmark probes were obtained in the original study conducted by Mooney et al. (2013b). Means, standard deviations, and low and high scores are reported (Table 3). Pearson correlations indicated that for probe 19, the mean for the final sample of cases was slightly higher at 8.79 than the means for the remaining probes.

Table 3: Descriptive Statistics for VM Scores and 6th Grade iLEAP Scores

<table>
<thead>
<tr>
<th>Measure</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probe 7</td>
<td>194</td>
<td>8.0</td>
<td>4.4</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Probe 11</td>
<td>155</td>
<td>6.0</td>
<td>3.3</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Probe 15</td>
<td>177</td>
<td>7.8</td>
<td>3.9</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Probe 19</td>
<td>135</td>
<td>8.8</td>
<td>5.0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Probe 22</td>
<td>191</td>
<td>7.1</td>
<td>5.24</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>iLEAP Social Studies 2009</td>
<td>224</td>
<td>308.23</td>
<td>39.70</td>
<td>100</td>
<td>441</td>
</tr>
<tr>
<td>iLEAP Science 2009</td>
<td>224</td>
<td>307.42</td>
<td>44.45</td>
<td>114</td>
<td>410</td>
</tr>
<tr>
<td>iLEAP English Language Arts 2009</td>
<td>224</td>
<td>299.74</td>
<td>42.98</td>
<td>156</td>
<td>435</td>
</tr>
<tr>
<td>iLEAP Reading 2009</td>
<td>224</td>
<td>298.68</td>
<td>57.25</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>iLEAP Math 2009</td>
<td>224</td>
<td>293.96</td>
<td>57.38</td>
<td>100</td>
<td>486</td>
</tr>
</tbody>
</table>

Note: VM = vocabulary matching curriculum based measurement raw scores; VM scores constitute the number of correct matches in five minutes; state test = integrated Louisiana Educational Assessment Program (iLEAP) standard scores
Multiple Linear Regression Model Assumptions

The assumption of linearity was met for the demographic variables. Since skewness and kurtosis of all randomly drawn variables were respectively close to zero and three, assumptions of normality were met. After viewing SPSS produced scatterplots it was determined that homoscedasticity was met and no cases needed to be removed. Independence: level-1 residuals and level-2 residuals were uncorrelated. Durbin-Watson confirmed non-independence or errors for predictor variables as all DV values were very close to 2.0. Limitations are later discussed.

Relationships between Measures

As none of the correlations reached the .80 thresholds and no R squared statistic was close to 1.0, all MLR models comparing subgroups showed that no two variables were closely related in the analyses. Regression analyses for probes 7, 11, 15, 19, and 22 produced strong to very strong correlations between the VM scores with the 6th grade iLEAP SS subtest (.65, .52, .67, .70, and .65, respectively). Probe scores also produced correlations high in magnitude with other sections of the 6th grade iLEAP with correlations in the following ranges: science (.57 - .70), English (.61 - .75), reading (.55 - .67), and math (.54 - .65) even though the content of the probes was unrelated to that material. The majority of correlations replicated previous findings related to the strengths of the relationship between a five-minute content-area vocabulary CBM and statewide assessments reported as r = .70 for the Mooney et al. (2010) and .68 for Mooney et al. (2013b).

Table 4 illustrates correlations and 95% confidence intervals of the five probe score predictor variables. All confidence intervals (CI) that did not include zero suggest that the corresponding measures were different at the .05 significance level. If the CI did include zero, then there was insufficient evidence to conclude that the corresponding correlations were
different at the .05 level of significance. Of all the predictor variables, probes 15 [95% CI, -3.43 - 2.13], 19 [95% CI, 1.614 - 7.11], and 22 [95% CI, -2 - 2.9] had the highest correlations with the 2009 social studies subtest at .7. Only Probe 19 had a CI that did not include zero, so there is insufficient evidence to conclude that the corresponding correlations were different at the .05 level of significance for probes 15 and 22.

Table 4: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Pearson Correlation</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>iLEAP 2008</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>iLEAP 2009</strong></td>
<td>152</td>
<td>.50</td>
<td>[112.37, 293.73]</td>
</tr>
<tr>
<td></td>
<td>194</td>
<td>.60</td>
<td>[.96, 6.8]</td>
</tr>
<tr>
<td>Probe 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>.50</td>
<td>[-.38, 2.3]</td>
</tr>
<tr>
<td>Probe 11</td>
<td>177</td>
<td>.70</td>
<td>[-3.43, 2.13]</td>
</tr>
<tr>
<td>Probe 15</td>
<td>135</td>
<td>.70</td>
<td>[1.614, 7.11]</td>
</tr>
<tr>
<td>Probe 19</td>
<td>191</td>
<td>.66</td>
<td>[-2, 2.9]</td>
</tr>
<tr>
<td>Probe 22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple regression analyses produced Pearson correlations for each VM probe score and ethnicity (minority or Caucasian) and on eligibility (Gifted/Talented, Non Gifted/Talented, or Disabilities; Table 5).

Table 5: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores

<table>
<thead>
<tr>
<th>Probe</th>
<th>Minority</th>
<th>Caucasian</th>
<th>Gifted/Talented (GT)</th>
<th>Non GT</th>
<th>Disabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>#7</td>
<td>.60</td>
<td>.66</td>
<td>.29</td>
<td>.64</td>
<td>.64</td>
</tr>
<tr>
<td>n</td>
<td>98</td>
<td>96</td>
<td>9</td>
<td>185</td>
<td>22</td>
</tr>
</tbody>
</table>
Additional multiple regression analyses produced Pearson correlations between each VM probe score and SES (free/reduced lunch or full pay lunch), gender (males or females), and for those who repeated 6th grade and for those who had not repeated 6th grade (Table 6).

Table 6: Pearson Correlations between VM Cutscores and iLEAP Social Studies Standard Scores

<table>
<thead>
<tr>
<th>Probe</th>
<th>SES: Free/Reduced Lunch</th>
<th>SES: Full Pay Lunch</th>
<th>Males</th>
<th>Females</th>
<th>Repeated 6th grade</th>
<th>Had not repeated 6th grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#7</td>
<td>.60</td>
<td>.66</td>
<td>.70</td>
<td>.68</td>
<td>.25</td>
<td>.67</td>
</tr>
<tr>
<td>n</td>
<td>133</td>
<td>61</td>
<td>101</td>
<td>93</td>
<td>13</td>
<td>181</td>
</tr>
<tr>
<td>#11</td>
<td>.58</td>
<td>.38</td>
<td>.55</td>
<td>.51</td>
<td>.27</td>
<td>.52</td>
</tr>
<tr>
<td>n</td>
<td>110</td>
<td>44</td>
<td>65</td>
<td>89</td>
<td>9</td>
<td>145</td>
</tr>
</tbody>
</table>
The final step involved checking the diagnostic accuracy of the ROC derived cutscores for sensitivity, specificity, PPP, and NPP using the Diagnostic Utility Statistics application. Data produced by the crosstabs feature in SPSS was inputted into the application for each cutscore separately. An example is illustrated below (Table 7).

Table 7: Diagnostic Utility Statistics Calculation

<table>
<thead>
<tr>
<th>CBM_CutScore_7</th>
<th>Disorder Present</th>
<th>Disorder Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Fail iLEAP-did not achieve 292_BASIC)</td>
<td>(Pass iLEAP-achieved 292_BASIC)</td>
</tr>
<tr>
<td>Test Positive</td>
<td>50 (True Positive)</td>
<td>45 (False Positive)</td>
</tr>
<tr>
<td>(did not have Cutscore_6 words correct)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probe</th>
<th></th>
<th>SES:</th>
<th>SES:</th>
<th>Males</th>
<th>Females</th>
<th>Repeated 6th grade</th>
<th>Had not repeated 6th grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Free/Reduced Lunch</td>
<td>Full Pay Lunch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#15</td>
<td></td>
<td>.64</td>
<td>.61</td>
<td>.73</td>
<td>.62</td>
<td>.20</td>
<td>.68</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>119</td>
<td>58</td>
<td>75</td>
<td>102</td>
<td>7</td>
<td>170</td>
</tr>
<tr>
<td>#19</td>
<td></td>
<td>.69</td>
<td>.68</td>
<td>.72</td>
<td>.69</td>
<td>.64</td>
<td>.70</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>86</td>
<td>49</td>
<td>56</td>
<td>79</td>
<td>6</td>
<td>129</td>
</tr>
<tr>
<td>#23</td>
<td></td>
<td>.60</td>
<td>.64</td>
<td>.70</td>
<td>.62</td>
<td>.40</td>
<td>.65</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>122</td>
<td>61</td>
<td>85</td>
<td>98</td>
<td>10</td>
<td>173</td>
</tr>
</tbody>
</table>
After the information was entered into the Diagnostic Utility Statistics Application, the calculations produced for each cutscore were produced (Table 8). A true positive outcome correctly identified the students who would fail the iLEAP and who actually failed - sensitivity (19%; n = 42). A false positive outcome was one where a student predicted to fail actually passed the iLEAP (20%; n = 45). A false negative occurred for students who were predicted to pass but who failed the iLEAP (13%; n = 30). A true negative represented those identified to pass who did pass - specificity (44%; n = 99). With large number of false positives (45) and many false negatives (30), a positive VM diagnostic cutscore is in itself poor at confirming iLEAP failure because it over-identifies. The cutscore incorrectly identified 20% (n = 45) as going to fail when they passed the state exam (false positives) and 13% (n = 30) of students to pass but who ended up failing (false negatives). The cutscore did, however, correctly identify 22% [50/(50+30) = 63%] of those students who failed as going to fail (sensitivity). As a predictor test, a negative result is almost acceptable at reassuring that a student will pass the iLEAP since it correctly predicted 69% who passed the test- (99 / = 45+99) = negative predictive power (NPP), or specificity. No cutscore was sensitive and specific enough to be used as a ‘good’ diagnostic indicator of performance. Probes 19 and 20 had just enough sensitivity but were weak
in specificity. Specificity and sensitivity have to be met with at least .70 (or .75 and .25 when using SPSS as I had in this study).

Table 8: Diagnostic Accuracy Across 6th Grade Students for Predicting Achievement

<table>
<thead>
<tr>
<th>Content Area Measurement: Vocabulary Matching</th>
<th>Benchmark Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probe 7</td>
</tr>
<tr>
<td></td>
<td>(N = 194)</td>
</tr>
<tr>
<td>Cutscore</td>
<td>6</td>
</tr>
<tr>
<td>AUC</td>
<td>.802</td>
</tr>
<tr>
<td>Sensitivity (TPR)</td>
<td>.6250</td>
</tr>
<tr>
<td>Specificity (TNR)</td>
<td>.6875</td>
</tr>
<tr>
<td>False positive rate</td>
<td>.3125</td>
</tr>
<tr>
<td>False negative rate</td>
<td>.375</td>
</tr>
<tr>
<td>PPP</td>
<td>.5263</td>
</tr>
<tr>
<td>NPP</td>
<td>.7674</td>
</tr>
</tbody>
</table>

Note: FAIL = below ‘BASIC’ achievement level on iLEAP; Cutscore = # of correct matches out of 20; AUC = Area Under the Curve, TPR = true positive rate, TNR = true negative rate; PPP = positive predictive power; NPP = negative predictive power
CHAPTER 5: DISCUSSION

The purpose of this study was to extend Mooney et al. (2013b) by developing VM cutscores for predicting future performance. Reliability and validity of VM probe scores as valid indicators of performance for a middle school 6th grade state standardized social studies test were examined. All probes showed correlations at the .01 significance level, even though probes 1 and 2 showed lower Pearson correlations below .50 suggesting that there are strong relationships between weekly vocabulary probe scores and the iLEAP social studies subtest. The .65 (probes 17 & 18), .66 (probe 22), .67 (probe 15), and .70 (probe 19), correlations with the social studies subtest were stronger in magnitude and reproduced previous findings related to the strengths of the relationship between a five-minute content area vocabulary CBM and a statewide assessment ($r = .70$; Mooney et al., 2010; .68; Mooney et al., 2013b). Researchers may wish to consider developing probes that have stronger construct validity at least in terms of prediction.

In answering research question number two, findings were strong across demographic variables such as ethnicity, disability category, and SES. Linear relationships of students not identified with exceptionalities and those identified with a disability for all five probes ranged between .53 and .68; for free/reduced lunch between .58 and .69; for minority students between .51 and .64; for those who repeated 6th grade between .20 and .64. The author hypothesized that VM #19 was the highest because a large number of students ($n = 79; 39.7\%$) were missing from the data because school B did not complete the probe. School C had only moderate correlations all year.

The predictive validity for social studies VM probe scores was examined by using the ROC statistic analyses for the first time in order to determine if valid cutscores could be developed. This is the main contribution of this present research to the existing literature.
Receiver Operating Characteristic curve analyses were a valid method for determining if cutscores could predict student achievement. Cutscores had weak sensitivity and specificity that predicted a high number of false positives, indicating that at this time, VM probe scores are not valid indicators of performance because they incorrectly identified 33% (false positives 20% + 13% false negatives) of students’ iLEAP performance.

Teachers want to know, with as much certainty as possible, which students are likely to fail an accountability test based on a valid diagnostic indicator. Students first need to be accurately identified, so that service delivery within a middle school RTI model focuses on the use of targeted interventions (Johnson, Mellard, Fuchs, & McKnight, 2006). It is better to intervene, even in error, with those students who are predicted to fail based on a CBM but then passed as opposed to providing intervention to students who are likely to pass in the first place (Keller-Margulis et al, 2008). For example, if teachers used the cutscore for probe seven they would respond by implementing a vocabulary intervention for 95 students to who they thought were at-risk for failing the iLEAP when in fact only 50 of them really needed one. In addition, if interventions were assigned to students solely based on this one cutscore, 30 students would miss out on needed intervention (the 30 false positives who were predicted to pass but who actually failed the exam). These students (13%) would not benefit from RTI if only probe seven was used for initial identification since they would never have been identified as at-risk for failure. A concern that emerged was regarding what exactly the probes were measuring when considering students only needed six correct choices out of 20 on probe seven to pass the state exam that was given seven months later. Results extend the prior research on VM as both sensitive and specific cutscores for predicting achievement on a summative state test were not successively created.
Limitations and Future Research Recommendations

Even though elementary CBM assessments have demonstrated the ability to do so, Espin et al. (2010) warns practitioners that it is not sufficient to simply assume CBM will work as such in middle or high schools. Sixth grade world history students from one small rural district in a southeastern state limit generalizability to other grades, districts, demographic populations, regions of the country, and to other VM probes with different states’ content area courses and standardized assessments. One, even though almost all students who took the 6th grade social studies iLEAP were included in the sample, it only made up about 96.5 percent of total district 6th graders because some students were not required to take the iLEAP. United States (US) government, US History, and ELA probes could extend construct validity of VM-CBM for predicting achievement (and progress over time), especially considering when the correlation between these VM probes showed higher correlations with ELA than they did with the social studies portion of the iLEAP.

Two, due to missing data, readers are cautioned in interpreting results. The percentages of missing individual student probe data from week-to-week between weeks one and 23 ranged from 10.7% (week three) to 54.5% (week one; these data were not included) averaging 21.3%. Using the SPSS missing data function or calculating student mean scores between weeks (e.g. every four weeks) may help address the large amount of missing VM scores. In calculating ROC curves, every blank VM score cell in SPSS was counted as a student not having the CBM_cutscore, which could have negatively affected levels of sensitivity and specificity. A data analyses comparing students with high percentages of missing data with those who have minimal probe scores missing could highlight patterns for future qualitative exploration (e.g. students, teachers, school-absenteeism rates).
Three, while teachers were more than willing to participate in the yearlong probe administration; their knowledge and experience of assessment design, fidelity of implementation, and personal ‘buy-in’ for vocabulary instruction may have affected VM scores. Teachers reported that they did not always give the probes on Fridays even though they understood they were all supposed to administer them on the same day each week. Likewise, the rating system used to determine which terms to include and in which semesters may unintentionally not have been aligned with the instruction pace of the three administrating teachers. Raters’ personal biases toward selecting vocabulary terms should be explored in future probe development. Additional studies implemented solely by university staff specializing in protocol development, testing administration and scoring, and who have personal investments in the project, may provide slightly different results.

Three, the time and manpower in which to ensure accurate scoring procedures needs to improve if a VM probe score is to meet all of the criteria of a valid indicator of performance. Formative evaluation is a lengthy and difficult process that strains teachers who are already in demanding roles (Burns, Klingbeil, & Ysseldyke, 2010). Special education teachers have a high number of students on their caseloads and may not have access to all students on a daily, weekly, or even monthly basis. Web-based applications that decrease the ‘man-power’ for secondary teachers may increase the face validity and the alternate-form reliability of CBM measures.

Unlike in earlier studies, I emailed classroom teachers their students’ progress monitoring Excel graphs monthly so that they could share results with students, but scores were not part of students’ grades. O’Brian (1926) suggests that “the individual graph made one of the strongest appeals to pupils and proved one of the most effective instruments in stimulating speed in reading.” Current computer based CBM programs like AIMSweb and EASYCBM also include
graphs for easier interpretation by parents, teachers, and students. DiGangi, Maag, and Rutherford (1991) further highlight “self-graphing appears to be a potentially powerful variable for enhancing reactivity of self-monitoring for both on-task behavior and academic performance” (p.228). Mooney et al. (2013a) report social validity data from student feedback that indicated student appreciation for immediate feedback received as soon as they finished their online VM probes. Studies including immediate graphed progress may influence VM scores and their future correlations with summative assessments.

Four, once reliable and valid measures are fully developed and implemented, it will be necessary to study whether middle school teachers not only use the CBM produced data, but if their interventions improve content course performance for those students whose data illustrate significant academic struggles. Assessment data should identify strengths in addition to student weaknesses, so that students who are achieving at faster rates are provided with an appropriate level of instructional changes (Fuchs, Fuchs, Hosp, & Hamlett, 2003). Quality assessment is only a part of quality instruction (Salvia, Ysseldyke, & Bolt, 2009). I recommend that peer-reviewed research based academic interventions based on VM scores be included in future studies to determine how effective those interventions are on growth, predictability, and achievement.

Five, probe cutscores produced a high number of false positives (31%, 45%, 33%, 52%, & 34%) and false negatives (38%, 39%, 31%, 30%, and 29%). The average percentage of false positives was 39% and the average percentage of false negatives was 33.4%. Schools have limited resources to provide interventions. Once sensitivity and specificity issues have been resolved by researchers, providing practitioners the ability to produce and change their own developed cutscores may help them make better informed decisions (LeBlanc, Dufre, &
McDougal, 2012) about to whom vocabulary intervention should be provided. The ability to predict which students will not only pass or fail but who may fall in other achievement categories (e.g. ‘mastery’ and ‘advanced’) could potentially improve service delivery for the higher level students. Although higher specificity values would increase the overall efficiency of the predictive model, it is preferable to not risk better sensitivity in exchange for enhanced specificity (Dorman, 2012).

Six, while results support the criterion-related validity and reliability of a CBM in content area world history, researchers may want to examine how VM tools can be used by school based support services (RTI) and special education teams to determine if students identified at-risk or classified with disabilities are in fact benefitting from current educational placements and/or IEP services. Johnson, Jenkins, and Petsher (2010) suggest that classification in reading increases when multiple predictive measures are used in comparison with a single screening tool. When using a combination of assessments, including VM probes, the likelihood of improving sufficiency in making accurate predications is more likely (Burns, Scholin, & Zaslofsky, 2011). The utilization of various content measures could potentially enhance the decision-making process (Scholin & Burns, 2012). If growth over time is not demonstrating sufficient progress in meeting IEP goals and CCSS, decisions made often by IEP teams, such as immediate instructional and/or placement changes, should take place including a change in students’ curricula. Aligning goals on IEPs with performance and progress in special education or with CCSS could meet current IDEIA mandates (McMaster & Espin, 2007; Nolet & McLaughlin, 2000). Such decisions can be drawn from the data that formative assessments produce; a second step in the RTI approach.
Conclusion

Results of the present study provide additional support for the utility of a five-minute, group–administered CBM for content subjects. Statistically significant correlations were reported between researcher-created VM probe scores and a statewide social studies subtest scaled score. Similar correlations were found between the probe scores and the statewide ELA, math, science, and reading subtests. Researchers may choose to investigate any connection between researcher designed social studies VM probe cutscores with other courses to further determine what VM probes are truly indicators of at the middle school level. The current study represents the first step in the development of a content-specific CBM measure as a diagnostic indicator of performance. Accurate choice of interventions is then what could ‘make or break’ a school-wide RTI program in terms of validity; poor intervention planning or weak fidelity of implementation can result in poor student responses (VanDerHeyden, Witt, & Barnett, 2005).

These findings, along with those results found in earlier literature (Bursuck, 2010; Mooney et al., 2010; Mooney et al., 2013b), provide researchers with an assessment tool that could potentially be used in the middle school grades as part of an RTI framework in content subject areas. It would be fascinating to determine whether VM could reach a point where scores illustrate for students who made progress throughout the year also scored higher on the iLEAP than those students who showed only minimal growth from week to week similar to what Ditkowsky and Koonce (2010) demonstrated with ORF probes.

Middle school RTI models demonstrate that there is little support to secondary students in content classes (Prewett, Mellard, Deshler, Alexander, & Stern, 2012) possibly because there are no benchmarks for monitoring student performance (Johnson et al., 2013). If curricula are
tied to CCSS and from which state standardized tests are designed, it makes sense that by increasing achievement on measures, like VM, there is an enhanced likelihood of increasing the number of students meeting state proficiency levels (Burns, Klingbeil, & Ysseldyke, 2010). Social studies and science teachers have had to rely on criterion referenced tests (CRTs) with minimal progress monitoring toward current learning goals (Vannest, Parker, & Dyer, 2011). Teachers need a means by which to identify students who are not going to keep up with the rigor of CCSS because they are not proficient readers of content textbooks.

An increasing focus on 21st century careers in science and technological fields illustrates that all students will be required to pass high stake science tests (Johnson et al., 2013) and warrants interventions to address struggling students. In 2005 only 29% of 4th and 8th graders scored proficient while 12th graders dropped to 18% with an even higher decline for students with disabilities (Vannest, Parker, & Dyer, 2011). Less than one third of 8th graders in 2009 scored at or above proficient levels. Missing from middle school classrooms are content area progress monitoring tools. When 67% of students with disabilities score below basic on state science tests compared to only 36% of their peers, early identification protocol for intervention is needed (Espin et al., 2013). Alarming to education researchers is that despite the identification of struggling readers and interventions implemented with fidelity in elementary schools, many of these students will continue to struggle in content area classrooms in the secondary grades (Torgesen & Miller, 2009).

In reference to research question number one: Can a VM probe, given over multiple points during the school year, predict the likelihood of meeting ‘basic’ requirements for passing a state standardized social studies content criterion assessment, results indicate that “no” it cannot due to insufficient sensitivity in predictive power. Vannest, Parker, and Dyer (2011)
suggest that curricular validity be considered in addition to a test’s sensitivity to improvement over time. Future researchers can address this problem by working more closely with social studies teachers within the classroom. It is possible that any shortfalls in instruction may have reduced probe sensitivity.

Certain accommodations may need to be provided in order for some students to access the reading material on the probes (e.g. if a student cannot read, they are unable to comprehend the probes and textbooks which may result in lower scores). Future studies that include an evaluation of modified probe validity written in other languages (e.g. Spanish), given orally via a computer, or as a recording may address these concerns (see Vannest, Parker, & Dyer, 2011). An investigation into teaching methods and vocabulary interventions that are based on probe scores is heavily warranted especially when determining whether or not they are the probes themselves that are not sensitive enough to learning or if it was the level of vocabulary instruction that negatively impacted diagnostic accuracy of scores.

In terms of predictive validity, the effectiveness of educational decisions rests on the accuracy of a CBM and screening protocol in predicting achievement and even student classifications (Burns, Scholin, & Zaslofsky, 2011). Consideration of the impact of false-positive and false-negative identification and potential for disproportionate identification rates of cultural and language minorities is highly recommended (Burns, Scholin, & Zaslofsky, 2011).

Until research identifies pre-intervention screenings that accurately identify children who need high levels of tiered interventions, teachers should only implement interventions with lower intensity while continuing progress monitoring except in rare cases (Scholin & Burns, 2012). Progress monitoring alone is not enough to address the requirements of effective formative evaluation (Burns, Klingbeil, & Ysseldyke, 2010). Additional research focusing on using probes
as dynamic measures rather than just as static measures in a middle school classroom is warranted, especially when determining if probes can be designed in such a way that is both sensitive and specific to student learning for purposes of performance prediction.

This study represents the first step in the development of a vocabulary CBM for predicting future content achievement of middle school students. Interestingly enough, while not the purpose of this study, probes 7, 11, 15, 19, and 22 scores produced correlations high in magnitude with the other 6th grade iLEAP subtests even though the content of the probes was unrelated to the material that made up those tests. Therefore, some may infer that a formative vocabulary CBM can be used in other subjects. However, one may question if VM probes in one content area are truly measuring learning progress in the area for which they were designed considering that some of the above correlations were actually higher than they were with the social studies subtest.

Results of the present study provide additional evidence for continuing VM research. Current findings, along with those results found in earlier literature (Mooney et al., 2013b), provide researchers with a solid assessment tool that can be used in middle school as a screening component within an RTI framework and as a means to evaluate the learning of CCSS. Even though predictive power is unreliable at this point, vocabulary probes could be used as an initial screening measure during levels one and two of RTI to determine appropriate levels of services. Studies examining potential cut scores ‘based on a relationship with a criterion measure with ways to improve diagnostic accuracy’ (p.110) are necessary in education (Burns, Scholin, & Zaslofsky, 2011).

“By increasing the efficiency of direct observation and reducing the labor, resources, and expertise required to use it, we may increase its utility…” (Taylor & Romanczyk, 1994, p. 252)
This is a valuable point when one considers the caseload sizes of not only middle school special education teachers, but also for RTI case managers. Probes may become a powerful tool as states implement CCSS across the board considering that vocabulary knowledge is important in understanding content.

Even though this small amount of research is potentially positive, more is needed (Burns, Scholin, & Zaslofsky, 2011) to increase the validity of VM before it may be recognized as a valid diagnostic tool. It is possible that RTI initiatives have moved assessment away from identification concerns to how the identification leads to increased learning (Burns, Scholin, & Zaslofsky, 2011). The immense growth of formative assessment will seriously alter evaluations as much as it might change the overall structuring of schooling (Dorn, 2010). This study contributed to the existing literature base because there may be potential for VM as a performance indicator. This study extended previous research, because it was the first to include a valid method for deriving cutscores through ROC on a VM-CBM.
REFERENCES


University of Georgia (n.d.). Institute for Evidence-Based Health Professions Education: Free Online Course in Evidence Based Practice. Retrieved from: http://ebp.uga.edu/courses/Chapter%204%20-%20Diagnosis%20I/8%20-%20ROC%20curves.html


Application for Exemption from Institutional Oversight

1) Principal Investigator: Jodie Schraven
Rank: Ph.D. Student
Student? Y/N Y

Dept.: ETPP
Ph: 287-1850
E-mail: jschra1@tigers.lsu.edu

2) Co Investigator(s): please include department, rank, phone and e-mail for each
Dr. Paul Mooney (ETPP) 578-2360 Assistant Professor
pmooney@lsu.edu

3) Project Title: Measuring Academic Achievement in Sixth Grade Social Studies

4) LSU Proposal? (yes or no) No
If Yes, LSU Proposal Number ______________________________________

Also, if YES, either:
☐ This application completely matches the scope of work in the grant
☐ More IRB Applications will be filed later

5) Subject pool (e.g. Psychology Students): 6th grade students
☐ Circle any "vulnerable populations" to be used (children <18, the mentally impaired, pregnant women, the aged, other). Projects with incarcerated persons cannot be exempted.

6) PI Signature
Date 2/9/09 (no per signatures)

"I certify my responses are accurate and complete. If the project scope or design is later changed I will resubmit for review. I will obtain written approval from the Authorized Representative of all non-LSU institutions in which the study is conducted. I also understand that it is my responsibility to maintain copies of all consent forms at LSU for three years after completion of the study. If I leave LSU before that time the consent forms should be preserved in the Departmental Office.

Screening Committee Action: Exempted 
Not Exempted 
Category/Paragraph
Reviewer: 
Signature: 
Date: 2/9/09
Youth Assent Form

Measuring Academic Achievement in Sixth Grade Social Studies

I have talked to my parent(s) and teacher(s) about the study. I know that I am contributing to research on social studies assessment procedures. I agree to be in the assessment study. My teacher has told me all about this study, which will involve being assessed in reading and social studies vocabulary. I agree to do all the activities of the study that I was told about. I know that I can talk to my parents or teachers if I have any concerns. I know that I can quit the study at any time without penalty. I know that my reading progress scores will be provided to both my teacher and my parent(s), if requested.

Signed: ___________________________  Date: ________________________________
    (Your Name)

Study Exempted By:
Dr. Robert C. Mathews, Chairman
Institutional Review Board
Louisiana State University
203 B-1 David Boyd Hall
225-578-8852 | www.lsu.edu/irb
Exemption Expires: 3-25-2021
Parent Consent Form

PARENT CONSENT FORM FOR YOUTH PARTICIPATION

Title: Measuring Academic Achievement in Sixth Grade Social Studies.

Project Director: Jodie Schraven, (225) 287-1850 jschrav@tigers.lsu.edu
available: anytime.

IRB Contact Information: This study has been approved by the LSU Institutional Review Board. For questions about participants’ rights, please contact the chair, Dr. Robert Matthews, 578-8692, or at irb@lsu.edu.

Purpose: The present study aims to examine the ability of multiple literacy assessment instruments (i.e., vocabulary-matching probe, two free writing probes, a reading comprehension maze probe, and the Gates-MacGinitie standardized reading assessment) to predict performance on the iLEAP social studies and English/language arts tests.

Research Procedures: A trained person will group administer each instrument to your youth and his or her classmates during a regularly-scheduled social studies class. The assessments will take place over parts of three class days in one week late in the spring semester of the 2008-09 academic year. The LSU graduate student administering the tests will be assisted by your youth’s social studies teacher. Once all assessments are completed your youth will be permitted to participate in a pizza party that will be attended by the researcher and social studies teacher.

Potential Risks: There are no apparent risks to any participants.

Potential Benefits: There are no direct benefits for any participants. Indirectly, each participant will have contributed to advancement in assessment practice research in sixth grade world history classrooms.

Participation: You are free to choose to not have your youth participate in the study. Also, your youth can quit the study at any time without penalty. You or your youth’s relationship with the school, investigators, or Louisiana State University will not be damaged in any way if you choose not to participate in the study or if your youth decides at any time to quit.

Confidentiality: The confidentiality of your youth’s reply will be ensured. Names will only be released to research team members (i.e., investigators). Data will be kept in a locked file cabinet when not being gathered.
Signature: "I have been fully informed of the above-described procedure, its possible benefits and risks, and I give my permission for my youth to participate in the study."

Parent Signature __________________________ Youth’s Name (Please Print) __________________________ Date ____________

Study Exempted By:  
Dr. Robert C. Mathews, Chairman  
Institutional Review Board  
Louisiana State University  
203 B-1 David Boyd Hall  
225-578-8692 l www.lsu.edu/irb  
Exemption Expires: 5.25.2012
VITA

Jodie Schraven, a native of Buffalo, New York, received her bachelor’s degree at the State University of New York at Buffalo State College in 1999 in criminal justice and in philosophy. Thereafter, she received her master’s degree in social work from the State University of New York at the University at Buffalo in 2002 with a focus on children and youth. She continued helping youth after receiving her MSW working as a clinical and school social worker in Erie County for four years until 2006. As her interests in special education law, data-driven assessment, and autism spectrum disorders grew, she made the decision to enter a doctoral program at Louisiana State University. She will receive her Ph.D. in Curriculum and Instruction in May 2014. She is presently a special education teacher in Las Vegas, Nevada.