2012

Using behavior screening data to predict scores on statewide assessments

Jeffrey Steven Chenier
Louisiana State University and Agricultural and Mechanical College, jcheni1@tigers.lsu.edu

USING BEHAVIOR SCREENING DATA
TO PREDICT SCORES ON
STATEWIDE ASSESSMENTS

A Dissertation

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

in

The Department of Psychology

by
Jeffrey S. Chenier
B.S., Louisiana State University, 2007
M.A., Louisiana State University, 2010
December 2012

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Federal and state initiatives (No Child Left Behind, 2001) require schools and districts to set high standards for student growth and achievement. Currently, student growth and progress are measured in Louisiana via statewide achievement tests. In 4th and 8th grades these assessments are considered to be ‘high-stakes’, as promotion and retention decisions are made based on how well students perform on these assessments. Making day-to-day decisions based on one assessment per year is not best practice (Jenkins, Deno, & Markin, 1979); therefore, screening instruments known as curriculum based measures (CBMs) were devised and tailored for school-based implementation. CBMs of academic skills have been shown to predict scores on statewide achievement tests (e.g. Good, Simmons, and Kameenui, 2001; Shaw & Shaw, 2002; Keller-Margulis, Shapiro, and Hintze, 2008). However, less research has been conducted using behavior screening instruments, despite the fact that the relationship among behavior and academic achievement has been extensively documented. The current study adds to the literature base by assessing the predictive validity of commercially available behavior screening instruments for statewide achievement test scores in a school district in Louisiana. Results show that two of four behavior screenings within the program are independent predictors of statewide testing scores in addition to academic screenings and prior achievement in their respective content areas. Implications of these findings are that it may prove beneficial for schools to proactively screen for and intervene with behavior problems as early and frequently as possible.

INTRODUCTION

NCLB and Accountability

The No Child Left Behind Act (NCLB, 2001) requires states to hold schools accountable for their students' academic progression. NCLB called for states to set standards for what students should know in addition to goals by which the state, districts, and schools can measure students' progress. In Louisiana, the accountability system uses annual test scores as part of its protocol to assign schools, districts, and the state a performance score. Students in grades 3-8, the focus of this study, are tested annually based on Grade-Level Expectations (GLEs), and different grades' tests have different implications. GLEs, in Louisiana, "identify what all students should know or be able to do by the end of each grade from prekindergarten through grade 12 in math, English, science, and social studies" (Louisiana Department of Education, 2011). Students in grades 4 and 8 take the *Louisiana Educational Assessment Program* (LEAP), which is considered a high stakes assessment for the student, due to the fact that scores on this test aid in the determination as to whether he/she passes, needs to attend summer school and take portions of the test again, or is retained. Students in grades 3, 5, 6, 7, and 9 take the *Integrated Louisiana Educational Assessment Program* (*i*LEAP), which assesses students in the same content areas as the annual LEAP test, but promotion or retention decisions are not made based on performance on this measure.

In addition to the student-level implications, in Louisiana, schools are assigned numerical scores known as School Performance Scores (SPS; LADOE, 2011). These scores are calculated using student test scores (90%) and attendance (10%) for schools with grades K-6. Test scores (90%), dropouts (5%), and attendance (5%) determine the SPS for schools with grades 7-8. Finally, high schools receive SPS based on test scores (70%) and Graduation Index (30%) (LADOE, 2011). Schools may receive a score anywhere from 0-200. Louisiana, for the first time in 2010-2011, assigned letter grades to these scores as well. Scores from 0-64.9 received an F, 65.0-89.9 a D, 90.0-104.9 a C, 105.0- 119.9 a B, and 120.0-200.0 an A. Schools are also assigned a plus or minus, depending on whether the school met their state assigned growth target (it should be noted that Louisiana has recently been granted a waiver from NCLB; therefore, the grading rubric is subject to change). Schools that perform well may receive recognition and additional funding from the state; and if the score is low enough for a school to be considered *Academically Unacceptable* across multiple years, the school is at risk for losing funding and eventually being taken over by the state's Recovery School District (RSD). The RSD is state-run and "designed to take underperforming schools and transform them into successful places for children to learn" (Louisiana Recovery School District, retrieved from http://www.rsdla.net/Home.aspx). Teachers may also be at-risk of eventually losing their jobs if students are not showing sufficient growth on these measures.

RTI to Increase Data-Based Decision Making

Schools and districts are currently assigned scores that judge their overall performance based primarily on students' performance on a single test. Considering that the implications of these tests extend from molecular to molar levels (i.e. implications for individual students/teacher and implications for entire school districts), schools are ultimately responsible for identifying and intervening with at-risk students as early as possible in their educational careers. Standardized, high-stakes assessments do not provide information regarding student performance until the end of the school year (McGlinchey & Hixson, 2004). Good, Simmons, and Kame'enui (2001) state that students as well as teachers should be given feedback constantly throughout the school year, so that methods and techniques that are effective can continue to be used and methods and techniques that are not working can be removed.

In order to accommodate these recommendations, the current method for identifying these students has shifted from a wait-to-fail system to proactive, universal screenings of entire schools in order to determine needs of students more frequently. Universal screening is a cornerstone in the current framework for providing services to students called Response to Intervention (RTI). The National Center on Response to Intervention states that RTI uses screening data to "identify students at risk for poor learning outcomes, monitor student progress, provide evidence-based interventions, adjust the intensity and nature of those interventions depending on a student's responsiveness, and identify students with learning disabilities or other disabilities" (National Center on Response to Intervention, 2010).

RTI is typically conceptualized as a tiered framework of service delivery based on the needs of the student in particular content areas. NCLB calls for scientific-based instructional practices and interventions to be used throughout the tiers, which is a hallmark of RTI. Additionally, within an RTI model, a student should receive more or less intervention in a content area based on his/her responding to an evidence-based intervention that is implemented with integrity (Gresham, 2005). RTI is a tiered mechanism of service delivery. Tier 1 in RTI consists of the general education curriculum which each student in the school receives. Using RTI, students are screened multiple times per year in order to determine if they are acquiring and performing the skills taught through the general curriculum at a level commensurate with either peers at the same school or a criterion set by the screening measure. If the student is not making adequate progress based on screening data, the student progresses to Tier 2, which is more focused instruction, typically via a small pull-out group in a particular academic area, in addition to continuing to receive the general curriculum. Students in Tier 2 are 'screened' more frequently to determine whether the services are helping the student grow toward the criterion that he/she failed to meet in the original screening. This practice is called progress monitoring. Tier 2 interventions need to be changed if a student fails to grow at a quick enough rate to catch the student up to the criterion. If the student's data from Tier 2 shows inadequate progress, the student is moved to Tier 3. In Tier 3, the services a student receives are more intense, which could mean that more individuals are involved in providing services (both in and out of school) and/or that time devoted to these services is increased. The student still receives the services provided in Tiers 1 and 2; and progress monitoring continues, often at a more frequent rate. To summarize, RTI uses a problem solving model in determining whether differences between baseline and post-intervention are sufficient to a degree to call "response" (Gresham, 2005). Screening and data collection throughout this process is the backbone of RTI, as the student's data is used to make decisions in respect to what services the student receives.

What Makes a Useful and Sound Screening Instrument

In order for screening instruments to be useful for decision making, the instruments must have sound psychometric properties (evaluated via reliability and validity), must have sound predictive validity by being able to identify true positives and negatives while failing to identify false negatives or false positives, and be both efficient and cost-effective.

The validity of an instrument, according to Messick's unified theory of validity (1989), is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p.13). The reliability of an instrument "refers to its degree of stability, consistency, predictability, and accuracy" (Groth-Marnat, 2009, p. 11). Another method by which to define reliability is the degree to which that instrument will detect similar results across different administrations over time.

A frequently used method by which to evaluate and interpret the predictive validity of tests/measures are conditional probability analyses, also known as the sensitivity, specificity, and predictive value model or diagnostic efficiency calculations (Kettler, Elliott, Davies, & Griffin, 2011; Glaros & Kline, 1988). This model allows for a specific score on a measure to function as a cut-off to predict whether an individual would either qualify or not qualify for a specific condition. This model was originally utilized in the medical field with laboratory screening procedures (Glaros & Kline, 1998; p. 1013). In this model, the binary outcome allows for a measure to use a cut-score to predict an individual case in four different ways. A "true positive" signifies the measure both predicted presence of a condition, and the person has that condition. A "false positive" then would be that the measure predicted the person having the condition, but the person does not have the condition. A "false negative" signifies that a person that has the condition, but he was identified by the measure as not having the condition. Finally, a "true negative" on a measure signifies that that a person is identified by a measure as not having a condition when he actually does not have a condition. Methods to quantify these results include sensitivity, specificity, positive predictive value, and negative predictive value. Sensitivity is the ability for a measure to accurately identify a condition when an individual actually has that condition. Specificity is the ability for a measure to accurately identify when an individual does not have that condition. These statistics are typically reported as percentages. The predictive values in this model are divided into positive predictive value and negative predictive value, where positive predictive value is the likelihood that an individual who tests positive actually has that condition. Negative predictive value is the likelihood that an individual who tests negative actually does not have that condition. Positive and negative predictive values are important to consider due to the fact that, typically, diagnostic, eligibility and assessment decisions are made based on a single individual's score. A psychologist or clinician would want to know how confident he/she could be in assigning a student to a condition based on a test score (p. 1015). Glover and Albers (2007) remark that measures reporting indices of sensitivity, specificity, and predictive values of below .75 or 75% should be utilized with caution; and Shapiro, Keller, Edwards, Lutz, and Hintze (2006) used .60 or 60% as a criteria to evaluate screening instruments.

Witt (2007) remarked that screening tools help guide schools intervention decision-making by using the "least dangerous assumption." Should screeners not be able to identify all students' scores or condition as either true positives or true negatives. Witt postulated that screeners should identify more students who may potentially need intervention (false positives) at the expense of minimizing false negatives. While an excessive amount of false positives presents problems (taxation of school resources and/or mislabeling a student), failure to identify a student using screening that actually needs intervention is unacceptable given the provisions of NCLB (Witt, 2007). Schools cannot recover the time lost between screening periods should a child actually need intervention.

Finally, screening measures must be time and cost efficient. Screening measures should be able to be frequently administered and sensitive to change in order for the data to be utilized to make frequent decisions regarding student progress (Hosp, Hosp, & Howell, 2007). Additionally, given budget constraints across the country, screening measures should be low-cost both monetarily and for staff resources.

CBMs for Academic Screening within RTI

Screening for Academics. The majority of the literature on screening in schools has been dedicated to the academic domain (Cook, Volpe, & Livanis, 2010). Screening for academics is typically done using curriculum-based measurements (CBMs), which are quick, reliable, and valid methods of assessing students' functioning in basic skill areas deemed to be critical for student success (Deno, 1985). These measures are shown to represent general outcome measures (GOMs), which are indicative of a student's overall functioning in the domain being assessed, rather than in a particular skill. CBM has been demonstrated to be an accepted method of screening students in academics, identifying potential strengths and weaknesses, and subsequently progress monitoring within interventions (Foegen, Jiban, & Deno, 2007). The most common academic domains assessed using CBM are reading, mathematics, and writing (Marston, 1989; Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002).

CBM has been adopted as a screening mechanism (an "academic thermometer") in an RTI model due to numerous advantages that CBM has compared to the utilization of standardized achievement tests. First, achievement tests may not sufficiently measure what is actually being taught in a particular student's general curriculum (Jenkins, Deno, & Markin, 1979; Jenkins & Pany, 1976). CBMs are designed to mimic the format and content of what is being taught and thus directly assess the student's current curriculum (Gansle, Noell, VanDerHayden, Slider, Hoffpauir, & Whitmarsh, 2004; Jenkins et. al., 1979). Additionally, standardized achievement tests cannot be administered at a frequent enough interval to appropriately inform decision-makers whether a student's curriculum is appropriate on that particular day, week, etc. (Jenkins et. al, 1979). Jenkins and colleagues remark that data may need to be available at least daily in order to evaluate whether a student's curriculum is appropriate (1979). CBMs are structured to be given more frequently, due to numerous reliable and valid different probes (Gansle et. al., 2004). Finally, CBMs take much less time to both administer and score compared to achievement tests (Gansle et al, 2004), which is appreciated in school systems where a single individual likely has multiple responsibilities.

Academic Screening Predicts Scores on Statewide Assessments. Beyond their utility for screening and monitoring progress in the academic domain, data from CBMs have been found to correlate with and predict performance on statewide assessments. Shaw and Shaw (2002) administered the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency (ORF) CBM at three intervals (fall, winter, spring) during the 2001-2002 school year to a sample of 58 third-grade students. The authors describe this ORF assessment as three passages read aloud for one-minute; with errors considered to be “words omitted, substitutions, and/or hesitations for more than three seconds.” Each word read correctly is scored and the sum of words read correctly is the final outcome measure. The authors found that 91 percent of students scoring at or above 90 words in a minute also scored at “proficient” or “advanced” on the Colorado State Assessment Program. Likewise, 73 percent of students who scored below 90 words per minute scored “unsatisfactory” or “partially proficient.” 86 percent of students were classified correctly as either “proficient/advanced” or “unsatisfactory/partially proficient” simply based on their DIBELS ORF.

These findings have been replicated across different states. Good, Simmons, and Kameenui (2001) found that 96 percent of students who met benchmark criteria for CBM ORF performed at criteria or beyond on the Oregon Statewide Assessment, while 72 percent of students who did not meet CBM benchmark criteria performed below criteria on the statewide assessment. . Buck and Torgesen (2003) found that 91 percent of students who read at or above 110 words per minute on CBM ORF scored at or above adequate on the reading subtests of the Florida Comprehensive Assessment Test- Sunshine State Standards (FCAT-SSS), and 81 percent of students who were at “high risk” (reading less than 80 words per minute) based on their ORF CBM performance did not score at or above adequate on the FCAT-SSS. Hintze and Silberglitt (2005) found this relationship for ORF and performance on the Minnesota Comprehensive Assessment for students from 1st through 3rd grade. Another example of this is from Ditkowsky and Koonce (2010), in which ORF predicted reading scale scores on the Illinois Standards achievement (ISAT). These authors also found that as students receiving special education services progressed in ORF, their chances of passing these statewide assessments increased.

Stage and Jacobsen (2001) used ORF on state developed and normed passages to determine whether students passed or failed the Washington Assessment of Student Learning (WASL). The authors found that their set cut score had a sensitivity of 76 percent, meaning that 76 percent of students who passed the WASL scored above the cut score. The specificity, the percent of students who failed the WASL when scoring below the cut score on ORF, was 66 percent. ORF had a positive predictive value of .90 and a negative predictive value of .41, and the overall hit rate of correct classification was 74 percent. The authors noted that ORF increased the ability to predict passing or failing the WASL by 30% over the base rate. McGlinchey and Hixson (2004) replicated the Stage and Jacobsen (2001) study in Michigan using the Michigan Educational Assessment Program (MEAP) as the criterion. The probes used were from the Macmillan Connections Reading Program (Arnold & Smith, 1987). Probes were administered across eight grade levels during the final two weeks before the MEAP was taken. The authors used 100 words per minute as their cut score. The sensitivity of 100 words per minute to identify students who scored at “satisfactory” or above on the MEAP was 75 percent, and the specificity of 100 words per minute to identify students scoring below satisfactory was 74 percent. The positive predictive value was 77 percent, the negative predictive power was 74percent, and the overall correct classification was 74 percent (McGlinchey & Hixson, 2004).

Yeo (2010) conducted a meta-analysis investigating prior research regarding reading CBM and scores on statewide reading assessments. Yeo's analysis came from 27 studies that met inclusion criteria of (but not limited to) CBM probes administered before the statewide assessment, a group design, and sufficient data provided to calculate effect sizes (p. 419). Also, articles that were not peer-reviewed were eligible for inclusion in the analysis, which may call into question the overall validity of the findings. Yeo found that there was a strong relationship between reading CBM and scores on reading portions of statewide tests (r = .69). Yeo also investigated whether there were moderating variables involved with this large effect. One finding was that studies that included high amounts of students with disabilities or English Language Learners in their sample size saw a reduced correlation coefficient. Another finding was that as time increased between administration of reading CBM and the statewide tests, the correlation coefficient decreased. A final finding was that this large effect stayed consistent across different states, which included different types of reading CBM passages (commercially available vs. state-generated) and standardized tests (multiple choice vs. multiple choice and open-ended questions) (pp. 419-420).

Relationships among measures of math CBM and achievement on statewide assessments have been demonstrated as well. Helwig et al. (2002) found a strong relation (r = .80) between scores on math CBM probes with conceptual problems and performance on a test that mimicked the Oregon statewide assessment. Shapiro et al. (2006) found similar relationships between math CBM probes and performance on the Pennsylvania statewide assessment. The authors used receiver operating characteristic (ROC) analyses to predict whether students would pass or fail the statewide assessment based on their scores on math CBM and found a sensitivity and specificity of .65. Jiban and Deno (2007) found similar results for math CBM, as it explained 52 percent of the variance in 5th grade and 27 percent of the variance in 3rd grade on the Minnesota Comprehensive Assessment in Mathematics. Finally, an unpublished dissertation by Menessess (2011) found that math CBM probes correctly classified between 61 and 72 percent of 3rd-5th grade student's scores on math portions of Louisiana statewide assessments.

A study by Keller-Margulis, Shapiro, and Hintze (2008) used ORF, math computation probes, and math concepts and applications probes to investigate the predictive ability of CBM on the Pennsylvania statewide assessment. ORF probes from AIMSweb® correctly classified 78 percent of students' passing or failing on the assessment. Math CBM probes, which consisted of math computation and math concepts and applications probes from AIMSweb®, were able to classify students as passing or failing the assessment with 70 percent accuracy.

Writing CBM has also been shown to correlate with scores on statewide assessments. Gansle et. al. (2002) demonstrated moderate to strong relationships using the most common scoring methodologies of three-minute writing CBM probes and different statewide assessments in Louisiana for students in 3rd and 4th grade. An unpublished dissertation by Henderson (2009) found similar relationships for commonly used scoring methods of writing fluency and scores on statewide testing for elementary aged students in Louisiana. Jewell and Malecki (2005) also found strong relationships between writing CBM and scores on the Stanford Achievement Test (SAT; Harcourt Brace Educational Measurement, 1997) for students in grades 2, 4, and 6. Epsin, Wallace, and colleagues (2008) had 10th grade students write for 10 minutes and scored the probes at 3, 5, 7, and 10 minutes using three different scoring methods. The authors found that scoring probes after 7 minutes using Correct Minus Incorrect Word Sequences (CMISs) was a reliable predictor of scores on the written expression section on the Minnesota Basic Standards Test/Minnesota Comprehensive Assessments (MBST/MCA).

These findings are significant in that screening an individual student in reading, math, and writing can take under 10-minutes to complete. The ability of screening measures to predict in the fall how a student may score on a statewide assessment given in the spring allows for appropriate goals and interventions to be utilized for that student within RTI, in addition to their utility in instructional planning and progress monitoring,

Behavior Screening and Response to Intervention. The previous studies have opened the door for utilizing reading, math, and writing CBM beyond universal screening and progress monitoring in an RTI model. The current study seeks to explore whether behavior CBMs can serve the same function.

Walker, Ramsey, and Gresham (2004) estimate that close to 20 percent of children in schools are at-risk for developing behavior problems, and that only 1-3 percent of those students are receiving appropriate services. Proper screening for these students should increase the services provided to these students. Severson, Walker, Hope-Doolittle, Kratochwill, and Gresham (2007) discuss a number of developments over the past decade that brought about the movement to screen for behavior and emotional problems in schools. The authors list the first development as the “shock and trauma” that the school shootings of the 1990s, such as Columbine, produced in America. These shootings forced lawmakers to fortify schools and identify potential students who could potentially perform such violent behaviors in the future. These shootings also may have contributed to NCLB suggesting to proactively screen for and intervene with students at-risk for both academic and behavior problems (2007), in addition to legislation requiring states to screen all children who are Medicaid eligible for social/emotional concerns (Rosie D. vs. Romney, 2006). The second development cited by the authors is pressure from the community for more visible returns on their investment in research-based interventions for mental health issues in school-aged children. The third and final development is the adoption of school systems' use of multi-tiered models of prevention for academic and behavior problems. Screening is fundamental for this process to be successful (Severson et. al., 2007).

An example of a state incorporating this practice is Louisiana's adopting an RTI model as a requirement in the multi-disciplinary assessment process for exceptionalities that include a behavioral concern. For example, within the criteria for Emotional Disturbance, *Bulletin 1508: Louisiana Pupil Appraisal Handbook* (2009) states that educational performance must be significantly affected *and* "behavioral patterns, consistent with the definition, exist after behavior intervention and/or counseling and educational assistance implemented through the RTI process which includes documented research-based interventions targeting specific behaviors of concern (p. 31)." As previously stated, at the forefront of the RTI process for both academics and behavior is proactive, universal screening (Fuchs, Fuchs, & Speece, 2002).

In order to carry out the RTI process for behavior, most public schools utilize implementation of multi-tiered approaches such as School-Wide Positive Behavior Support (SWPBS). These types of programs are reported to be in place across 30 states and 7,900 schools in the United States alone (Spaulding, Horner, May & Vincent, 2008). SWPBS programs are incorporated into schools' RTI models of proactive, evidence-based intervention (Sugai & Horner, 2009). Sugai and Horner (2009) state that SWPBS has five core components: behavioral theory and applied behavior analysis, focus on prevention, instructional focus, evidence-based behavioral practices, and systems approach.

Multi-tiered systems of behavior support look similar to tiered systems of support for academics. Tier 1 consists of a school's universal approach to defining school expectations and means by which to reward those students who comply. Typically, in SWPBS programs, schools post expectations of students in each area of a school and students who are caught behaving appropriately are rewarded using a token economy system (PBS bucks linked to the ability to "purchase" preferred items at a PBS store). As with academics, students are screened multiple times per year to determine whether their response to an evidence-based system of school-wide/classroom management is sufficient. Tier 2 interventions are implemented for students who do not respond appropriately to the Tier 1 program. These interventions are typically delivered by the classroom teacher within the classroom as designed through consultation from a school psychologist or other team member trained in behavioral interventions. As with academic interventions, continuous data collection guides decision making in regards to the student's response to this intervention. Tier 3 for behavior typically calls for a Functional Behavior Assessment to inform a Behavior Intervention Plan, as well as wraparound services such as counseling services for the student or services provided to the family through inter-agency coordination. The effectiveness of SWPBS programs has been investigated by a handful of researchers. A meta-analysis evaluating SWPBS by Solomon, Klein, Hintze, Cressey, and Peller (2012) found effect sizes across categories such as outcome variable, setting, duration, type of intervention, grade level, and demographic to range from $r^2 = .27$ to $r^2 = .60$

Behavior Screening Instruments. As previously mentioned, the current study seeks to evaluate whether behavior screenings could lend information as to how students would achieve on yearly, statewide testing. Until recently, most research on the development and utilization of behavior screening instruments has focused on identifying students at-risk for displaying externalizing behavior problems (Cook et. al., 2011). Some of the more widely used screeners for behavior are the Systematic Screening for Behavior Disorders (SSBD; Walker & Severson, 1990), Student Risk Screening Scale (SRSS; Drummond, 1994), Strengths and Differences Questionnaire (SDQ; Goodman, 1997), Student Internalizing Behavior Screener (Cook et. al., 2011), Social Skills Improvement System: Performance Screening Guide (PSG; Elliott & Gresham, 2007); and the BASC-2 Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2008).

The SSBD uses multiple gating through three stages moving from teacher nomination to questionnaires to direct observations of students who pass through the first two gates (Walker & Severson, 1990). The SSBD was originally normed for students in grades K-6, but Calderella et. al (2008) normed the instrument for middle school students as well. The Student Risk Screening Scale (SRSS) is a teacher-completed externalizing behavior screener which takes approximately 10 minutes per class. The SRSS has teachers rate each student using a 4-point likert scale on 7 different behaviors; and if the student's score is above a pre-determined cut score, the student is at-risk for developing future externalizing behaviors without further intervention. Like the SSBD, the SRSS was normed originally with students in grades K-6, and Lane and colleagues (2008) normed the instrument to be used with middle and high school students. The SIBS is similar in format to the SRSS, except that its function is to identify students who are at-risk of developing an internalizing behavior problem (Cook et. al., 2011). The SDQ is a behavior screener that can be used with children ages 3-16. The SDQ has been researched with clinic samples in the UK and has been found to effectively identify children at risk for developing psychiatric symptoms (Goodman, Ford, & Simmons, 2000).

The final two measures are utilized as the independent variables for this study. The BESS and PSG are included in the AIMSweb® Behavior module, which is a commercially available web-based program that can be utilized by schools and districts to track behavior screening data and monitor intervention data based on items from the screeners. These measures are marketed as CBMs for behavior.

The BESS consists of teacher and student forms that can be completed either on-line or using pen and paper. The teacher form of the BESS contains 27 items for which teachers rate their students using a 4-point likert scale (Never, Sometimes, Often, Always occurring). The form was normed with students from Pre-Kindergarten through 12th grade. The student form contains 30 items and requires a 3rd grade reading level so is therefore normed for students in grades 3-12. Items on the BESS target both externalizing and internalizing behavior problems, as well as academic and social problems. The authors report that once the teacher/student is familiar with the form, it should take approximately 3-5 minutes per form per student to complete. The data is entered via the web-based module, and t-scores (M=50; SD=10) are produced for each form. T-scores below 61 are interpreted as “Meets or exceeds basic expectations,” t-scores from 61-70 are interpreted as “Consider need for individualized instruction,” and t-scores above 70 are interpreted as “Consult with behavior specialist.” The BESS’s reliability and validity information are presented in the methodology section.

The PSG was developed as a universal screening instrument for behavior focusing on four areas: Prosocial Behavior, Motivation to Learn, Reading Skills, and Math Skills (Elliott & Gresham, 2007). This measure was developed to accompany the release of the Social Skills Improvement System – Rating Scales (SSiS-RS; Gresham & Elliott, 2008), which is the revised and re-normed edition of the Social Skills Rating Scales (SSRS; Gresham & Elliott, 1990). The

Social Skills Improvement System was released with a Classwide Intervention Guide and the PSG (Elliott & Gresham, 2007) in order to have a means by which to teach social skills at the universal level. The PSG was developed to quickly screen students in the aforementioned areas pre- and post-intervention in order to determine whether further intervention is needed beyond the classroom program. The Prosocial Behavior and Motivation to Learn areas are included in the AIMSweb ® Behavior module. The PSG is a teacher-completed form that takes approximately 20 minutes per class to complete. For the PSG, teachers rate each student in their class on a 5-point likert scale [Very limited/extreme difficulty/poor (1), Frequent difficulty/limited/little (2), Occasional difficulty/somewhat less than expected (3), General competence/adequate/ appropriate (4), and excellent/high (5)] for behaviors described to define prosocial behavior or motivation to learn. The authors define prosocial behavior as "behavior directed toward other people that involve effective communication skills, cooperative acts, and self-control in difficult situations (2007).

The authors define motivation to learn as "a state of excitement and activity directed toward learning and completing classroom tasks or activities" (2007). A score of 4 or 5 is interpreted as "Meets or exceeds basic expectations," a score of 2 or 3 is interpreted as "Consider need for individualized instruction," and a score of 1is interpreted as "Consult with behavior specialist." The current study will use the motivation to learn subscale of the PSG to determine if teacher-ratings of students' motivation add to the prediction of scores on a statewide assessment through behavioral measures. Motivation has been hypothesized as key factor in learning and competence in a specific skill area (Sternberg, 2005; Wentzel, 2005). In screening using CBMs, consideration is taken as to whether the presenting problem is a skill acquisition deficit, "can't do" problem, or a skill performance deficit. Interventions that target skill acquisition deficits actually teach the skill to the student. Interventions that target skill performance deficits, better known as "won't do" problems, typically alter the student's environment so that reinforcement is removed for the maintaining behavior that one wishes to decrease and is added for the behaviors that one wishes to increase (Gresham & Elliott, 2008). Therefore, it could be said that teacher-rated motivation is a subjective judgment of natural reinforcement that a student receives for performing well academically. In a pilot study using the Social Skills Improvement System – Classwide Intervention Program (Elliott & Gresham, 2007), it was discovered that teacher's ratings of students' motivation to learn at the beginning of a 10-week class-wide social skills intervention was significantly related to increases in prosocial behavior ratings following the intervention ($F(4,367) = 4.47$, $p<.05$) with pre-intervention scores for prosocial behavior used as a covariate (Patty, Hunter, & Chenier, 2011). It was hypothesized that teacher-rated motivation would generalize to performance in academic subjects as well.

Relationship between Behavior and Academic Achievement

The current study seeks to investigate the utility of behavior screening beyond identifying students at-risk for developing social-emotional problems and informing intervention. The research documenting the relationship between behavior and academic achievement is extensive.

Relationship between Social Behavior and Academic Achievement. The theory that social behavior and academic achievement may be directly related is linked to the work of Vygotsky (1978) and Bandura (1997) and the idea of social learning (Malecki & Elliott, 2002). Children learn through observing their peers and either listening to those peers or copying their behaviors (2002 p. 2). These researchers postulated that children learn whether certain behaviors their peers exhibit are either reinforced or punished, and this theory extended into the academic domain, where students who learned to work cooperatively with peers and teachers would exhibit higher levels of academic learning (2002 pp. 2-3).

The literature base that links high levels of social-behavioral competence with increased academic achievement is extensive. Feshbach and Feshbach (1987) found that teacher ratings of students' empathy when they were 8 or 9 years old were related to those same students' academic achievement when they were aged 10-11. Soli and Devine (1976) found that observed behaviors such as initiating to the teacher, self-stimulation, and positive social interactions were able to predict academic achievement in reading and math in third and fourth grade students. Cobb (1972) found similar relationships with specific on- and off-task behaviors and scores on both arithmetic and reading/spelling subtests on the Stanford Achievement Test (SAT). Lambert and Nicoli (1977) used correlation and regression statistics to demonstrate that "nonintellectual characteristics" of children, such as teacher ratings of whether students get in fights, are easily distracted, and have no enthusiasm toward school can negatively predict performance on reading assessments.

Wentzel's (1991, 1993) research has demonstrated direct positive relationships between prosocial behavior and both achievement scores and grade point average. In 1991, she used regression analyses to show that socially responsible behavior in 12-13 year old students is significantly related to student's grades when accounting for their IQ, sex, ethnicity, school absence, and family structure. She noted that socially responsible behaviors may foster an environment in which student's social goals align with academic goals. In 1993, she found that prosocial behavior was a significant, independent positive predictor of student's GPA; and she found that antisocial behavior was a significant, negative predictor. Other variables that positively predicted GPA were academic behaviors, IQ, and family structure. Prosocial behavior was also a positive predictor of standardized test scores. A discussion of the directionality of these correlations led Wentzel to argue that level of social competence may be predicting achievement, due to multiple factors. First, since achievement scores are not typically disseminated to students, it would not be expected that higher scores on these tests would foster more positive interactions. Second, IQ did not predict prosocial behavior (Wentzel, 1991). Finally, she remarked that interventions targeting social behaviors have collateral effects of increasing achievement scores or grades; but there is less evidence that interventions targeting achievement scores or grades have as strong of a collateral effect on social behavior.

Agostin and Bain (1997) used the Social Skills Rating System (SRSS; Gresham & Elliott, 1992) along with a screening tool, the Early Prevention of School Failure (EPSF, George & Wilkeson, 1989), to predict achievement scores on the Stanford Achievement Test (SAT) and grade retention/promotion in kindergarten and first grade students. Results of the 2 year study found that the Cooperation and Self-Control subscales of the SSRS, along with a measure of fine motor skills from the EPSF, were three of the four variables that accounted for the most variance when the model significantly identified students as at-risk for academic failure. The combination of assessment instruments correctly identified 76.2 percent of students as at-risk for either being retained or having low achievement scores.

Caprara, Barbaranelli, Pastorelli, Bandura and Zimbardo (2000) acquired both prosocial ratings and academic achievement scores of 294 3rd graders in Rome, Italy, in order to determine a model of academic achievement of these students in 8th grade. To acquire a rating of prosocial behavior, students rated themselves on a 10-item scale, students rated other students sociometrically, and teachers rated the students on the same 10-item scale. For academic achievement, the students had six different teachers grade them in each of their six courses, compiling a comprehensive score. Using structural equation modeling, the authors found that academic achievement in 8th grade was predicted robustly by their 3rd grade prosocial behavior score, with an impact coefficient of .52. The authors also found that the impact of 3rd grade prosocial behavior was independent of those students' academic achievement in 3rd grade, and 3rd grade academic achievement was not significantly related to 8th grade academic achievement (p. 304).

Malecki and Elliott (2002) investigated this relationship in 139 students in grades 3 and 4. The students in this study were a diverse sample, with 54 percent female, 46 percent male; 69 percent minority, 31 percent white; and 95 percent of students qualifying for free or reduced lunch prices. These students were assessed in the fall and spring using the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) to assess teacher and student ratings of both social skills, problem behaviors, and academic competence, and using the Iowa Test of Basic Skills (ITBS; Hoover, Hieronymus, Frisbiw, & Dunbar, 1993) to assess academic achievement. The authors found similar results to Wentzel: teacher ratings of social behavior were related to academic variables as measured by the ITBS. Additionally, the authors found that student self-ratings of social competence and their ITBS scores were not significantly correlated. Like Wentzel (1993), the authors also found that Problem Behavior ratings were associated with lower academic scores, although these ratings were not a significant predictor of achievement scores when entered into a multiple regression. Finally, using regression analyses, the authors found that teacher ratings of social skills accounted for a significant amount of the variance in those teachers ratings of academic competence; and teacher ratings of academic competence significantly predicted academic achievement (p. 15).

Fleming, Haggerty, Catalano, Harachi, Mazza, and Gruman (2005) utilized behavior ratings of students in 7th grade to predict achievement in 10th grade. Ratings completed by teachers consisted of the antisocial behavior and attention regulation scales from the Walker-McConnell Scale of Social Competency and School Adjustment (Walker & McConnell, 1988) and the Achenbach Teacher Report Form (TRF; Achenbach, 1991). Academic achievement was measured in 10th grade by the Washington Assessment of Student Learning (WASL, 1988). The results showed that increased levels of attentiveness, peer relationships, and pro-social behaviors, as rated by teachers in 7th grade on the aforementioned measures, were significant predictors of scoring higher on the WASL (p < .05).

The previous studies demonstrate a relationship between teacher ratings of behavior and scoring higher on different tests of achievement. While the previous studies have documented increases in academic achievement scores relative to increases in prosocial behavior, other research has shown a negative relationship between externalizing behavior problems and scores on measures of achievement.

Relationship between Externalizing Problem Behaviors and Academic Achievement. Externalizing behavior problems refer to “under-controlled behaviors,” including attention problems, disobedience, aggression, and deliberate rule violation.” (Walker, Ramsey, & Gresham, 2004). In addition to having poor academic achievement, children with externalizing behavior problems are more likely to be rejected by their peers and display substance abuse (Fergusson, Horwood, & Ridder, 2005). These problems are distinguished from “over-controlled” internalizing behavior problems, which include behaviors such as social withdrawal, anxiety, depression, and somatic complaints (Sourander & Helstela, 2005). A major differentiator between externalizing and internalizing behavior problems is the amount of attention given to these types of problems in the classroom. Thomas, Presland, Grant, and Glynn (1978) remark that the extant literature suggests that teachers spend much more time addressing children exhibiting externalizing behavior problems. Internalizing behavior problems are often overlooked by teachers as behaviors consistent with internalizing problems actually mirror behaviors of the "*ideal* student: docile, quiet, and still" (Cook et al., 2010; Walker, Ramsey, & Gresham, 2004; Winett & Walker, 1972), while externalizing behavior problems are much more overt and call for teacher and staff attention to correct.

A large amount of research has been focused on the relationship between externalizing behavior problems in children and adolescents and substandard academic or intellectual functioning. Following a review investigating comorbidity among externalizing behavior problems and poor academic outcomes, Hinshaw (1992) stated that overlap among the two constructs are too significant to be simply due to chance. Hinshaw's review stated that students with academic deficiencies typically show externalizing behavior problems in the classroom as well. Metzler (1984) compared 53 'delinquent' adolescents (aged 13-16) who were committed to the department of Youth Services of the Commonwealth of Massachusetts to adolescents (mean age 14.6) who were enrolled in a junior high school in Watertown, Massachusetts on an educational inventory that assessed student's abilities in reading, spelling, written expression, and mathematics. File reviews for the 'delinquents' were conducted in addition to histories obtained through parent interviews. Following the assessments, the groups of students differed significantly on reading accuracy, reading comprehension, spelling, mathematics, and reading rates, as well as grade-level equivalents as estimated by the educational inventory. Results of the parent interviews revealed that the delinquent group displayed delays in academics as early as second grade, and one-third of the delinquent group had been retained by their third grade year.

Richards, Symons, Greene, and Szuszkiewicz (1995) hypothesized that for 43 students, ages 11-17 enrolled at a private school for students with learning disabilities, the relationship between externalizing behavior problems and academic achievement may actually be bi-directional. These students' parents and teachers completed the Children's Attention and Adjustment Survey (CAAS, Lambert, Hartsough, & Sandoval, 1990) and the Child Behavior Checklist (CBCL; Achenbach, 1983). The authors divided their sample into two cohorts based on the year they enrolled in the school. Cohort 1 was in their second year, and cohort 2 was in their first year at the school. Data was collected for the first year of both cohorts 1 and 2 and for the second year for cohort 1. Regression analyses showed that teacher ratings of inattention in year one was significantly negatively related to reading achievement and spelling achievement measures as estimated by the Wide Range Achievement Test (WRAT-R) for both cohorts. Ratings of externalizing behavior problems on the CBCL and TRF accounted for as much as 39 percent of the variance in predicting academic achievement in the following year for cohort 1, and ratings of internalizing behavior problems did not significantly contribute to the model.

McIntosh, Horner, Chard, Boland, and Good (2006) used number of major office discipline referrals (ODRs) and reading CBM in students in grades K, 2, and 4 to predict number of major ODRs in those same students when they were in 5th grade. The authors used logistic regression analyses in order to determine response to SWPBS in these 5th grade students based on the aforementioned predictor variables. The authors found that ODRs (R=0.56, OR=0.99, p < .0005) and ORF (R=0.30, OR = 1.63, p < .0005) from the students' 4th grade year predicted whether students received 2 or more major ODRs in their 5th grade year. The authors also found that ODRs (R=0.13, OR=1.20, p =.01) and oral reading fluency (ORF) (R=0.54, OR=0.98, p < .0005) in 2nd grade significantly predicted whether students received 2 or more major ODRs in

5th grade. Finally, DIBELS Phoneme Segmentation Fluency scores (*R*=0.52, OR=0.97, *p* < .001) measured in kindergarten significantly predicted whether students would have more than 2 major ODRs in 5th grade, while number of ODRs in kindergarten did not significantly predict ODRs in 5th grade. The authors note that ODRs are not the gold-standard for screening for or measuring behavior in schools due to the inability for ODRs to capture all behavior in schools; but given that collecting and using ODRs was a criterion to evaluate the efficacy of the district SWPBS plan, the authors used ODRs as their behavior predictor (p. 279).

Trout, Nordness, Pierce, and Epstein (2003) conducted a review of 65 articles from 1961-2000 aimed at assessing the literature base for the current academic standing of students with emotional and behavioral disorders (E/BD). Sixteen studies reported on the academic functioning of students with E/BD. No study reported that these students were functioning at either age or grade level, and 91percent of studies reported that these students were actually functioning at least 1 grade level or year behind their peers. There were 84 ‘cases’ in which students with E/BD were compared to another group (typically developing, learning disabled, intellectually disabled, or attention deficit hyperactivity disorder). Compared to students with learning disabilities and ADHD, students with E/BD performed at the same level in reading, arithmetic and written expression. Compared to students with intellectual disabilities, students with E/BD functioned at a higher level in both written expression and arithmetic (2003, p.8). Reid, Gonzales, Nordness, Trout, and Epstein (2004) followed up on the previous study with 25 studies published between 1961 and 2000. The authors utilized studies that provided data for effect size calculation. These 25 articles included 2,486 students with E/BD, 82 percent of those students male, 69 percent Caucasian, 27 percent African American, 3 percent Hispanic, and 1 percent mixed ethnicities. The authors found a significant difference between students with E/BD and typically developing students in regards to academic achievement (ES= -0.69). Students with E/BD performed worse than students without E/BD in all subjects. Therefore, early identification of students at-risk for students may be beneficial in helping students with future behavior and academic problems.

Nelson, Benner, Lane, and Smith (2004) investigated the relationship among students with E/BD and their academic achievement. The authors utilized a cross-sectional design in a sample of 155 students aged K-12 in an urban school district in the Midwest. Data was collected regarding social adjustment using the Achenbach TRF; for academic achievement as measured by the Woodcock-Johnson Tests of Achievement – Third Edition (WJ-III); and regarding ethnicity, hours of special education per day, and IQ via record reviews. The authors found that nearly 83 percent of students classified as E/BD had achievement scores below the control group of non-disordered peers. No gender differences were found in regards to academic achievement, but older students scored lower on the math portion of the WJ-III. Using multiple regression methods, it was found that students who were rated high on externalizing problem subscales of the TRF had a significantly greater chance of having lower scores on the WJ-III in reading, written language, and math than students who were rated as only having internalizing behavior problems.

Fleming, Harachi, Cortes, Abbott, and Catalano (2004) investigated a model by which they reviewed the stability of reading scores and teacher-reported attention problems from elementary to middle school and evaluated if these scores/ratings in elementary school predicted problem behaviors when these students entered middle school. Their participants were 783 students enrolled in the Raising Healthy Children Project in the Pacific Northwest. Reading achievement data was collected via Northwest Evaluation Association: Achievement Level Tests (NWEA, 1997), and data regarding attention problems were collected via a teacher survey called the Teacher Observation of Classroom Adaptation-Revised (Werthamer-Larsoon, Kellam, & Ovesen-McGregor, 1990). Problem behavior was measured by a student survey form assessing substance use, covert antisocial behaviors, and physical aggression. Latent growth curve models were used to analyze the dataset. The authors found that reading ability and attention problems ratings were generally stable over time, as 62 percent of the variance in reading scores in grade 6 was explained by reading scores in grade 3; and 23 percent of the variance in attention problems in grade 6 was explained by attention problems in grade 3. The authors also found that attention problems predicted problem behaviors, but students with high scores for attention problems in grade 3 with decreasing scores as they advanced to grade 6 were less likely to exhibit problems in 7th grade, further highlighting the importance of screening and early intervention.

The previous studies demonstrate a relationship between externalizing problem behaviors and deficits in academic ability. A key point is that, if left un-treated, these externalizing problems, as well as their co-morbid academic deficiencies, do not disappear with age. The Fleming et. al. (2004) study provided promising data for intervening with students who exhibit externalizing behavior problems prior to completion of 6th grade. These studies suggest that intervening in areas such as attention and externalizing problem behaviors can have a positive impact on academic competence in addition to remediation of behavior problems in the classroom. Other skills that impact academic success but that are not themselves academic skills are considered to be academic enablers (Diperna & Elliott, 2002).

Academic Enablers

In an attempt to integrate and further explain the relationship between academics and behavior, Diperna and Elliott (2002) investigated a model of academic competence that included both academic skills and academic enablers. Academic skills included in the model were reading, mathematics, and critical thinking. Academic enablers, “attitudes and behaviors that allow a student to participate in, and ultimately benefit from academic instruction in the classroom,” were engagement, study skills, motivation, and interpersonal (social) skills. An evaluation of this model by Malecki (1998) and Malecki & Elliott (2002) found that increased social skills significantly predicted higher academic competence, which in turn significantly predicted academic achievement.

Volpe, DuPaul, and colleagues (2006) had parents and teachers of students with and without ADHD complete the ADHD Rating Scale –IV (ADHD-IV) and parents of these students complete the Diagnostic Interview Schedule for Children (Shaffer et. al., 1998) in order to determine how symptoms of ADHD affect academic achievement in reading and math, as measured by the WJ-III. The authors found that elevated ratings of academic enablers (motivation, study skills) mediated the effect between ADHD and reading and math achievement. This would mean that students with ADHD are not predisposed to low achievement, but students with ADHD often have deficits in academic enablers and are therefore more at risk to score lower than students without deficits in academic enablers. The academic enabler research is further evidence that proactive screening for both academic and nonacademic behaviors may prove invaluable in providing the most optimal early intervention program.

Using Behavior Screening Data to Predict Achievement

A limited number of studies have utilized evidence-based behavior screening instruments to predict academic achievement. Guzman, Jellinek, and colleagues (2011) utilized the Teacher Observation of Classroom Adaptation-Revised (TOCA-R) and Pediatric Symptom Checklist (PSC-CI) in order to determine whether mental health screening scores when Chilean students are in first grade can predict the same students’ achievement scores in fourth grade after accounting for individual and family risk factors. The authors found that, in their sample of over 7,000 students, after controlling for these factors, that students rated at-risk for mental health problems on one screener in 1st grade scored approximately 1/3 standard deviations lower on the national achievement tests in 4th grade than those students who were not rated at risk. If the students were screened at-risk on both screeners, they scored approximately 2/3 standard deviations lower than those not rated at-risk. Behavior ratings were found to be the 2nd strongest predictor, with teacher-ratings of academic competence on the TOCA-R when students were in 1st grade being the strongest predictor.

Two studies have used the behavior screening instruments relevant to the current study to attempt to predict achievement on a state or national assessment. An unpublished dissertation conducted by Emens (2009) investigated whether the Behavior Assessment System for Children–Teacher Rating Scale–Child Screener (BASC-TRS-C Screener; Kamphaus, 2009) could successfully predict whether students would pass or fail the reading or math sections of the Georgia Criterion Reference Competency Test (CRCT; GDOE, 2004). For the entire sample of 2nd-5th grade students (N=636), students who failed at least one portion of the CRCT had a significantly higher mean score on the BASC-TRS-C. Results of logistic regression analyses were that the BASC-TRS-C predicted with 90% accuracy whether a student would pass or fail the CRCT reading subtest. Significant predictors in the model were the screening score, being of African American ethnicity, and being of Hispanic ethnicity. While prior achievement scores were not utilized in the analyses, an ability measure, the Cognitive Ability Test, was used as a predictor and did not significantly predict results on the CRCT.

Kettler, Elliott, Davies, and Griffin (2009) used the PSG and the Social Skills Improvement System – Rating Scales (SSiS-RS; Gresham & Elliott, 2007) to predict Australian student achievement on a national achievement test. The authors found that the PSG and SSiS-RS both produced correlations among academic scores and prosocial behavior at around the same degree (r=.57) as the Caprara et al. (2000) study. Additionally, the authors found that the prosocial behavior score on the PSG had high sensitivity (.95), meaning it correctly identified students who scored below criteria on the achievement test, and high negative predictive value (.99), meaning that a high rate of students identified as not at risk by the PSG scored above criteria on the achievement test. The prosocial behavior score on the PSG had low scores in specificity (.44), meaning that a large amount of students who scored above criteria on the achievement test were rated as at-risk on the PSG, and positive predictive value (.18), identifying a large amount of students as at-risk on the PSG who scored above criteria on the achievement test. Finally, the PSG compiled a hit rate of .5 with a base rate of .11 (p.8). Another finding from this study was that scores on the SSiS-RS, which takes 12-15 minutes per student, did not add much to the variance explained in predicting achievement than the PSG, a quick screening instrument that takes approximately 25 minutes per classroom. The PSG may over-identify students as at-risk for underperforming on an achievement test; therefore, additional assessment may be needed before placing students into intervention groups.

Rationale and Research Questions

Given the movement of districts and schools, in addition to state and federal governments, to screen students for behavioral concerns, an increased knowledge of what this data can tell personnel would be useful. Scores on statewide assessments have implications along multiple levels; therefore, the ability for school personnel to both identify students who are more likely to perform poorly on these assessments, and thus, intervene with those students early and in as many areas as possible, is in high demand. The research base for identifying these students based on reading, math, and writing CBM scores is growing, but the relationship between behavior screening and results on statewide assessments is still not clear, despite the extensive documented relationship between behavior and academics. This study has two purposes: (a) to assess the predictive validity of behavior screening data from results on the statewide achievement tests in Louisiana; and (b) to extend our knowledge of the relationship between teacher/student ratings of behavior of students and scores on statewide assessments. This study was guided by the following research questions.

1. What is the relationship between behavior screening scores and outcomes in ELA and Mathematics on Louisiana statewide assessments?

2. Do behavior screening scores predict scores on statewide assessments?

3. Are the author-prescribed cut-scores for behavior screening useful in classifying whether a student passes or fails the statewide assessment to a better degree than chance?

4. Can behavior screening scores be combined with prior scores on statewide assessments and reading screening scores to lend a more accurate prediction of student outcomes on statewide assessments?

The following hypotheses were tested in investigating the aforementioned research questions.

H1: BESS Teacher, PSG Prosocial behavior, and PSG Motivation to Learn (fall and winter) will correlate significantly with each other and both *i*LEAP/LEAP ELA scores and *i*LEAP/LEAP math scores for the total sample and across different grade levels.

H2: BESS Teacher, PSG Prosocial behavior, and PSG Motivation to Learn (fall and winter, with winter accounting for a greater percent of the variance) will each significantly predict both *i*LEAP/LEAP ELA scores and *i*LEAP/LEAP math scores for the total sample and across grade levels.

H3: A model with BESS Teacher, PSG Prosocial behavior, and PSG Motivation to Learn will accurately classify students either passing or failing the criterion measures to a better degree than chance (i.e. postulating that each student will pass the assessment).

H4: The results will not differ across grade level.

METHOD

Participants and Setting

Four schools from a school district in Louisiana participated in the study. Data from 750 students in third through eighth grade were used for analysis. A power analysis was conducted using G*Power 3. Given an effect size of 0.15, alpha = 0.05, and a power of 0.80, it was determined a sample size of 85 participants was needed to conduct these analyses. Students' data were eligible for inclusion in the study if the student had behavior screening data from either the winter or fall screening and if their scores on the *i*LEAP and LEAP were available from the LEAP reporting system. Data from students who take a LEAP Alternative Assessment such as the LAA1 (students with severe cognitive disabilities) and LAA2 (students "with persistent academic difficulties") (Louisiana Department of Education, retrieved from: http://www.doe.state.la.us/topics/laa2.html) were not used for analysis, due to the test being significantly different from the non-alternative assessment. The final sample's demographic information is presented in Appendix B.

Measures

Behavior Assessment Scale for Children – II - Behavioral and Emotional Screening System. The *Behavioral and Emotional Screening System* (BESS, Kamphaus & Reynolds, 2007) is an instrument used to quickly and reliably assess the behavioral and emotional functioning of an entire school (grades pre-kindergarten through 12). The items for the measure came from the original item set that comprises the *Behavior Assessment Scale for Children – II* (BASC-2) teacher and student rating forms (Dowdy et. al, 2011). The teacher form consists of 27 items which, according to the authors, can be completed in 3-5 minutes. The student form consists of 30 items and requires students to be reading at a third grade level to complete. Raw scores are converted into T-scores (M=50, SD=10), with T-scores of 61-70 representing “elevated risk” and T-scores of 71 or higher representing “extremely elevated risk.” Psychometric data for the BESS are as follows. Split-half reliability estimates range from .96-.97 for the teacher form and .90-.93 for the student form. Test-retest reliability is .91 for the teacher form and .80 for the student form. Interrater reliability for the teacher form is .70. Both teacher and student BESS forms correlate with the Behavior Symptoms Index scores of the BASC-2 at *r*=.90.

Social Skills Improvement System - Performance Screening Guide. *The Social Skills Improvement System – Performance Screening Guide* (PSG, Elliott & Gresham, 2007) is a measure used to screen for students who “may be at risk for developing or having prosocial behavior or motivation to learn problems.” (2007). Teachers rate each student in their class on “Prosocial Behavior” and “Motivation to Learn”. The items are rated using a five point likert-scale, with scores of 4-5 signifying “no risk,” scores of 2-3 representing “elevated risk,” and a score of 1 representing “extremely elevated risk.” Test-retest reliability for the Prosocial Behavior and Motivation to Learn scales of the PSG range from *r*=.69 to *r*=.72 and *r*=.73 to *r*=.74, respectively, depending on the grade range of the student. Interrater reliability coefficients range from *r*=.37 to *r*=.55 for the Prosocial Behavior scale and *r*=.59 to *r*=.62 for the Motivation to Learn scale. The Prosocial Behavior scale correlates at *r*=.70 to the Social Skills scale and at *r*=-.58 to the Problem Behavior scales of the *Social Skills Improvement System: Rating Scales* (SSIS:RS; Gresham & Elliott, 2008). The Motivation to Learn scale of the PSG correlated at *r*=.58 to the Social Skills scale and at *r*=-.56 to the Problem Behavior scales on the SSIS:RS.

Louisiana Educational Assessment Program. The *Louisiana Educational Assessment Program* (LEAP) is a series of tests which determine whether fourth and eighth grade students are eligible to proceed to the next grade. These tests are criterion-referenced measures that determine the extent to which students have “mastered the state content standards” (Louisiana Department of Education, retrieved from: http://www.doe.state.la.us/testing/). Louisiana began administering these tests in 1997 (Mitzel & Borden, 2000). There are four sections: English Language Arts (ELA), Mathematics, Science, and Social Studies. Students may score, from lowest achievement to highest achievement, *Unsatisfactory, Approaching Basic, Basic, Mastery,* or *Advanced*. In order to proceed to the next grade level, students must score *Basic* in either ELA or math and *Approaching Basic* or above in the other content areas. The specific scores to delineate these achievement levels for each grade, based on the 2011 testing year, are located in Appendix F. The science and social studies standards are not included because those content areas are not used in determining whether a student passes or fails the test in high-stakes testing years (grades 4-8). On English Language Arts tests, there are four portions: writing, using information resources, reading and responding, and proofreading. On the math tests, six “strands” are assessed: number and number relations; algebra; measurement; geometry; data analysis, probability, and discrete math; and patterns, relations, and functions.

The 2010 technical manual for the LEAP test is the most recent available manual and can be accessed via the Louisiana Department of Education website (Louisiana Department of Education, retrieved from: http://www.louisianaschools.net/lde/uploads/18004.pdf). The report states that content validity was established by having in-state committees define the content that the test should cover, and then sending those content standards statewide for review. Then, the blueprint for the test was designed, based on the content standards set forth by the committee. Each item on the test was analyzed through field tests and by advisory committees to determine their content validity. Reliability estimates on the grades 4 and 8 tests as measured by Cronbach’s alpha range from 0.89-0.93.

ELA and math scores from the 2012 administration of the LEAP were used as the primary dependent variable in the analyses. ELA and math scores were also collected for each student from the 2011 administration. Using regression analyses, Noell and Burns (2006) found that prior year's testing scores correlated with the current year's testing score with r=.718 in ELA and r=.773 in math. Using these data in the analysis should aid in determining the unique variance contributed by the remaining predictor variables.

Integrated Louisiana Educational Assessment Program. The *Integrated Louisiana Educational Assessment Program* (*i*LEAP) is administered to students in grades 3, 5, 6, and 7. The test was designed using items from both the *Iowa Tests of Basic Skills* (ITBS; Hoover & Dunbar, 2007) and items crafted by specialists with training in test construction and design. The tests cover Louisiana's Grade-Level Expectations (GLEs) and content standards in ELA, Mathematics, Science, and Social Studies. The test is "integrated" because it functions as both a norm-referenced (ITBS) and criterion referenced assessment (items added to ensure all GLEs and content standards are covered by the test). Similar to the LEAP, students may score along a continuum of *Unsatisfactory, Approaching Basic, Basic, Mastery,* and *Advanced.* The range of scaled scores for each classification range can be found in Appendix B. The math strands assessed are: number and number relations; algebra; measurement; geometry; data analysis, probability, and discrete math; and patterns, relations, and functions.

The English Language Arts content standards are:

- Students read, comprehend, and respond to a range of materials, using a variety of strategies for different purposes;
- students write competently for a variety of purposes and audiences;

- students communicate using standard English grammar, usage, sentence structure, punctuation, capitalization, spelling, and handwriting;
- students locate, select, and synthesize information from a variety of texts, media, references, and technological sources to acquire and communicate knowledge; and
- students read, analyze, and respond to literature as a record of life experiences/

The reliability coefficients, as estimated using Cronbach's alpha, range from 0.82 to 0.93 depending on the test and grade level (Louisiana Department of Education, retrieved from: http://www.louisianaschools.net/lde/uploads/18005.pdf). 2011 and 2012 testing data were used in the analyses.

Curriculum-Based Measures. *Dynamic Indicators of Basic Early Literacy Skills* **(**DIBELS; University of Oregon, 2003) Oral Reading Fluency and Reading maze (Daze) screening instruments were utilized by the school district during the 2011-2012 school year. Oral reading fluency is measured as the number of words read correctly by the student in one minute. The same passages are administered to each student in their particular grade level. Students in grades 3, 4, and 5 were administered DIBELS Oral Reading Fluency during the fall, winter, and spring. Students in grades 6, 7, and 8 were administered DIBELS Reading Daze which is a multiple choice task whereby students read a passage silently and select the most appropriate word out of three to complete sentences within the story. After the first sentence, every seventh word in the passage is replaced with a choice between the correct word and two distractor words. Students have 3 minutes to complete the passage. Their score is comprised of the total number of words chosen correctly minus the total number of words chosen incorrectly.

Procedure

Measure Administration and Data Collection. As a part of the district's behavior RTI initiative, each teacher completed a BESS and PSG behavior screening on each student, in the fall (September, 2011) and winter (January-February, 2012). Teachers completed these measures within three weeks and school-assigned team leaders entered the data into a school-wide database. Teachers administered the student BESS to their homeroom class and followed a similar procedure regarding returning the data and data entry. The reading CBM probes were administered to each student in August, 2011 and January, 2012.

The LEAP was administered to students in two phases. In phase one, students were tested on March 20, 2012 on writing and math constructed response items. Phase two occurred on April 12, 13, 16, and 17, 2012. Phase two consisted of the bulk of the test, as students were tested on remaining items in the ELA and math sections as well as on the science and social studies sections. Students who took the *i*LEAP were tested during the same time period as phase two, and these students were tested in ELA, math, science and social studies. LEAP and *i*LEAP scores were matched to behavioral screening data by means of each student's state-issued identification number. Following this process, each student was assigned a unique identification number for purposes of analysis for this study; and their names and state-issued identification numbers were removed from the master data file.

Other archival data collected was the student's LEAP/*i*LEAP score from the previous school year via the LEAP data reporting system.

Standardization of Data. Scores across grade level on the LEAP and *i*LEAP, as well as scores across grade level on DIBELS oral reading fluency and DIBELS reading Daze, are not comparable (i.e. each grade level measure has a slightly different mean and standard deviation).

Therefore, in order to compare the LEAP and *i*LEAP scores accurately, the reported standard scores were converted into *z* scores based on the mean and standard deviation of each grade level's test (Noell & Burns, 2006). The *z*-scores were then converted to a standard score with a mean of 300 and a standard deviation of 50, which are the approximate mean and standard deviation of the LEAP and *i*LEAP. Finally, reading scores were converted to normal curve equivalents (NCEs) for more accurate comparison.

Analyses

After collecting the data from the district, it was discovered that there were cases with missing data. The total number of cases in this study was 750, but there were 281 cases with at least one predictor variable missing: 21.8 percent of fall student BESS, 11.7 percent of fall teacher BESS, 12.3 percent of fall PSG Motivation to Learn, 12.1 percent of fall PSG Prosocial Behavior, 7.6 percent of winter student BESS, 10.1 percent of winter teacher BESS, 8.5 percent of winter PSG Motivation to Learn, and 8.7 percent of winter PSG Prosocial Behavior scores were missing. Also, 6 percent of fall reading scores and 4.8 percent of winter reading scores were missing. All LEAP/iLEAP scores were present for each case. In order to determine the means by which to work with these cases, Little's Missing Completely at Random test was run to aid in determining the pattern of missing data, to see if the data were missing completely at random, missing at random, or not missing at random. In data that are missing at random, the "missingness" could depend on observed data, but not on unobserved data (Graham, 2009, p.552). Data that are missing completely at random are not dependent on observed or unobserved values in the dataset (Graham, 2009; Howell, 2009). Data that are not missing at random are dependent on unobserved data, and the absence of those data may cause interpretation of the data to be biased. Data that are missing completely at random may be eligible for listwise deletion (removing the entire case from the analyses should the case have any missing predictor/criterion variables), due to the analysis remaining unbiased and the variability in the data not being affected (Graham, 2009; Howell, 2009). The results of Little's Missing Completely at Random test revealed that the data were determined to not be missing completely at random (*p*<.001), and the separate variance t-tests indicated that the missingness can be predicted by variables other than the criterion test scores (Tabachnick & Fidell, 2007). Due to the data being inferred as missing at random, using listwise deletion would not be appropriate for analyzing and drawing conclusions from this dataset, due to the risk of biased results by potentially removing relevant sources of variation (Bennett, 2001). Also, Graham (2009) does not recommend listwise deletion when there are more than 5 percent of data missing in the sample.

Given the limitations of using listwise deletion in dealing with missing data, multiple imputation was used to estimate the missing data. The IBM SPSS Missing Values manual (IBM; 2011) reports that multiple imputation is the preferred method for handling data that are not missing completely at random. In multiple imputation, missing values for relevant variables are predicted using values from existing variables (Wayman, 2003). The specific method used is described in the next paragraph. These predictions are calculated multiple times, and consequently multiple datasets are produced. Rubin (1996) recommends that five imputations (creating five new datasets) be created as this is a sufficient number in most cases (Tabachinick & Fidell, 2007). Each imputed dataset should be used in the analyses. Statistical analyses of choice are performed on each dataset; and the results from each dataset are pooled, leaving one set of results for interpretation (Wayman, 2003). Because existing data (and its parameters) are used to predict the values of the imputed data, new, imputed data points in multiple imputation are not ‘guessed;’ and therefore, multiple imputation is more efficient than other methods of estimating missing data.

A fully conditional specification model using Monte Carlo methods based on sampling using Markov chains (MCMC) was run to create the imputed datasets using the Missing Values add-on in SPSS 20. MCMC is completed in four steps as described by Azur and colleagues: “(the program) 1) creates “place holders” by imputing the mean for every missing value in the dataset, 2) the ‘place holder’ mean imputations for one variable are set back to missing; 3) observed values from that one variable are regressed on the other variables in the imputation model; and 4) The missing values from step 2 are replaced with predictions from the regression model” (Azur, Stuart, Frangakis, & Leaf, 2011, p. 42). Multiple imputations were used on the entire dataset to calculate missing values for the behavior screening scores and reading screening scores.

The following analyses were run to answer the aforementioned research questions. In order to determine the linear relationship between behavior screening scores, Pearson product-moment correlations were calculated with the behavior screening scores (fall and winter ratings) and 2011 ELA and math scores. In order to determine the predictive relationship between the predictor variables (behavior screening) and criterion statewide testing scores, multiple regression analyses were utilized. The predictor variables were fall and winter behavior screening scores, and the criterion variables were LEAP/*i*LEAP ELA and math scores. With multiple regression analyses, the goal is to produce the linear combination of predictor variables that best correlate with the criterion variables (Field, 2005). Field states that when using multiple regression analyses, care should go toward selection of predictors entered into the regression equation, due to both the need for a theoretical basis for using predictors and the high level of inter-correlation among variables in social science research (p. 160). In order to tend to these concerns, hierarchical (blockwise) multiple regression analyses were employed using a stepwise method to determine order of predictors. With hierarchical multiple regression, predictors are entered into the equation as blocks, allowing for the first block of predictors to be analyzed before the second block is accounted for. This method is applicable for data collected in schools because data can be evaluated as it becomes available. In the current study, fall behavior scores were entered into the first block, and spring behavior scores were entered into the second block. Stepwise regression was employed to determine the order of entry of variables within each block into the model, as well as how many predictors entered the model. Using stepwise methodology, the program searches for the predictor variable that is most related to the criterion, followed by adding in the predictor with the largest relationship to the criterion after accounting for the initial predictor, and so forth. Stepwise models are accepted for model building (Field, 2005), but opponents to this method have criticized the means by which stepwise regression orders and selects variables for inclusion in that the technique may capitalize on chance (Flom & Cassell, 2007). While this study still capitalizes on chance to an extent, the ordering method of blocks of variables was chosen a priori based on the order in which data became available to decision makers in the schools. Additionally, in this study, variables other than the behavior scores such as prior achievement (Noell & Burns, 2006) and reading screening scores (Shaw & Shaw, 2002; Good, Simmons,& Kameenui, 2001; Buck and Torgesen, 2003) have been shown to be related to the criterion statewide tests. Although behavior was hypothesized to be a significant predictor even accounting for these other variables, the order of importance of the variables was not able to be determined a priori due to inconsistencies of past research, thus stepwise selection was used within blocks.

To determine the behavior screening score's ability to predict 'passing' or 'failing' the LEAP/iLEAP, logistic regression analyses were conducted. Logistic regression is a technique utilized to predict a binary outcome from a dataset of variables that could be one or a combination of continuous, discrete, dichotomous variables (Tabachnick & Fidell, 2007). The authors note that the goal of logistic regression is to accurately predict the probability of an individual case being in one category or the other (2007). In order to complete this analysis, each student's scores in ELA and math were analyzed to determine whether he/she would have passed/failed the test, regardless of grade. In order to pass the high-stakes LEAP tests, students must score *Basic* or above in either ELA or math and at least *Approaching Basic* or above in the other content area. Therefore, each student, regardless of grade level, had his score categorized based on whether he met criteria to pass the LEAP test. This dichotomous 'pass/fail' variable was the criterion for the logistic regression. The predictor variables were entered in blocks using the same technique as the multiple regression analyses. Finally, conditional probability or diagnostic efficiency models were run to assess the accuracy of the independent variables in predicting success or failure on the criterion tests. A score of 61 and above on the BESS is considered 'at risk,' and scores of 1-3 on the PSG are considered 'at risk'. These cut scores were used to determine the sensitivity, specificity, positive predictive power, and negative predictive power of the behavior screening scores.

The dataset was divided into four groups for analysis. The first set of analyses was run on the entire dataset, using only the behavior screening scores as predictors and test scores as criterion variables. This was done in attempt to answer the first three research questions. In the regression analyses, the fall administration of the behavior screening data was entered into block one; and the spring data were entered into block two. For the conditional probability/diagnostic effiency analyses, the author recommended cut scores to determine whether a child is ‘at-risk’ were used to assess the predictability of the measures.

The following three groups of data and their analyses included other known variables that have been previously demonstrated as being related to results on statewide tests. These analyses were run to determine whether behavior screening scores would be significant, independent predictors of results on statewide tests. The dataset was divided into three groups: third grade, fourth and fifth grades, and sixth through eighth grades. These divisions were made due to the available data for each grade level. At the time of this study, third grade students in Louisiana did not have prior statewide testing scores; but third grade students are included in the study due to the reality that third grade students are tested with a statewide assessment in Louisiana, and knowledge about whether screening data are related to outcomes on statewide assessments may highlight more areas in which to intervene with these students. The fourth-eighth grades’ analyses included statewide testing scores from the prior year in the analyses, due to the documented elevated relationship between prior and current years’ testing scores. The fourth and fifth grade data was analyzed separate from the sixth-eighth grade data due to the grade levels taking different reading screening measures. The fourth and fifth grades were administered oral reading fluency passages, while the sixth-eighth grade students were administered reading Daze passages. The same analyses that were run with the entire dataset were run with each subset of data. In the multiple regression and logistic regression analyses for the third grade students, fall reading and behavior screening scores were entered in block one; and winter reading and behavior screening scores were entered in block two. For the fourth-eighth grade students, the prior statewide testing score was entered into block one, fall reading and behavior screening scores were entered into block two; and winter reading and behavior screening scores were entered in block three.

RESULTS

Relationship Between Behavior Scores and Statewide Test Scores

Means and standard deviations of the behavior screening variables from the fall and winter and the means and standard deviations of the LEAP/*i*LEAP are reported in Table 1. Scores on the behavior screening measures did not differ greatly for the sample across screening periods, and the scores were within the average range compared to the standardization sample. Mean scores on the LEAP/*i*LEAP would fall in the *Basic* or *Approaching Basic* achievement level, depending on the student's grade level. *Basic/Approaching Basic* are the achievement levels in which a student must score to meet criteria for passing the LEAP/*i*LEAP. The exact achievement levels and their score ranges are located in Appendix B.

Table 1

Means and Standard Deviations for Behavior Screening and LEAP/iLEAP Scores by Measurement Period

	Measurement Period		
	Fall	Winter	Spring
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Behavior Screening Scores			
Student BESS	52.54 (10.1)	51.74 (10.4)	
Teacher BESS	52.31 (10.5)	52.53 (10.4)	
PSG - Motivation to Learn	3.61 (1.06)	3.66 (1.00)	
PSG - Prosocial Behavior	3.65 (1.03)	3.63 (1.02)	
*i*LEAP			
ELA Scaled Score			292.28 (49.9)
Math Scaled Score			300.35 (58.7)

Table 2 displays the Pearson correlations between the Fall/Spring behavior screening scores and scaled scores on the LEAP/*i*LEAP. Each behavior screening score is significantly correlated with the sample's LEAP/*i*LEAP score ($p < .01$). Scores on the BESS are negatively correlated, meaning that as scores on the BESS increase (ratings of behavior move toward 'at-risk'), scores on the LEAP/*i*LEAP decrease. Scores on the measures of the PSG and statewide tests are positively correlated. While statistically significant, scores on the student-rated screenings were less highly correlated with statewide assessment outcomes than the teacher-rated BESS and PSG. For scores on the ELA portion of the test, the fall behavior ratings were slightly more correlated then the winter ratings. The winter ratings were slightly more correlated with scores on the math portion of the test. The fall PSG-Motivation to Learn rating had the highest correlation with LEAP/*i*LEAP ELA scores, and the winter PSG-Motivation to Learn had the highest correlation with LEAP/*i*LEAP math scores. The 30-item teacher-BESS had higher correlations overall than the 1-item PSG Prosocial Behavior rating. Overall, the fall and winter behavior scores are closely correlated.

Table 2

Correlations between Behavior Screening and LEAP/iLEAP Scores

Behavior Screening Score	LEAP/*i*LEAP Scaled Score	
	ELA	Math
Fall Student BESS	-.18*	-.11*
Fall Teacher BESS	-.42*	-.38*
Fall PSG - Motivation to Learn	.44*	.37*
Fall PSG - Prosocial Behavior	.38*	.32*
Winter Student BESS	-.18*	-.14*
Winter Teacher BESS	-.40*	-.42*
Winter PSG - Motivation to Learn	.43*	.40*
Winter PSG - Prosocial Behavior	.37*	.34*
ELA Scaled Score	*	.73*
Math Scaled Score	.73*	*

* Correlation is significant at the .01 level

Tables 3 and 4 report the results of stepwise, forward hierarchical multiple regressions, which were conducted in order to further investigate the relationship between behavior screening scores and results on statewide assessments for the entire dataset. Fall PSG Motivation to Learn, fall teacher BESS, and winter PSG Motivation to Learn were entered into the regression model for the ELA scaled score. The change in R^2 for each of these variables entering the equation was significant. These behavior screening variables accounted for 24.3% of the variance for ELA scaled score, F (3, 746) = 88.89, p < .001. Each variable had significant β values: fall PSG

Motivation to Learn, t (749) = 3.17, $p < .01$; fall teacher BESS, t (749) = -3.77, $p < .001$; and winter PSG Motivation to Learn, t (749) = 4.38, $p < .001$. Winter teacher BESS was added to the aforementioned variables in the regression equation for math scaled score, and the variables accounted for almost 21% of the variance, F (4, 745) = 50.30 $p < .001$. After each variable was entered in the final model, winter PSG Motivation to Learn, t (749) = 2.52, $p < .051$, and winter teacher BESS, t (749) = -3.33, $p < .001$, contributed significant β values. PSG prosocial behavior and student ratings were not included in the equations for either ELA or math due to not contributing sufficient unique variance.

Table 3

Summary of Stepwise, Forward Multiple Regression Analyses of Behavior Screening Scores Related to ELA Scaled Score on the LEAP/iLEAP

	Independent Variable	*SE B*	β	$\underline{R}^2$	ΔR^2
Predictor	Model 1:				
	Fall PSG Motivation to Learn	1.663	0.442	0.195	
	Model 2:				
	Fall PSG Motivation to Learn	2.308	0.292	0.218	0.023
	Fall Teacher BESS	0.230	-0.216		
	Model 3:				
	Fall PSG Motivation to Learn	2.770	0.183	0.243	0.025
	Fall Teacher BESS	0.225	-0.180		
	Winter PSG Motivation to Learn	2.439	0.212		

Table 4

Summary of Stepwise, Forward Multiple Regression Analyses of Behavior Screening Scores Related to Math Scaled Score on the LEAP/iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	Fall Teacher BESS	0.194	-0.373	0.138	
	Model 2:				
	Fall Teacher BESS	0.264	-0.233	0.159	0.021
	Fall PSG Motivation to Learn	2.614	0.203		
	Model 3:				
	Fall Teacher BESS	0.278	-0.081	0.200	0.041
	Fall PSG Motivation to Learn	2.597	0.147		
	Winter Teacher BESS	0.257	-0.284		
	Model 4:				
	Fall Teacher BESS	0.292	-0.097	0.209	0.009
	Fall PSG Motivation to Learn	2.844	0.083		
	Winter PSG Motivation to Learn	3.030	0.158		
	Winter Teacher BESS	0.302	-0.198		

In order to predict the likelihood of students meeting criteria of passing or failing the LEAP/*i*LEAP test based on behavior screening scores, a forward, stepwise logistic regression analysis was conducted. Results are displayed in Table 5 and reported for the pooled imputation model. The beginning block for the entire sample (n=750) had a hit rate of 63.7%. The hit rate is the number of correct classifications divided by the sample size. In other words, for this sample, if one guessed that all students would pass the leap, the hit rate of 63.7% signifies that one would be correct 63.7% of the time if he guessed that all students passed the *i*LEAP/LEAP. Fall teacher BESS entered the equation first in block one, raising the correct classification rate to 69.3%. When fall PSG Motivation to Learn was added to block one, the correct classification rate of block one increased by 0.5%. Winter PSG Motivation to Learn entered into block two, and the correct classification of pass/fail increased to 70.7%, resulting in a total increase in classification accuracy of 7%. Fall teacher BESS and winter PSG Motivation to Learn reliably predicted passing or failing the LEAP, according to the Wald statistic. Odds ratios of 1.04 and .56 were calculated for fall teacher BESS and winter PSG Motivation to Learn, respectively. An odds ratio of 1.04 for fall teacher BESS signifies that for every 1 point increase on the BESS, a student is 4 percent more likely to not meet criteria on the LEAP/*i*LEAP, when accounting for other variables. Each unit decrease on the Motivation to Learn would increase the student's odds of not meeting criteria by 44 percent, when accounting for the other variables.

Table 5

Summary of Logistic Regression Analyses for Predicting Pass/Fail of the LEAP/iLEAP Tests from Behavior Screening Scores

Variable	β	SE	Wald	*p*	TN	FN	TP	FP	Hit Rate
BLOCK 1									
Step 1					93.6	52	426	178.4	69.3
Fall T BESS	0.072	0.009	73.628	0.000					
Step 2					99.8	54.4	423.6	172.2	69.8
Fall T BESS	0.05	0.011	21.522	0.000					
Fall PSG MTL	-0.325	0.108	9.491	0.003					
BLOCK 2									
Step 1					111.6	59	419	160.5	70.7
Fall T BESS	0.042	0.012	12.278	0.000					
Fall PSG MTL	-0.049	0.128	0.265	0.702					
Win PSG MTL	-0.573	0.121	26.2	0.000					

Following logistic regression analyses, conditional probability models were run to determine the diagnostic efficiency of each behavior screening variable. Cut scores recommended by the authors of the measures were used for the analyses. Scores of 61 and above on each BESS measure and scores of 3 and below on each PSG measure are considered 'at-risk,' therefore those cut scores were used in the analyses. The results of these analyses are displayed in table 6.

Table 6

Diagnostic Efficiency for Behavioral and Emotional Screening System and SSIS: Performance Screening Guide for All Grades

	Fall				Winter			
	Student BESS	Teacher BESS	PSG MTL	PSG PSB	Student BESS	Teacher BESS	PSG MTL	PSG PSB
Sensitivity	24%	35%	63%	53%	26%	32%	60%	58%
Specificity	82%	90%	67%	71%	83%	86%	72%	68%
Positive Predictive Power	43%	65%	52%	51%	47%	57%	55%	51%
Negative Predictive Power	65%	71%	76%	73%	66%	69%	76%	74%

For these results, sensitivity is the probability that a student who does not meet criteria to pass the LEAP/*i*LEAP will be identified as "at-risk" by the screening measure. Specificity is the probability that a student who does meet criteria to pass the LEAP/*i*LEAP, is not classified by the screening measure as "at-risk." Positive predictive power is the likelihood that a student who is rated "at-risk" by the screener did not meet criteria to pass the LEAP/*i*LEAP, and negative predictive power is the likelihood that a student who was not rated as "at-risk" by the screener did not meet criteria to pass the LEAP/*i*LEAP. Shapiro and colleagues (2006) used 60% as a cut-off to evaluate the usefulness of a screening measure for diagnostic purposes. The behavior screening measures exceeded this criterion for specificity and negative predictive power across screening periods. The teacher BESS was more effective than student BESS across each statistic in correctly identifying students. The teacher BESS had the highest scores of any measure in positive predictive value, meaning that the teacher BESS was the strongest if the question being asked is, “If a student scores at-risk on a behavior screener, what is the probability of that student not meeting criteria of passing the LEAP/*i*LEAP?” Sixty-five percent of students who were rated as “at-risk” on the fall teacher BESS did in fact fail to meet criteria on the test (positive predictive power), whereas 71% of students who scored in the not “at-risk” range on the teacher BESS met criteria to pass the test (negative predictive power). Scores on the PSG-Motivation to Learn have higher percentages in negative predictive power and sensitivity than the teacher BESS, while the teacher BESS has higher percentages in specificity and positive predictive power.

Predicting Scores on Statewide Assessments for 3rd Grade Students

Descriptive statistics for third grade students are presented in Table 7. The results for third grade students include a measure of reading screening, DIBELS oral reading fluency, which has been well documented to be related to outcomes on statewide assessments across the country (Shaw & Shaw, 2002; Good, Simmons,& Kameenui, 2001; Buck and Torgesen, 2003). As stated above, prior achievement scores were not included in the analyses for the third grade students, due to there being no prior scores available. Means and standard deviations of the predictor and outcome variables are reported in Table 7. Third grade students rated themselves in both screening periods at ½ standard deviations above the mean (BESS has a mean of 50 and standard deviation of 10). Teacher ratings were closer to the average score of the measure. Students’ reading scores were slightly below average. Students’ mean scores on the *i*LEAP in ELA and math fall within the *Basic* achievement level.

Table 7

Means and Standard Deviations for Behavior Screening, Reading Screening, and LEAP/iLEAP Scores by Measurement Period

	Measurement Period		
	Fall	Winter	Spring
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Behavior Screening Scores			
Student BESS	56.11 (9.2)	55.59(10.7)	
Teacher BESS	53.11(11.4)	52.27(11.1)	
PSG - Motivation to Learn	3.33(1.2)	3.56(1.1)	
PSG - Prosocial Behavior	3.52(1.0)	3.54(1.1)	
Reading ORF (NCE)	44.66 (18.2)	43.75 (19.1)	
*i*LEAP			
ELA Scaled Score			280.21 (58.09)
Math Scaled Score			293.25 (60.08)

Pearson correlations are reported in Table 8. Student ratings of behavior were not as highly correlated with testing scores and oral reading fluency. Each other behavior screening score was correlated at the 0.01 level with both ELA and math scaled scores and oral reading fluency. The correlation coefficients were higher for the third grade sample than for the dataset as a whole. For example, fall teacher BESS had a correlation of *r*=-.60 with *i*LEAP scaled score for third grade students, and fall teacher BESS correlated with LEAP/*i*LEAP scaled scores at *r*=.42 for the entire dataset.

Table 8

Correlations between Behavior Screening and LEAP/iLEAP Scores

Behavior Screening Score	*i*LEAP Scaled Score		Reading ORF	
	ELA	Math	Fall	Winter
Fall Student BESS	-.20*	-.12	-.19*	-.22*
Fall Teacher BESS	-.60**	-.61**	-.31**	-.42**
Fall PSG - Motivation to Learn	.59**	.53**	.47**	.51**
Fall PSG - Prosocial Behavior	.51**	.51**	.41**	.46**
Winter Student BESS	-.14	-.21**	-.10	-.13
Winter Teacher BESS	-.60**	-.64**	-.38**	-.45**
Winter PSG - Motivation to Learn	.56**	.60**	.42**	.47**
Winter PSG - Prosocial Behavior	.47**	.47**	.36**	.40**
Reading ORF Fall	.63**	.50**	*	.91**
Reading ORF Winter	.69**	.55**	.91**	*

* Correlation is significant at the .05 level
** Correlation is significant at the .01 level

Stepwise, forward multiple regressions were run for the third grade students' data. Similar to the previous regressions, fall measures were entered into block one, and winter measures were entered into block two. Each student's measure of oral reading fluency was included in each block with the reading screening measures. Two separate regressions were run with ELA scaled score as the criterion variable in the first and math scaled score as the criterion variable in the second. Results are displayed in Tables 9 and 10. For the ELA portion of the *i*LEAP, fall oral reading fluency, fall teacher BESS, and winter oral reading fluency entered the model. These three variables accounted for 58% of the variability for the *i*LEAP ELA scaled score, *F* (3,149) = 72.11, *p* < .001. In the final model, fall teacher BESS, *t* (151) = -6.44, *p* < .001, and winter ORF *t* (151) = 2.58, *p* < .01, contributed significant beta values. For the math scaled score on the *i*LEAP, fall teacher BESS entered the regression equation first, accounting for 36% of the variance. The total model, with fall ORF and winter teacher BESS entered in the model, accounting for 51% of the variance for *i*LEAP math scaled score, *F* (3,149) = 53.46, *p* < .001. Each variable in the final model had significant beta values: fall teacher BESS, *t* (151) = -3.77, *p* < .05; fall ORF, *t* (151) = 4.78, *p* < .01; and winter teacher BESS, *t* (151) = -3.35, *p* < .01.

Table 9

Summary of Stepwise, Forward Multiple Regression Analyses of Behavior Screening and Reading Oral Reading Fluency Related to ELA Scaled Scores on the iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	Fall ORF	0.204	0.631	0.394	
	Model 2:				
	Fall ORF	0.184	0.494	0.567	0.173
	Fall Teacher BESS	0.295	-0.440		
	Model 3:				
	Fall ORF	0.415	0.187	0.584	0.017
	Fall Teacher BESS	0.310	-0.387		
	Winter ORF	0.416	0.351		

Table 10

Summary of Stepwise, Forward Multiple Regression Analyses of Behavior Screening and Reading Oral Reading Fluency Related to Math Scaled Scores on the iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	Fall Teacher BESS	0.352	-0.609	0.366	
	Model 2:				
	Fall Teacher BESS	0.342	-0.500	0.472	0.106
	Fall ORF	0.209	0.346		
	Model 3:				
	Fall Teacher BESS	0.505	-0.270	0.509	0.037
	Fall ORF	0.203	0.297		
	Winter Teacher BESS	0.518	-0.319		

A forward, stepwise logistic regression analysis was conducted in order to determine the degree to which behavior screening scores, in addition to reading screening scores, can predict whether third graders would be considered as passing or failing on the *i*LEAP test. Similar to the previous logistic regression, scores were entered into a fall block and spring block, with reading screening scores added to the behavior scores in the two blocks. Results are displayed in Table 11. For the third grade students, if one would have guessed that each student would meet criteria to pass the *i*LEAP test, he would have been 69.9% accurate. Fall teacher BESS and fall reading ORF were reliable predictors according to the Wald criterion. Fall teacher BESS was entered into step one of the logistic regression, and the correct classification improved to 80.5%. When fall ORF was entered in step two, that percentage increased to 84.7%. The total improvement in classification accuracy increased 14.8%. Fall ORF had an odds ratio of .94, and fall teacher BESS had an odds ratio of 1.14. No variables from winter screening were entered into the regression.

Table 11

Summary of Logistic Regression Analyses for Predicting Pass/Fail of the iLEAP Test from Reading and Behavior Screening Scores

Variable	β	SE	Wald	*p*	TN	FN	TP	FP	Hit Rate
BLOCK 1									
Step 1					25	8.8	98.2	21	80.5
Fall T BESS	0.134	0.024	32.942	0.000					
Step 2					30.6	8	99	15.4	84.7
Fall ORF	-0.064	0.016	16.167	0.000					
Fall T BESS	0.131	0.027	24.732	0.000					

Results from diagnostic efficiency tests are presented in Table 12. The fall teacher BESS had scores in each efficiency measure above 60 percent. The teacher BESS across fall and winter had high scores (nearly 9/10 across fall and winter) in specificity, which is the probability that a student who met criteria to pass the *i*LEAP also was rated “not at-risk” by scores on the teacher BESS. Negative predictive value scores were also high for the teacher BESS, but PSG Motivation to Learn had the highest probability of a student being rated “not at-risk” and subsequently meeting criteria to pass the *i*LEAP test. These negative predictive values were at least 10% higher than the values observed for the total dataset. The teacher BESS also had good positive predictive power for the third grade students, with nearly a 7 out of 10 chance of identifying a student not meeting criteria on the *i*LEAP based on being rated as “at-risk” on the screener. Nearly each screening measure for the third grade performed better than the screening measures of the entire sample for each statistic.

Table 12

Diagnostic Efficiency for Behavioral and Emotional Screening System and SSIS: Performance Screening Guide for Third Grade Students

	Fall				Winter			
	Student BESS	Teacher BESS	PSG MTL	PSG PSB	Student BESS	Teacher BESS	PSG MTL	PSG PSB
Sensitivity	41%	61%	85%	72%	43%	50%	78%	70%
Specificity	70%	89%	64%	69%	69%	90%	73%	68%
Positive Predictive Power	37%	70%	50%	50%	38%	68%	55%	48%
Negative Predictive Power	74%	84%	91%	85%	74%	81%	89%	84%

Predicting Scores on Statewide Assessments for 4th-5th Grade Students

The next subset of student's data that was analyzed to determine the relationship between behavior screening and scores on statewide assessments in Louisiana was the fourth and fifth grade. These grades were separated from third grade due to the absence of prior achievement scores for third grade students. They were separated from grades 6-8 because their reading screening outcome measure was different. Fourth and fifth grade students were administered DIBELS oral reading fluency measures, and sixth-eighth grade students were given DIBELS Daze probes. Also, in many instances these groups of students are in different school buildings, so knowledge of the relationship between predictor and criterion variables parsed in this way has practical implications and may aid administrators in interpretation of the results. There are 284 fourth-fifth grade students in these analyses.

Descriptive statistics for fourth-fifth grade students are displayed in table 13. Scores on the behavior screening measures are within the average range compared to the norming sample. Scores on the reading measures are within the low average range, and scores on the LEAP/*i*LEAP are either within the *Approaching Basic* or *Basic* scoring range, depending on the grade the student was in when he took the test.

Table 13

Means and Standard Deviations for Behavior Screening, Reading Screening, and LEAP/iLEAP Scores by Measurement Period

	Measurement Period		
	Fall	Winter	Spring
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Behavior Screening Scores			
Student BESS	52.25 (10.2)	51.47 (10.4)	
Teacher BESS	51.75 (9.7)	52.5 (10.3)	
PSG - Motivation to Learn	3.8 (1.00)	3.74 (1.06)	
PSG - Prosocial Behavior	3.75 (1.04)	3.72 (1.08)	
Reading ORF (NCE)	46.06 (17.7)	46.04 (17.6)	
2011 LEAP/*i*LEAP			
ELA Scaled Score			302.43 (46.0)
Math Scaled Score			305.51 (61.7)
2012 LEAP/*i*LEAP			
ELA Scaled Score			302.58 (46.6)
Math Scaled Score			310.66 (62.2)

Pearson correlations for fourth-fifth grade students are presented in table 14. Significant correlations at the .01 level are observed for all variables except for fall and winter student BESS and 2012 LEAP/*i*LEAP math scaled score (*r*=-.11;-.09) and winter student BESS and fall reading ORF (*r*=-.11). Winter reading ORF had the highest correlation with the 2012 LEAP/*i*LEAP ELA scaled score (*r*=.63), and 2011 LEAP/*i*LEAP ELA scaled score had the highest correlation with 2012 LEAP/*i*LEAP math scaled score (*r*=.48.)

Table 14

Correlations between Behavior Screening, Reading Screening, and LEAP/iLEAP Scores

Behavior Screening Score	2012 LEAP/*i*LEAP Scaled Score		Reading ORF	
	ELA	Math	Fall	Winter
Fall Student BESS	-.23**	-.11	-.16**	-.17**
Fall Teacher BESS	-.50**	-.39**	-.31**	-.33**
Fall PSG - Motivation to Learn	.49**	.39**	.41**	.39**
Fall PSG - Prosocial Behavior	.45**	.33**	.34**	.35**
Winter Student BESS	-.17**	-.09	-.11	-.16**
Winter Teacher BESS	-.47**	-.41**	-.34**	-.37**
Winter PSG - Motivation to Learn	.46**	.36**	.41**	.42**
Winter PSG - Prosocial Behavior	.39**	.30**	.32**	.32**
2011 LEAP/*i*LEAP ELA Scaled Score	.58**	.48**	.47**	.46**
2011 LEAP/*i*LEAP Math Scaled Score	.48**	.44**	.23**	.23**
Reading ORF Fall	.62**	.43**	*	.93**
Reading ORF Winter	.63**	.43**	.93**	*

* Correlation is significant at the .05 level
** Correlation is significant at the .01 level

Forward, stepwise regressions were run for the fourth-fifth grade student's data. Similar to the previous analyses, variables were grouped into blocks by their availability. Therefore, ELA and math scaled scores from the 2011 testing year were entered into block one, fall screening variables were entered into block two, and winter screening variables were entered into block three. Regressions were run for both 2012 ELA and 2012 math scores. The results are presented in Tables 15 and 16. For ELA scores, testing scores from the year before accounted for 34% of the variance. Behavior screening added a 0.05 percent *R*-square change when added to the model. In the overall model, fall and winter oral reading fluency in addition to the fall teacher BESS accounted for nearly 55% of the variance, *F* (5,278) = 69.85, *p* < .001. Three variables, ELA Scaled, *t* (282) = 5.17, p < .001; fall teacher BESS, *t* (282) = -5.43, *p* < .01; and winter ORF, *t* (282) = 3.13, *p* < .001 had significant beta values in the final model. The final regression model for math included prior achievement scores, the fall period's reading fluency measure, and fall/winter teacher BESS, accounting for almost 35% of the variance, *F* (5,278) = 31.17, *p* < .001. In the final model, math Scaled, *t* (282) = 3.69, *p* < .001; fall ORF, *t* (282) = 4.41, *p* < .001; and winter teacher BESS, *t* (282) = -2.743, *p* < .01, had significant beta values.

Table 15

Summary of Stepwise, Forward Multiple Regression Analyses of Prior Achievement, Behavior Screening, Oral Reading Fluency Related to ELA Scaled Scores on the LEAP/iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	2011 ELA Scaled	0.049	0.576	0.330	
	Model 2:				
	2011 ELA Scaled	0.071	0.702	0.341	0.011
	2011 Math Scaled	0.053	-0.173		
	Model 3:				
	2011 ELA Scaled	0.071	0.425	0.482	0.141
	2011 Math Scaled	0.048	-0.071		
	Reading ORF Fall	0.130	0.434		
	Model 4:				
	2011 ELA Scaled	0.068	0.353	0.537	0.055
	2011 Math Scaled	0.046	-0.093		
	Reading ORF Fall	0.124	0.393		
	Fall Teacher BESS	0.220	-0.261		
	Model 5:				
	2011 ELA Scaled	0.068	0.342	0.549	0.012
	2011 Math Scaled	0.045	-0.085		
	Reading ORF Fall	0.282	0.116		
	Fall Teacher BESS	0.219	-0.239		
	Reading ORF Winter	0.283	0.309		

Table 16

Summary of Stepwise, Forward Multiple Regression Analyses of Prior Achievement, Behavior Screening, and Oral Reading Fluency Related to Math Scaled Scores on the LEAP/iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	2011 ELA Scaled	0.071	0.475	0.223	
	Model 2:				
	2011 ELA Scaled	1.030	0.332	0.238	0.015
	2011 Math Scaled	0.077	0.196		
	Model 3:				
	2011 ELA Scaled	0.110	0.139	0.304	0.066
	2011 Math Scaled	0.074	0.267		
	Reading ORF Fall	0.201	0.301		
	Model 4:				
	2011 ELA Scaled	0.109	0.085	0.333	0.029
	2011 Math Scaled	0.073	0.250		
	Reading ORF Fall	0.199	0.270		
	Fall Teacher BESS	0.348	-0.196		
	Model 5:				
	2011 ELA Scaled	0.108	0.073	0.348	0.015
	2011 Math Scaled	0.072	0.261		
	Reading ORF Fall	0.199	0.248		
	Fall Teacher BESS	0.459	-0.071		
	Winter Teacher BESS	0.424	-0.187		

A forward, stepwise logistic regression analysis was conducted in order to determine the predictor variables' ability to identify students as either passing or failing the LEAP/*i*LEAP. The predictor variables were entered into the same blocks as for the regression analyses. Results are displayed in Table 17. Approximately 64% percent of fourth and fifth grade students met criteria to pass the LEAP/*i*LEAP. These students' 2011 ELA scaled score entered the equation first, and improved the classification rate to 75%. The addition of fall PSG Motivation to Learn improved the correct classification rate to 77.3%. In the final block, fall oral reading fluency and winter

PSG motivation to learn were entered into the model, but the overall classification rate remained approximately the same, resulting in a final improvement of 12.8%, 2% less than the improvement measured in the third grade model. According to the Wald criterion, 2011 ELA, fall ORF, and winter PSG Motivation to Learn were reliable predictors. 2011 ELA had an odds ratio of .98, fall ORF .98, and winter PSG Motivation to Learn had an odds ratio of .66.

Table 17

Summary of Logistic Regression Analyses for Predicting Pass/Fail of the LEAP/iLEAP Test from Prior Achievement, Reading ORF, and Behavior Screening Scores

Variable	β	SE	Wald	*p*	TN	FN	TP	FP	Hit Rate
BLOCK 1									
Step 1					52	22	161	49	75.0
2011 ELA	-0.030	1.238	51.320	0.000					
BLOCK 2									
Step 1									
2011 ELA	-0.027	0.004	39.097	0.000	60.8	24.4	158.6	40.2	77.3
Fall PSG MTL	-0.564	0.159	12.601	0.000					
Step 2									
2011 ELA	-0.024	0.004	30.238	0.000	59.8	26	157	41.2	76.3
Fall ORF	-0.028	0.011	6.851	0.009					
Fall PSG MTL	-0.475	0.166	8.257	0.004					
BLOCK 3									
Step 1									
2011 ELA	-0.024	0.004	28.727	0.000	60	23.8	159.2	41	77.2
Fall ORF	-0.025	0.011	5.384	0.020					
Fall PSG MTL	-0.215	0.204	1.115	0.292					
Win PSG MTL	-0.424	0.196	4.691	0.031					

Table 18

Diagnostic Efficiency for Behavioral and Emotional Screening System and SSIS: Performance Screening Guide for Fourth and Fifth Grades

	Fall				Winter			
	Student BESS	Teacher BESS	PSG MTL	PSG PSB	Student BESS	Teacher BESS	PSG MTL	PSG PSB
Sensitivity	29%	31%	57%	50%	25%	34%	59%	54%
Specificity	82%	92%	76%	74%	85%	83%	75%	72%
Positive Predictive Power	47%	67%	57%	51%	48%	52%	57%	51%
Negative Predictive Power	68%	71%	76%	73%	67%	69%	77%	74%

Conditional probability/diagnostic efficiency results are presented in Table 18. The teacher BESS again had high scores in specificity (92 and 83 percent across fall and spring). It is interesting to note that all measures saw decreased scores in sensitivity and negative predictive power relative to the results for third grade students. The third grade screening scores for teacher BESS and PSG Motivation to learn were generally better overall than for the fourth-fifth grade students.

Predicting Scores on Statewide Assessments for 6th-8th Grade Students

The final subset of the data that was analyzed was that of the sixth-eighth grade students (n=313). Descriptive statistics are reported in Table 19. Students scored slightly above average on the BESS teacher ratings, and they rated themselves as closer to average. Students' reading scores were higher by nearly 12 normal curve equivalents in the fall screening period. Students' scores on the LEAP/*i*LEAP were in the *Basic* or *Approaching Basic* achievement level, depending on the student's grade level at the time of the test.

Table 19

Means and Standard Deviations for Behavior Screening, Reading Screening, and LEAP/iLEAP Scores by Measurement Period

	Measurement Period		
	Fall	Winter	Spring
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Behavior Screening Scores			
Student BESS	51.05 (9.7)	50.10 (9.6)	
Teacher BESS	52.43 (10.9)	52.68 (10.1)	
PSG - Motivation to Learn	3.57 (1.01)	3.63 (.888)	
PSG - Prosocial Behavior	3.61 (1.00)	3.59 (.903)	
Reading Daze (NCE)	53.67 (22.84)	41.64 (19.82)	
2011 LEAP/*i*LEAP			
ELA Scaled Score			284.80 (43.8)
Math Scaled Score			294.46 (62.7)
2012 LEAP/*i*LEAP			
ELA Scaled Score			288.83 (46.8)
Math Scaled Score			294.46 (53.4)

Table 20

Correlations between Behavior Screening, Reading Screening, and LEAP/iLEAP Scores

Behavior Screening Score	2012 LEAP/*i*LEAP Scaled Score		Reading Daze	
	ELA	Math	Fall	Winter
Fall Student BESS	-.09	-.10	-.07	-.11
Fall Teacher BESS	-.24**	-.22**	-.01	-.03
Fall PSG - Motivation to Learn	.25**	.22**	-.06	.05
Fall PSG - Prosocial Behavior	.21**	.17**	-.01	-.01
Winter Student BESS	-.18*	-.16*	-.10	-.10
Winter Teacher BESS	-.23**	-.31**	-.06	-.19**
Winter PSG - Motivation to Learn	.29**	.31**	.08	.17**
Winter PSG - Prosocial Behavior	.26**	.28**	.07	.20**
2011 LEAP/*i*LEAP ELA Scaled Score	.75**	.65**	.38**	.44**
2011 LEAP/*i*LEAP Math Scaled Score	.62**	.73**	.23**	.29**
Reading ORF Fall	.37**	.29**	*	.46**
Reading ORF Winter	.43**	.38**	.46**	*

* Correlation is significant at the .05 level
** Correlation is significant at the .01 level

Pearson correlations are reported in Table 20. Fall student BESS was not significantly correlated with both 2012 LEAP/*i*LEAP ELA and math scaled scores or reading Daze. Winter student BESS was correlated at 0.05 with ELA and math scaled scores. None of the fall behavior screening variables were correlated with fall or winter reading screening measures. Winter behavior screening measures were also not correlated with fall reading screening measures, but winter teacher BESS, winter PSG motivation to learn, and winter PSG prosocial behavior were significantly correlated with reading Daze's winter administration (p<.01). Fall and winter teacher BESS, PSG motivation to learn, and PSG prosocial behavior were significantly correlated with 2012 ELA and math scores. The highest correlations were found between the prior year's testing scores and the current year's testing scores

Forward, stepwise regressions were run to further investigate these relationships. Similar to the fourth-fifth grade model, data were entered into blocks in the order that they become available to school personnel. 2011 testing scores were entered into block one, fall screening scores were entered into block two, and winter screening scores were entered into block three. Regression results for ELA and math scores on the LEAP/*i*LEAP are reported in Tables 21 and 22, respectively.

Table 21

Summary of Stepwise, Forward Multiple Regression Analyses of Prior Achievement, Behavior Screening, and Reading Daze Related to ELA Scaled Scores on the LEAP/iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	2011 ELA Scaled	0.040	0.749	0.559	
	Model 2:				
	2011 ELA Scaled	0.053	0.611	0.580	0.021
	2011 Math Scaled	0.037	0.204		
	Model 3:				
	2011 ELA Scaled	0.056	0.565	0.589	0.009
	2011 Math Scaled	0.037	0.21		
	Reading Daze Fall	0.083	0.109		
	Model 4:				
	2011 ELA Scaled	0.057	0.533	0.595	0.006
	2011 Math Scaled	0.037	0.210		
	Reading Daze Fall	0.086	0.073		
	Reading Daze Winter	0.101	0.102		

Table 22

Summary of Stepwise, Forward Multiple Regression Analyses of Prior Achievement, Behavior Screening, and Reading Daze Related to Math Scaled Scores on the LEAP/iLEAP

	Independent Variable	*SE B*	β	R^2	ΔR^2
Predictor	Model 1:				
	2011 Math Scaled	0.033	0.725	0.524	
	Model 2:				
	2011 Math Scaled	0.043	0.531	0.568	0.044
	2011 ELA Scaled	0.062	0.287		
	Model 3:				
	2011 Math Scaled	0.043	0.535	0.578	0.010
	Reading Daze Winter	0.110	0.118		
	2011 ELA Scaled	0.065	0.239		
	Model 4:				
	2011 Math Scaled	0.042	0.532	0.584	0.006
	Reading Daze Winter	0.114	0.110		
	2011 ELA Scaled	0.067	0.216		
	Winter Teacher BESS	0.212	-0.095		

For ELA and math scores on the LEAP/*i*LEAP, the prior year's score on that particular section of the test entered into the regression first, accounting for 55.9 and 52.4 percent of the variance, respectively. For ELA scores, math scores from the previous year increased r-squared 2%, and the addition of reading screening scores increased r-squared another 1.5%, with 59.5% of the variance accounted for, F (4,308) = 115.73, $p < .001$. No behavior screening scores entered the equation for ELA. The variables 2011 ELA scaled, t (311) = 10.01, $p < .001$; 2011 math scaled, t (311) = 4.30, $p < .001$; and reading Daze winter, t (311) = 2,39, $p < .001$, had significant beta values in the equation. For math scores, adding the ELA scaled score to math scaled score added 4.4% to the variance accounted for. Adding a reading screening measure increased r-squared by 1%, and winter teacher BESS was the only behavior screening score to enter the model, adding 0.6% to the variance accounted for, *F* 4,308) = 110.45, *p* < .001. 2011 math scaled, *t* (311) = 10.71, *p* < .001; reading Daze winter, *t* (311) = 2.41, *p* < .01; 2011 ELA scaled, *t* (311) = 3.85, *p* < .001; and winter teacher BESS, *t* (311) = -2.48, *p* < .05, each had significant beta coefficients in the final model.

In order to determine the predictive ability of this model for scoring at criteria to pass or fail the test, a stepwise forward logistic regression was run. Variables were entered into the logistic regression in blocks in the same fashion that they were entered into the multiple regressions. Results of the logistic regression are reported in Table 23. Approximately 60% of students met criteria to pass the LEAP/*i*LEAP. Prior testing scores entered into block one of the logistic regression, and the addition of those scores saw the correct classification rate increase to 79.9%. In block two, fall reading Daze screening entered the equation, and the classification rate increased to 80.4%. Two of the imputation models entered winter teacher BESS into block three, while two other imputation models entered winter PSG motivation to learn into block three. The other imputation model did not enter any variables into block three. Both behavior screening variables were reliable predictors according to the Wald criterion (*p*<.05), while all other predictors were also reliable as well (*p*<.01). The model with winter teacher BESS increased the correct classification rate by 0.8%, while the model with winter PSG motivation to learn increased the correct classification rate by 1.4%. Odds ratios were .98 for 2011 ELA, .98 for 2011 math, .98 for fall reading Daze, 1.04 for winter teacher BESS, and .68 for winter PSG Motivation to Learn. Both total models increased the classification accuracy for these grade levels by more than 20% over the base rate.

Table 23

Summary of Logistic Regression Analyses for Predicting Pass/Fail of the LEAP/iLEAP Test from Prior Achievement, Reading ORF, and Behavior Screening Scores

Variable	β	SE	Wald	*p*	TN	FN	TP	FP	Hit Rate
BLOCK 1									
Step 1									
2011 ELA Scaled	-0.043	0.005	63.751	0.000	84	28	160	41	78.0
Step 2									
2011 ELA Scaled	-0.029	0.006	25.536	0.000	88	26	162	37	79.9
2011 Math Scaled	-0.020	0.004	23.432	0.000					
BLOCK 2									
Step 1									
2011 ELA Scaled	-0.026	0.006	18.373	0.000	87.2	23.4	165	37.8	80.4
2011 Math Scaled	-0.021	0.004	24.230	0.000					
Daze Fall	-0.020	0.007	8.075	0.006					
BLOCK 3a									
Step 1									
2011 ELA Scaled	-0.024	0.006	15.235	0.000	90.5	24.5	164	34.5	81.2
2011 Math Scaled	-0.022	0.004	24.425	0.000					
Daze Fall	-0.022	0.007	8.383	0.004					
Win T BESS	0.036	0.017	4.527	0.034					
BLOCK 3b									
Step 1									
2011 ELA Scaled	-0.025	0.006	16.311	0.000	91	23	165	34	81.8
2011 Math Scaled	-0.021	0.004	23.038	0.000					
Daze Fall	-0.022	0.007	8.659	0.004					
Win PSG MTL	-0.384	0.192	3.995	0.046					

Diagnostic efficiency statistics are presented in Table 24. The teacher BESS had high scores in sensitivity, but the measures collectively performed the worst for this grade range.

Table 24

Diagnostic Efficiency for Behavioral and Emotional Screening System and SSIS: Performance Screening Guide for All Grades

	Fall				Winter			
	Student BESS	Teacher BESS	PSG MTL	PSG PSB	Student BESS	Teacher BESS	PSG MTL	PSG PSB
Sensitivity	14%	28%	59%	50%	20%	26%	54%	72%
Specificity	88%	88%	61%	69%	90%	87%	68%	65%
Positive Predictive Power	45%	60%	50%	52%	57%	56%	53%	69%
Negative Predictive Power	61%	65%	69%	67%	63%	64%	69%	69%

DISCUSSION

The purpose of this study was to further explore the relationship between behavior and academic achievement by investigating the relationship between commercially available behavior screening scores and outcomes on statewide assessments. Screening scores and outcomes on statewide assessments were chosen as predictor and criterion variables because they are mandated to be administered yearly as part of Louisiana's state plan (Louisiana Department of Education, retrieved from http://www.louisianaschools.net/lde/uploads/16839.pdf & http://www.doe.state.la.us/topics/leap.html), and the data should be readily available to school and central office personnel. Four different behavior screening variables, which are part of a commercially available program used for screening, intervention, and progress monitoring, were used to investigate these questions (AIMSweb®, www.AIMSweb.com). Two variables were teacher and student ratings on a 27- or 30-item rating scale, using a 4-point likert type rating, on the BASC-II *Behavioral and Emotional Screening System* (BESS, Kamphaus & Reynolds, 2007). A score of 61 and above on the BESS signifies that a student is in the 'at-risk' range of exhibiting behavior problems. The other two variables were the Prosocial behavior and Motivation to Learn scales from the *Social Skills Improvement System: Performance Screening Guide* (PSG, Elliott & Gresham, 2007). These are both teacher-rated scales, with one item per student comprising the score on each scale (1-"extremely elevated risk", 5-"no risk"). The criterion variables were scores on the Louisiana statewide assessment, either the *Louisiana Educational Assessment Program* (LEAP) or the *Integrated Louisiana Educational Assessment Program* (*i*LEAP). Multiple analyses revealed that two of the four behavior screening variables may function as predictors of scores on statewide assessments, in addition to academic variables.

The study sought to answer four research questions. The first research question inquired about the relationship between behavior screening scores and outcomes on the ELA and

Mathematics sections of the LEAP/*i*LEAP tests. The second research question was closely related, asking if behavior screening scores can be used to predict outcomes on the statewide assessments. Multiple analyses were run in order to address the questions of whether behavior screening scores are related to outcomes on statewide assessments in Louisiana and if behavior screening scores can predict outcomes on these tests.

In order to answer the first two research questions, the first set of analyses focused on the relationship between behavioral screening scores (administered both in the fall and in the winter) and outcomes on statewide assessments, independent of other known variables that are related to scores on the tests, such as scores on the tests from the previous year and reading screening scores. It was hypothesized that the BESS teacher, PSG Prosocial Behavior, and PSG Motivation to Learn screening scores would correlate significantly with both ELA and math scaled scores on the LEAP/*i*LEAP tests, with winter scores screening scores being more highly correlated than fall screening scores. This hypothesis was partially supported, with the fall and winter correlations not differing significantly from each other. The PSG Motivation to Learn screening scores had higher correlations with scores on LEAP/*i*LEAP ELA, and winter teacher BESS having slightly higher correlations for math scaled score. The PSG Prosocial Behavior had the third highest. Two findings, the lack of difference between fall/winter screening scores, and large difference between teacher/student ratings are discussed in the ensuing paragraphs.

Generally, fall screening scores were slightly higher correlated with testing outcomes than winter screenings. Each screening variable was slightly higher correlated to ELA scaled scores than to math scaled scores, with the exception of winter teacher BESS. Prior research has shown that behavior problems in school-aged children are relatively stable (Fleming et. al., 2004; Hayling, Cook, Gresham, Slate, & Kern, 2007), and this district did not implement systematic interventions following the fall behavior screening, so it may have been unreasonable to expect winter screenings to be more accurate in predicting test scores. Another theory that may help explain these findings is Rosenthal's Pygmalion Effect (Rosenthal, 2002; Rosenthal & Jacobson, 1968). In the Pygmalion effect study, students who were expected by the teacher to experience large growth intellectually made significantly higher growth than the remainder of the students on a group-administered test of intellectual ability. Teachers, after 6-8 weeks of school, may identify their 'red-zone students,' for whom they have reduced expectations and subsequently view and teach differently than other students. These initial views and expectations may persist throughout the school year and subsequently have effects on academic outcomes.

Student screening scores, while significant at the .01 level, were markedly less correlated with outcomes on assessments than the teacher-rated items, with a difference of nearly r=0.25. The teacher/student BESS correlated at a low to moderate degree (r=.26). Gresham and colleagues (2010) found that teachers and students perceive the degree of a student's problem behavior differently, finding that teacher and student ratings of social skills on the *Social Skills Improvement System: Rating Scales* (SSIS;RS Gresham & Elliott, 2008) correlate at a low to moderate degree (r=.21). Malecki and Elliott (2002) found similar results, as student ratings of behavior on the SSRS were not predictive of academic outcomes, while teacher ratings were predictive. Thus, it should not be surprising that teacher and student ratings of student behavior on the BESS would differ; and teacher ratings would be more predictive of scores on the tests.

Following calculation of correlation coefficients, multiple regression analyses were conducted to determine the best set of behavior screening predictors of scores on the ELA and math sections of the LEAP/*i*LEAP tests. It was hypothesized that BESS teacher, PSG prosocial behavior, and PSG motivation to learn would be the best predictors in a regression; and winter screening variables would be better predictors of test scores than fall screening variables. This hypothesis was partially supported, due to that behavior screening scores were significant predictors of test scores, but winter scores did not account for more of the variance. Three variables (fall PSG Motivation to Learn, fall teacher BESS, and winter PSG Motivation to Learn) were entered into the last model for ELA score. Winter teacher BESS joined fall PSG Motivation to Learn and fall teacher BESS, and winter PSG Motivation to Learn as significant contributors to the math scaled score. Logistic regression analyses were used to determine the predictive validity of behavior screening variables for identifying students as passing/failing the LEAP/*i*LEAP. Similar to the model for ELA, in the multiple regression, fall teacher BESS, fall PSG Motivation to Learn, and winter PSG Motivation to Learn entered the model. Again, the fall variables accounted for more of the variance in the model. As stated earlier, based on student behavior, teachers may be able to identify those who will struggle academically early in the school year. Or possibly, teachers give differential attention to those students who are considered well-behaved at the outset of the school year. Also, the fact that both screenings were significantly related may signify the importance of completing multiple screenings of behavior as well as academics, so that the screening casts a wide enough net to catch students whose problems may have only been emerging at the beginning of the school year.

To further investigate the usefulness of the behavior screeners to identify students who are at-risk of failing to meet standards on statewide assessments, diagnostic efficiency or conditional probability statistics were run for the screening variables using the cut scores recommended by the authors for identifying students as “at-risk.” The fall teacher BESS had the highest scores of all of the measures in specificity (90%). This score in specificity signifies that 9/10 students who met criteria to pass the LEAP/*i*LEAP also scored in the ‘not at-risk’ range on the teacher BESS. The teacher BESS had low scores in sensitivity (35%), meaning that only 3.5/10 students who did not meet criteria to pass the tests were rated in the ‘at-risk’ range on the measure. The fall teacher BESS had acceptable scores in positive (65%) and negative (71%) predictive power. This screening variable on its own classified nearly 7/10 students correctly, significant for a nonacademic screening variable. The PSG Motivation to Learn had higher sensitivity scores (57-58%), but those scores are below the acceptable criteria of 60% (Shapiro et. al., 2006). The PSG Motivation to Learn and Prosocial Behavior did not have the elevated scores in sensitivity (95%) and negative predictive value (99%) or as low of scores in positive predictive value (18%) and specificity (44%) that Elliott and colleagues reported in their study (2009).

Each of the behavior screening measures performed better on the specificity and negative predictive powers, while sensitivity and positive predictive power (save fall teacher BESS – 65%) were below 60%. For a comparison, studies examining ORF as a predictor of scores on reading portions of statewide tests found negative predictive values around 90% and positive predictive values around 75% (Good et. al, 2001; Buck & Torgeson, 2003; Stage & Jacobson, 2001) While the teacher BESS and PSG Motivation to Learn show promise in classifying students, the results from this study show that these measures should not be used on their own to determine whether students are at-risk of failing to perform well on statewide assessments. Subsequent analyses investigated the relationship of other variables to statewide testing scores for this sample.

In summary, scores from the fall administration of the teacher BESS and PSG Motivation to Learn accounted for most of the variance in the relationship between the fall and winter behavior screening variables and scores on statewide assessments; but both fall and winter administrations were closely correlated and each related to results on the LEAP/*i*LEAP. The following analyses sought to determine whether these screening variables remained significant predictors after including academic variables previously identified as predictors of statewide testing scores.

The final research question asked if behavior screening scores could be combined with other variables already demonstrated to be significantly related to results on statewide assessments to lend a more accurate prediction of outcomes on the tests. In other words, are behavior screening scores significant, independent predictors of test results when accounting for and including other known significant predictors of test scores? To aid in the interpretation of the analyses, results from the analyses are partially summarized in Table 25. This table presents the division of data, analysis run, scores, and significant predictors in each equation. Also, correlations across grade level for the two best predictors (teacher BESS and PSG Motivation to Learn) are presented in Table 26.

It was hypothesized that behavior screening scores would be significant, independent predictors of test scores. Behavior screening data entered the equation for all analyses except for predicting ELA scores in sixth-eighth grade students; therefore, this hypothesis was partially supported. It was also hypothesized that results would not differ across grade levels, and that hypothesis was not supported.

Table 25

Summary of Study Results

	Adj R-squared		
Subset of Data	ELA	Math	Classification Accuracy
Total Sample	.243 (a)	.209 (b)	70.7% (c)
Grade 3	.584 (d)	.509 (e)	84.7% (f)
Grades 4-5	.549 (g)	.348 (h)	77.2% (i)
Grades 6-8	.595 (j)	.584 (k)	81.8% (l)

a (Fall PSG MTL, Fall T BESS, Win PSG MTL)

b (Win PSG MTL, Win T BESS)

c (Fall T BESS, Win PSG MTL)

d (Fall T BESS, Win ORF)

e (Fall T BESS, Fall ORF, Win T BESS)

f (Fall ORF, Fall T BESS)

g (2011 ELA, Fall T BESS, Win ORF)

h (2011 Math, Fall ORF, Win T BESS)

i (2011 ELA, Fall ORF, Win PSG MTL)

j (2011 ELA, 2011 Math, Daze Win)

k (2011 Math, Daze Win, 2011 ELA, Win T BESS)

l (2011 ELA, 2011 Math, Daze Fall, Win T BESS (a), Win PSG MTL (b))

Analyses for the grade 3 data included fall and winter measures of DIBELS ORF added to the regression equations. For ELA scaled score, winter ORF and fall teacher BESS entered the equation s as significant predictors. For grade 3 math scaled score, fall teacher BESS was the first variable to enter the equation; and fall ORF entered second. A nonacademic skill, teacher rating of behavior, was a better predictor of outcomes on *i*LEAP math tests than screening scores for oral reading fluency. The fall measure of ORF and teacher BESS were significant predictors of passing the *i*LEAP test, with a hit rate of 84.7%. It is hypothesized that a math screener would increase this percentage (Menessess, 2011), but math data was not available from the school district. For these third grade students, behavior screening adds to the model of predicting students who struggle on the statewide assessments, even with an academic variable included in the model.

In Louisiana at the time of this study, the third grade did not have prior achievement scores to use in determining how students may perform on the current year's statewide assessment. Fall teacher BESS entered into each equation in the analyses, and it actually entered in the equation before reading screening for math scaled scores. It is hypothesized that if a math screening score was entered, a similar regression equation to the reading regression equation would occur (math entering first then teacher BESS; due to the teacher BESS having higher correlations with math scaled score than oral reading fluency). The screening data for the third grade students had the highest correlations with scores on the statewide tests. The behavior screening scores for third grade also performed the best according to conditional probability/diagnostic efficiency calculations. Those data for the fall administration of the teacher BESS and PSG Motivation to Learn can be found in Table 27. For third grade students, these data show that accounting for and intervening with behavior may be useful in maximizing an RTI model for struggling students.

Table 26

Correlations between Teacher BESS/PSG Motivation to Learn and Results on the LEAP/iLEAP Tests across Grade Levels

	Total Sample		Grade 3		Grades 4-5		Grades 6-8	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
Fall T BESS	-.42	-.38	-.60	-.61	-.50	-.39	-.24	-.22
Win T BESS	-.40	-.42	-.60	-.64	-.47	-.41	-.23	-.31
Fall PSG MTL	.44	.37	.59	.53	.49	.39	.25	.22
Win PSG MTL	.43	.40	.56	.60	.36	.30	.29	.31

Screening scores in grades 4-5 were less highly correlated with testing scores than scores for the third grade students. Grades four and five had the previous year's testing results in the first block of the regression equations. 2011 ELA and 2011 math scaled scores were in the final model as significant predictors of their respective 2012 testing score. Fall teacher BESS and winter reading screening entered the final equation for ELA, and winter teacher BESS and fall reading screening entered the final equation as significant predictors for math. 2011 ELA scaled score, fall reading screening, and winter PSG Motivation to Learn were significant predictors in the logistic regression equation. The overall variance explained for math test scores was much lower than for grades 3 and 6-8, and the 2011 ELA scaled score surprisingly had a higher correlation with 2012 math scaled score than the 2011 math scaled score. Finally, the grade 4-5 screening scores did not perform as well as the third grade scores in terms of sensitivity and negative predictive power.

Table 27

Summary of Conditional Probability/Diagnostic Efficiency Calculations for Fall Teacher BESS and Fall PSG Motivation to Learn

	Total Sample		Grade 3		Grades 4-5		Grades 6-8	
	Teacher BESS	PSG MTL	Teacher BESS	PSG MTL	Teacher BESS	PSG MTL	Teacher BESS	PSG MTL
Sensitivity	35%	63%	61%	85%	31%	57%	28%	59%
Specificity	90%	67%	89%	64%	92%	76%	88%	61%
Positive Predictive Power	65%	52%	70%	50%	67%	57%	60%	50%
Negative Predictive Power	71%	76%	84%	91%	71%	76%	65%	69%

Grades 6-8 had DIBELS reading Daze as the reading screening variable instead of DIBELS ORF. The regression model for ELA scaled score did not contain a behavior screening measure, and the model for math included winter teacher BESS as the fourth variable in the equation. This is likely due to the correlation coefficients for this grade range being much lower than the correlations for grades 3 and 4-5. The logistic regression analysis included winter PSG Motivation to Learn as the final variable included in the model. Nearly 60% of the variance was accounted for in both ELA and math scaled scores, the highest among the four groups of data, despite not having much contribution from behavior screening data. The grades 6-8 data also did not perform as well as the third grade data in terms of sensitivity, positive predictive power, and negative predictive power; but only the sensitivity score was under the Shapiro et. al. (2006) recommendation of 60%.

Overall, each teacher-rated behavior screening measure was significantly related to outcomes on statewide assessments in Louisiana for the data in this study. Student ratings of their own behavior did not enter the regression equations in predicting scores. PSG Prosocial Behavior did not enter any of the regression equations either. The BESS may be higher correlated than the PSG Prosocial Behavior in part because the BESS contains items related to the description of the PSG Prosocial Behavior, and the BESS also includes classroom behaviors (e.g. breaks the rules, has trouble keeping up in class). The PSG Motivation to Learn screening instruments entered regression equations throughout the analyses. The teacher BESS and PSG Motivation to Learn were consistently one of top behavior screening predictors. There were no consistent patterns as to whether fall or winter screenings would enter the models for predicting a specific grade level’s score on a particular test, likely because the fall and winter administrations were had similar correlations to the test scores.

After accounting for prior achievement scores and reading screening scores in fourth-fifth and sixth-eighth grade, behavior screening contributed a smaller amount to the total variance accounted for when predicting ELA and math scores on the LEAP/*i*LEAP. Behavior scores did not enter the equation for sixth-eighth ELA, and it may not have entered the equation for sixth-eighth math if a math screening score were available. If math screening scores were available, behavior screening scores may not have entered the logistic regression equations at these grade levels as well. While behavior screening scores may not be as related to academic outcomes as their respective subject area's screening instrument, the fact that behavior screeners are significantly related to academic outcomes give school personnel more reason to systematically screen for behavior problems and subsequently intervene. Based on these data, behavior screening is a better predictor of outcomes on statewide tests in Louisiana for younger grades.

Implications

Based on these findings, two of the four measures on the AIMSweb® Behavior module (teacher BESS & PSG Motivation to Learn) show promise as being helpful tools to administrators and teachers. Not only are these measures reliable and valid for identifying students who are at risk for exhibiting behavior problems, but they are also significantly related to outcomes on statewide tests even when accounting for other academic variables.

Schools implementing School-wide Positive Behavior Intervention Support programs (PBIS) as part of their RTI initiative often rely on number of office discipline referrals (ODRs) to identify children who are at-risk for developing behavior problems (n>2; Sugai, Sprague, Horner, & Walker, 2000). Walker and colleagues (2005) note that problems with using ODRs as the outcome measure for decision making in a PBIS model are that research is unclear regarding the relationship between a high number of ODRs and future behavior problems and ODRs typically overlook internalizing behavior problems. Thus, behavior screening using reliable and valid measures is the preferred, or recommended, practice in schools to measure behavior change as ODRs do not provide enough information in identifying students as at-risk for developing a behavior problem nor do they lend information for intervention planning or progress monitoring (Walker, Cheney, Stage, & Blum, 2005). Additionally, ODRs are not technically sound, as definitions of problems and tolerance of behaviors may vary among staff members and across school buildings (Tidwell, Flannery, & Lewis-Palmer, 2003). ODRs are useful when combined with other measures in evaluating the effectiveness of school-wide programs, but ODRs are not as efficient, accurate, and sensitive to changes in behavior as reliable and valid screening instruments. Results from this study should urge school districts to adopt reliable and valid behavior screening instruments instead of relying on office discipline referrals.

Informal communication with administrators and teachers regarding the relationship between behavior and academic success almost always includes the administrator or teacher saying, “Well, obviously behavior and academics are related.” Research has supported this statement (Caprara et al., 2000; DiPerna & Elliott, 2002; Malecki & Elliott, 2002; Wentzel, 1993). In spite of this, most programs in schools are aimed at targeting only academics, leaving behavior and classroom management as almost an afterthought. There is legislation aimed at changing this way of thought, as part of a teacher’s now yearly evaluation will continue to be based on observation of the classroom environment, which partly consists of managing student behavior and managing classroom procedures (Louisiana Department of Education, retrieved from http://www.louisianaschools.net/compass/about_compass.html). Teacher’s scores on their evaluation will improve if students are behaving and following routines appropriately. The current study adds to the research by providing further support for the use of behavior screeners to proactively intervene with potential behavior problems by documenting the relationship between behavior and academics. Since school employees are focused on improving academic outcomes for their students, having evidence that behavior screening is important (demonstrating that students who are not at-risk for behavior problems score well on the tests that schools themselves are graded on) may lead to increased efforts in setting and enforcing behavioral expectations and standards in their buildings via systematic direct instruction and a tiered PBIS program implemented with fidelity. In other words, the study adds more evidence showing that dedicating time and resources to teach students how to behave, similar to how we teach students to read and write, may have collateral effects on scores on the tests on which teachers are evaluated and also on the overall academic success of students.

Based on these results, these screeners should not be used in isolation to identify students at-risk for struggling on the statewide test; but there is evidence to show that the teacher BESS is 60-70% accurate on its own in identifying students who are going to meet criteria on the statewide test. If a student is rated 3 or below on the motivation to learn scale of the PSG, there is approximately a 7 in 10 or greater chance he/she will not meet criteria on the test. While students scoring at risk on these screeners may not need to be placed in academic intervention groups, these students are at a greater risk for developing academic deficits and can be flagged for closer monitoring. Once a student is identified is ‘at-risk’ for behavior or motivation problems, it would seem that teachers would be motivated to remediate these problems as quickly as possible, as their jobs are now hinging both on their own classroom environment and scores on the tests. Using these two scales in collusion with other academic screeners may lend more information toward a certain “recipe” for students who are identified as at-risk by other screening tools and previous performance.

Limitations and Future Directions

This study has some limitations that should be considered when interpreting to results and that encourage further research on this topic. First, the study used data from four schools within one district in Louisiana. This district represents less than 1% of students in Louisiana. The district also qualifies more students for free/reduced lunch and has less highly qualified teachers per student than the state average. The behavior and reading screening data, as well as statewide testing scores, may not represent the overall student population in Louisiana.

Another limitation of the study lies inherently in the measurement method that behavior screening employs. Behavior screening depends on ratings of behavior from either the teacher or the student, while other curriculum-based measures are more direct (i.e. measuring the number of words correct that come out of a student's mouth). Ideally, a comprehensive direct assessment of all students would give the best overall picture of a student; but it would be nearly impossible to directly observe each student in the classroom, then determine whether that observation was truly representative of a student's functioning. For example, Hintze and Matthews (2004) determined that in order to obtain a reliability of .80 of on-task behavior for a student, four observations per day across 40 days would be needed. The time and resources required to obtain reliable observations for every student in a school would likely outweigh the benefits. Elliott, Busse, and Gresham (1993) affirm that behavior ratings scales are useful for identifying students with behavior concerns, but the authors recommend that the rating scale be practically useful, reliable, and valid; criteria that the screening measures utilized in this study fulfill.

An additional limitation of this study was that some behavior screening and reading data was missing. Multiple imputation is more accepted than listwise deletion if the data is not missing completely at random due to the loss of the contribution to the variability when cases are deleted; but multiple imputation may result in a decrease in power, diminishing some of the effects of the imputed variables (Graham, Olchowski, & Gilreath, 2007). Five imputations were run on this dataset as Rubin (1992) prescribed, but it is unclear if using this many imputations may have resulted in reduced loss in power, if any power was in fact lost (2007). If the effects of the variables were reduced, the relationships reported in the study may not be as strong as the true relationship.

States are not going to move away from using annual testing to measure student growth and teacher performance, so knowing what variables add to predicting how a student will do as early as possible will aid in giving teachers and staff as many areas as possible in which to intervene should there be a deficit. Future research should focus on asking the same research questions with a complete dataset, which would aid in determining if there was any loss in variability across grade levels. Also, math screening scores should be included in order to determine whether behavior screening scores remain significant predictors following their entry.

Other mediational research could be conducted in this area. One potential study could use a smaller sample size and look at direct measurement of specific behaviors to determine if a small subset of observable behaviors is mediating the relationship between the screening scores and outcomes on tests. Also, the new Compass observation and teacher effectiveness data could be used to examine whether effective teachers are mediating the relationship between behavior and achievement (Louisiana Department of Education, retrieved from http://www.louisianaschools.net/compass/about_compass.html). Finally, the results could be analyzed by socioeconomic status and gender.

REFERENCES

Achenbach, T.M. & Edelbrock, C.S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T.M. *Manual for the Child Behavior Checklist 2/3 and 1992 Profile*. Burlington, Vt.: University of Vermont, Dept. of Psychiatry;1992.

Agostin, T.M. & Bain, S.K. (1997). Predicting early school success with developmental and social skills screeners. *Psychology in the Schools, 34*(3), 219-228.

Azur, M.J., Stuart, E.A., Frangakis, C., & Leap, P.J. (2011). Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res, 20*(1), 40-49.

Bandura, A. (1972). *Self-efficacy: The exercise of control.* New York: Freeman.

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Benner, G.J., Nelson, J.R., Allor, J.H., Mooney, P., & Dai, T. (2008). Academic processing speed mediates the influence of both externalizing behavior and language skills on the academic skills of students with emotional disturbance. *Journal of Behavioral Education, 17*(1), 63-78.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*, 464 – 469

Buck, J., & Torgeson, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report 1). Tallahassee: Florida Center for Reading Research.

Bursuck, W.D. & Asher, S.R. (1986). The relationship between social competence and achievement in elementary school children. *Journal of Clinical Child Psychology, 15,*(1), 41-49.

Caldarella, P., Young, E. L., Richardson, M. J., Young, B. J, & Young, K. R. (2008). Validation of the systematic screening for behavior disorders in middle and junior high school. *Journal of Emotional and Behavioral Disorders, 16*(2), 105–117.

Caprara, G.V., Barbaranelli, C., Pastorelli, C., Bandura, A., & Zimbardo, P.G. (2000). Prosocial foundations of children’s academic achievement. *Psychological Science, 11,* 302-306.

Chen, X., Huang, X., Chang, L., Wang, L., & Li, D. (2010). Aggression, social competence, and academic achievement in Chinese children: A 5-year longitudinal study. *Development and Psychopathology, 22,* 583-592.

Cobb, J.A. (1972). Relationship of discrete classroom behaviors to fourth-grade academic achievement. *Journal of Educational Psychology, 63* (1), 74-89.

Cook, C.R., Volpe, R.J., & Livanis, A. (2010). Constructing a roadmap for future universal screening research beyond academics. *Assessment for Effective Intervention, 35*(4), 197-205.

Crooks, C.V. (2005). Predicting academic difficulties: Does a complex, multipdimensional model outperform a unidimensional teacher rating scale. *Canadian Journal of Behavioural Science, 37*(3), 170-180.

Del'Homme, M., Kasari, C., Forness, S., & Bagley, R. (1996). Prereferral intervention and students at risk for emotional and behavioral disorders. *Education and Treatment of Children, 19*(3), 272−285.

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

DiPerna, J.C. & Elliott, S.N. (2002). Promoting academic enablers to improve student achievement: An introduction to the Mini-Series. *School Psychology Review, 31*(3), 293-297.

DiPerna, J.C., Volpe, R.J., & Elliott, S.N. (2001). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review, 31*(3), 298-312.

DiPerna, J.C. (2006). Academic enablers and student achievement: Implications for assessment and intervention services in the schools. *Psychology in the Schools, 43*(1), 7-17.

Ditkowsky, B. & Koonce, D.A. (2010). Predicting performance on high-stakes assessment for proficient students and students at risk with oral reading fluency growth. *Assessment for Effective Intervention, 35*, 159-167.

Dowdy, E., Twyford, J.M., Chin, J.K., DiStefano, C.A., Kamphaus, R.W., & Mays, K.L. (2011). Factor structure of the BASC-2 Behavior and Emotional Screening System Student Form. *Psychological Assessment, 23*(2), 379-387.

Drummond, T. (1994). *The student risk screening scale (SRSS)*. Grants Pass, OR: Josephine County Mental Health Program.

DuPaul, G. J., Power. T. J., Anastopoulos. A. D., & Reid, R. (1998). *ADHD Rating Scale~lV.* New York: Guilford Press.

Elliott, S.N., Busse, R.T., & Gresham, F.M. (1993). Behavior rating scales: Issues of use and development. *School Psychology Review, 22*(2).

Elliott, S.N., DiPerna, J.C., Mroch, A.A., & Lang, S.C. (2004). Prevalence and patterns of academic enabling behaviors: An analysis of teachers' and students' ratings for a national sample of students. *School Psychology Review, 33*(2), 302-309.

Elliott, S.N., & Gresham, F.M. (2007). *Social skills improvement system: Classwide intervention program.* Bloomington, MN: Pearson.

Espin, C., Wallace, T., Campbell, H., Lembke, E.S., Long, J.D., & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children, 74*(2), 174-193.

Fergusson, D.M., Horwood, L.J., & Ridder, E.M. (2005). Show me the child at seven: The consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry*, 46, 837–849.

Feshbach, N.D. & Feshbach, S. (1987). Affective processes and academic achievement. *Child Development, 58*, 1335-1347.

Field, A. (2005). Discovering statistics using SPSS (2nd ed.). London Sage Publications Ltd.

Fleming, C.B., Harachi, T.W., Cortes, R.C., Abbott, R.D., & Catalano, R.F. (2004). Level and change in reading scores and attention problems during elementary school as predictors of problem behavior in schools. *Journal of Emotional and Behavioral Disorders, 12*(3), 130-144.

Fleming, C.B., Haggerty, K.P., Catalano, R.F., Harachi, T.W., Mazza, J.J., and Gruman, D.H. (2005). Do social and behavioral characteristics targeted by preventive interventions predict standardized test scores and grades? *Journal of School Health, 75*(9), 342-349.

Flom, P.L. & Cassell, D.L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. Paper presented at the meeting of the Northeast SAS User's Group.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, (2), 121-139.

Fuchs, L.S., Fuchs, D., & Speece, D.L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25*(1), 33-45.

Gansle, K.A., Noell, G.H., VanDerHeyden, A.M, Naquin, G.M., & Slider, N.J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology Review, 31*(4), 477-497.

Gansle, K.A., Noell, G.H., VanDerHeyden, A.M., Slider, N.J., Naquin, G.M., Hoffpauir, L. D., & Whitmarsh, E.L. (2004). An examination of the criterion validity and sensitivity of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41,* 291-300.

George, S., & Wilkeson, D. (1989). *Early prevention of school failure.* Peotone, IL: Early Prevention of School Failure.

Glaros, A.G. & Kline, R.B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*(6), 1013-1023.

Glover, T.A. & Albers, C.A. (2007). Considerations for evaluating universal screening instruments. *Journal of School Psychology, 45,* 117-135).

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

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 38*(5), 581–586.

Goodman, R., Ford, T., Simmons, H., Gatward, R., Meltzer H. (2000). Using the strengths and differences questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry, 177,* 534-539.

Graham, J.W., Olchowski, A.E., & Gilreath, T.D. (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213.

Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576.

Gresham, F.M., & Elliott, S.N. (1990). *Manual for the Social Skills Rating System*. Circle Pines, MN: American Guidance Service.

Gresham, F.M. (2005). Response to intervention: An alternative means of identifying students as emotionally disturbed. *Education and Treatment of Children, 28*(4), 328-344.

Gresham, F.M., & Elliott, S.N. (2008a). *Social Skills Improvement System-Rating Scales*. Minneapolis, MN: Pearson Assessments.

Gresham, F.M., & Elliott, S.N. (2008b). *Social Skills Improvement System-Rating Scales Manual*. Minneapolis, MN: Pearson Assessments.

Gresham, F.M., Elliott, S.N., Cook, C.R., Vance, M.J., & Kettler, R. (2010). Cross-informant agreement ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System – Rating Scales. *Psychological Assessment, 22*(1), 157-166.

Groth-Marnat, G. (2009). *Handbook of Psychological Assessment.* New York: Wiley.

Guzman, M.P., Jellinek, M., George, M., Hartley, M., Squicciarini, A.M., Canenguez, K.M., Kuhlthau, K.A., Yucel, R., White, G.W., Guzman, J., & Murphy, J.M. (2011). Mental health matters in elementary school: first grade screening predicts fourth grade achievement test scores. *Eur Child Adolesc Psychiatry, 20,* 401-411.

Hale, L., Berger, L.M, LeBourgeois, M.K., & Brooks-Gunn, J. (2011). A longitudinal study of preschoolers' language-based bedtime routines, sleep duration, and well-being. *Journal of Family Psychology, 25*(3), 423-433.

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

Hayling, C.C., Cook, C., Gresham, F.M., Slate, T., & Kern, L. (2007). An analysis of the status and stability of the behaviors of students with emotional and behavioral difficulties. *Journal of Behavioral Education, 17*, 24-42.

Helstela. L. & Sourander, A. (2005). Childhood predictors of externalizing and internalizing problems in adolescence. *Eur Child Adolesc Psychiatry*, 14, 415-423.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *Journal of Special Education, 36*, 102-112.

Henderson, M. (2009). Predicting performance on high stakes testing: Validity and accuracy of curriculum-based measurement of reading and writing. (Unpublished doctoral dissertation) Louisiana State University, Baton Rouge, Louisiana.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*, 127-155.

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

Hoge, R.D. & Luce, S. (1979). Predicting academic achievement from classroom behavior. *Review of Educational Research, 49*(3), 479-496.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *Iowa Test of Basic Skills*. Chicago: Riverside.

Hosp, M.K, Hosp, J.L., & Howell, K.W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement*. New York, NY. Guilford.

Howell, D.C. (2009, March). Treatment of missing data. Retrieved from: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

Hudziak, J.J., Copeland, W., Stanger, C., Wadsworth, M. (2004). Screening for DSM-IV externalizing disorders with the Child Behavior Checklist: a receiver-operating characteristic analysis. *Journal of Child Psychology and Psychiatry (45)*, 7, 1299-1307.

IBM (2011). IBM SPSS Missing Values 20. Retrieved from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Missing_Values.pdf

Ikeda, M.J., Neessen, E., & Witt, J.C. (2007). Best practices in universal screening. *Best Practices in School Psychology V*, Wakefield, CT: The Charlesworth Group

Jenkins, J., & Pany, D. (1978). Standardized achievement tests: How useful for special education? *Exceptional Children*, *44*(6), 448-453.

Jenkins, J.R., Deno, S.L., & Mirkin, P.K. (1979). Pupil progress: Measuring pupil progress toward the least restrictive alternative. *Learning Disability Quarterly*, *2*(4), 81-91.

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

Jiban, C.L. & Deno, S.L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate? *Assessment for Effective Interventions*, *32*(2), 78-89.

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior Assessment System for Children—Second Edition (BASC–2): Behavioral and Emotional Screening System (BESS)*. Bloomington, MN: Pearson.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 Behavioral and Emotional Screening System (BESS) manual*. Circle Pines, MN: Pearson.

Keller-Margulis, M. A., Shapiro, E. S., Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*, 374-390.

Kettler, R.J., Elliott, S.N., Davies, M., & Griffin, P. (2009). Using academic enabler nominations and social behavior ratings to predict students' performance level on Australia's national achievement test. Paper presented at the American Educational Research Association, San Diego, CA.

Kettler, R.J., Elliott, S.N., Davies, M., & Griffin, P. (2011). Testing a multi-stage screening system: Predicting performance on Australia's national achievement test using teachers' ratings of academic and social behaviors. *School Psychology International*, *33*(1), 93-111

Lambert, N.M. & Nicoll, R.C. (1977). Conceptual model for nonintellectual behavior and its relationship to early reading achievement. *Journal of Educational Psychology, 69*(5), 481-490.

Lambert, N., Hartsough, C., & Sandoval, J. (1990) *Manual for the Children's Attention and Adjustment Survey.* Palo Alto, CA: Consulting Psychologists Pree.

Lane, K. L., Kalberg, J.R., Parks, R.J., & Carter, E.W. (2008). Student risk screening scale: Initial evidence for score reliability and validity at the high school level. *Journal of Emotional and Behavioral Disorders, 16, 178-190*.

Lloyd, J.W., Kauffman, J.M., Landrum, T.J., & Roe, D.L. (1991). Why do teachers refer pupils for special education? An analysis of referral records. *Exceptionality, 2*(3), 115-126.

Louisiana Department of Education (2009). *Bulletin 1508: Louisiana Pupil Appraisal Handbook*. Retrieved from: http://www.doe.state.la.us/divisions/specialp/bulletin_1508_training.html.

Louisiana Department of Education (2012). *What are the LEAP, iLEAP, GEE, and End of Course Tests?* Retrieved from: http://www.doe.state.la.us/testing/

Mazzocco, M.M. & Thompson, R.E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice, 20*(3), 142-155.

Malecki, C.K. & Elliott, S.N. (2002). Children's social behaviors as predictors academic achievement: a longitudinal analysis. *School Psychology Quarterly*, *17*(1), 1-23.

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

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

McIntosh, R.H., Chard, D.J., Boland, J.B., & Good, R.H. (2006). The use of reading and behavior screening measures to predict nonresponse to school-wide positive behavior support: A longitudinal analysis. *School Psychology Review*, *35* (2), 275-291.

Meltzer, L. (1984). An analysis of the learning styles of adolescent delinquents. *Journal Of Learning Disabilities, 17* (10), 600-608.

Menesses, K.F. (2011). Using curriculum-based measures to predict math performance on a statewide assessment. (Unpublished doctoral dissertation). Louisiana State University, Baton Rouge, Louisiana.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741-749

Messick, S. (1989). Validity. In R.L. Linn (Ed). *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Minnesota Department of Education & Beck Evaluation and Testing Associates, Inc. (1997). *Minnesota Basic Skills Test of Written Composition/Minnesota Comprehensive Assessments (BST/MCA).* St. Paul, MN: Minnesota Department of Education.

Mitzel, H. C., & Borden, C. F. (2000). *LEAP for the 21st century: 1999 operational final technical report.* Monterey, CA: CTB/McGraw-Hill.

Moffitt, T. E. (1990). Juvenile delinquency and Attention Deficit Disorder: Boys'development trajectories from age 3 to age 15. *Child Development, 61*, 893–910

National Center on Response to Intervention (2010, April). Essential components of RTI – A closer look at response to intervention. Retrieved from http://www.rti4success.org/pdf/rtiessentialcomponents_042710.pdf

Nelson, R.J., Benner, G.J., Lane, K., & Smith, B.W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional Children, 71*(1), 59-73

No Child Left Behind (NCLB) Act of 2001. Pub. L. No. 107-110, H.R. 1, 115 Stat. 142.

Noell, G.H. & Burnes, J.L. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*, 37-49.

Patty, E.F., Hunter, K.K., Chenier, J.S., & Gresham, F.M. (2011). Influencing Pro-Social Change in Students: What Matters? Paper presented at the 43rd annual convention of the National Association of School Psychologists, San Francisco, CA.

Pearson (2000). *AIMSweb Behavior: Administration and Technical Manual.* Retrieved from: https://AIMSweb® .pearson.com/downloads/AIMSweb® _Behavior_Manual.pdf. PsychCorp. Bloomington, MN

Ponsor MA (2006) ROSIE D. et al. v. Mitt ROMNEY et al. In: United States District Court District of Massachusetts (ed) Civil Action No. 01 30199 MAP. Boston, MA, p 410 F.Supp.412d 418.

Reid, R., Gonzales, J.E., Nordness, P.D., Trout, A., & Epstein, M.H. (2004). A meta-analysis of the academic status of students with emotional/behavioral disturbance. *The Journal of Special Education, 38*(3), 130-143.

Richards, C.M., Symons, D.K., Greene, C.A., & Szuszkiewicz, T.A. (1996). The bidirectional relationship between achievement and externalizing behavior problems of students with learning disabilities. *Journal of Learning Disabilities, 28*(1), 8-17.

Rosenthal, R. (2002). The Pygmalion Effect and its mediating mechanisms. *Improving Academic Achievement*. Elsevier Science.

Rosenthal, R. & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports, 19,* 115-118.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489.

Shaffer, D.. Fisher. P.^ & Lucas. C. (1998). *Computerized Diagnostic Interview Schedule for Children (CDISC 4.0).* New York: New York State Psychiatric Institute.

Severson, H.H., Walker, H.M., Hope-Doolittle, J., Kratochwill, T.R., & Gresham, F.M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology, 45,* 193-223.

Shaffer, D., Fisher, P., & Lucas, C. (1998). *Computerized Diagnostic Interview Schedule for Children*. New York: New York State Psychiatric Institute.

Shapiro, E. S., Keller, M. A., Edwards, L., Lutz, G., & Hintze, J. M. (2006). General outcome measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *42,* 19-35.

Shaw, R. & Shaw, D. (2002) DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP). (Technical Report) Eugene, OR: University of Oregon.

Soli, S.D. & Devine, V.T. (1976). Behavioral correlates of schievements: A look at high and low achievers. *Journal of Educational Psychology, 68,* 335-341.

Solomon, B.G., Klein, S.A., Hintze, J.M., Cressey, J.M., & Peller, S.L. (2012). A meta-analysis of school-wide positive behavior support: An exploratory study using single-case synthesis. *Psychology in the Schools*, *49*(2), 105-121.

Spaulding, S. A., Horner, R. H., May, S. L., & Vincent, C. G. (2008, November). *Evaluation brief: Implementation of school-wide PBS across the United States*. OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. Web site: http://pbis.org/evaluation/evaluation_briefs/default.aspx

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

Sternberg, R. J. (2005). Intelligence, competence, and expertise. In A. J. Elliott & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 15-30). New York: The Guilford Press.

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality, 17*(4), 223 – 237.

Sugai, G., & Horner, R. H. (2009b). Defining and describing schoolwide positive behavior support. In W. Sailor, G. Dunlop, G. Sugai, & R. H. Horner (Eds.), *Handbook of positive behavior support* (pp. 307 – 326). New York: Springer Publishing.

Thomas, J.D., Presland, I.E., Grant, M.D., & Glynn, T. (1978). Natural rates of teacher approval and disapproval in grade-7 classrooms. *Journal of Applied Behavior Analysis*, 11, 91–94.

Trout, A.L, Nordness, P.D., Pierce, C.D., & Epstein, M.H. (2003). Research on the academic status of children with emotional and behavioral disorders: A review of the literature from 1961 to 2000. *Journal of Emotional and Behavioral Disorders, 11*(4), 198-210.

Volpe, R.J., DuPaul, G.J., DiPerna, J.C., Jitendra, A.K., Lutz, G., Tresco, K., & Junod, R.V. (2006). Attention deficit hyperactivity disorder and scholastic achievement: A model of mediation via academic enablers. *School Psychology Review, 35*(1), 47-61.

Vygotsky, L.S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Walker, H.M. & McConnell, S.R. (1988). *Walker-McConnell scale of social competence and school adjustment.* Pro-Ed.

Walker, H., & Severson, H. (1990). *Systematic screening for behavior disorders (SSBD).* Longmont, CO: Sopris West.

Walker, H. M., Ramsey, E., & Gresham, F. M. (2004). *Antisocial behavior in school: Evidence based practices* (2nd ed.). Belmont, CA: Wadsworth.

Wayman, J.C. (2003). Multiple imputation for missing data: What is it and how can I use it? Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.

Wentzel, K.R. (1991). Relations between social competence and academic achievement in early adolescent. *Child Development, 62,* 1066-1078.

Wentzel, K.R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology, 85*(2), 357-364.

Wentzel, K. R. (2005). Peer relationships, motivation, and academic performance at school. In A. Elliot & C. Dweck (Eds.)*, Handbook of Competence and Motivation* (pp. 279– 296). New York: Guilford.

Werthamer-Larsson, L., Kellam, S.G., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585-602.

Winett, R.A. & Winkler, R.C. (1972). Current behavior modification in the classroom: be still, be quiet, be docile. *Journal of Applied Behavior Analysis*, 5(4), 499-504.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*(6), 412-422.

APPENDIX A

Study Sample Demographic Information

Grade Levels	All	3	4	5	6	7	8
Number of Students	750	153	162	122	149	83	81
Sex							
Male	375	71	84	66	76	36	42
Female	375	82	78	56	73	47	39
Race							
African American	530	100	108	79	105	72	66
Caucasian	211	48	52	41	44	11	15
Asian	0	0	0	0	0	0	0
Hispanic	6	1	3	2	0	0	0
Am. Indian/Alaskan	1	1	0	0	0	0	0
Other	3	3	0	0	0	0	0
Socioeconomic Status							
Free/Reduced Lunch	655	132	139	100	134	75	73
Paid Lunch	95	19	23	22	15	8	8

APPENDIX B

Range of scaled scores associated with math achievement levels on *i*LEAP and LEAP tests for 2010-2011 school year

*i*LEAP Grade 3

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	383-500	386-500
Mastery	338-382	342-385
Basic	282-337	283-342
Approaching Basic	239-281	245-282
Unsatisfactory	100-238	100-244

*i*LEAP Grade 6

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	387-500	394-500
Mastery	341-386	358-393
Basic	280-340	281-357
Approaching Basic	239-279	248-280
Unsatisfactory	100-232	100-247

LEAP Grade 4

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	408-500	419-500
Mastery	354-407	370-418
Basic	301-353	315-369
Approaching Basic	263-300	282-314
Unsatisfactory	100-262	100-281

*i*LEAP Grade 7

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	383-500	421-500
Mastery	344-382	376-420
Basic	286-343	292-375
Approaching Basic	236-285	255-291
Unsatisfactory	100-235	100-254

iLEAP Grade 5

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	386-500	405-500
Mastery	341-385	355-404
Basic	286-340	282-354
Approaching Basic	247-285	250-281
Unsatisfactory	100-246	100-249

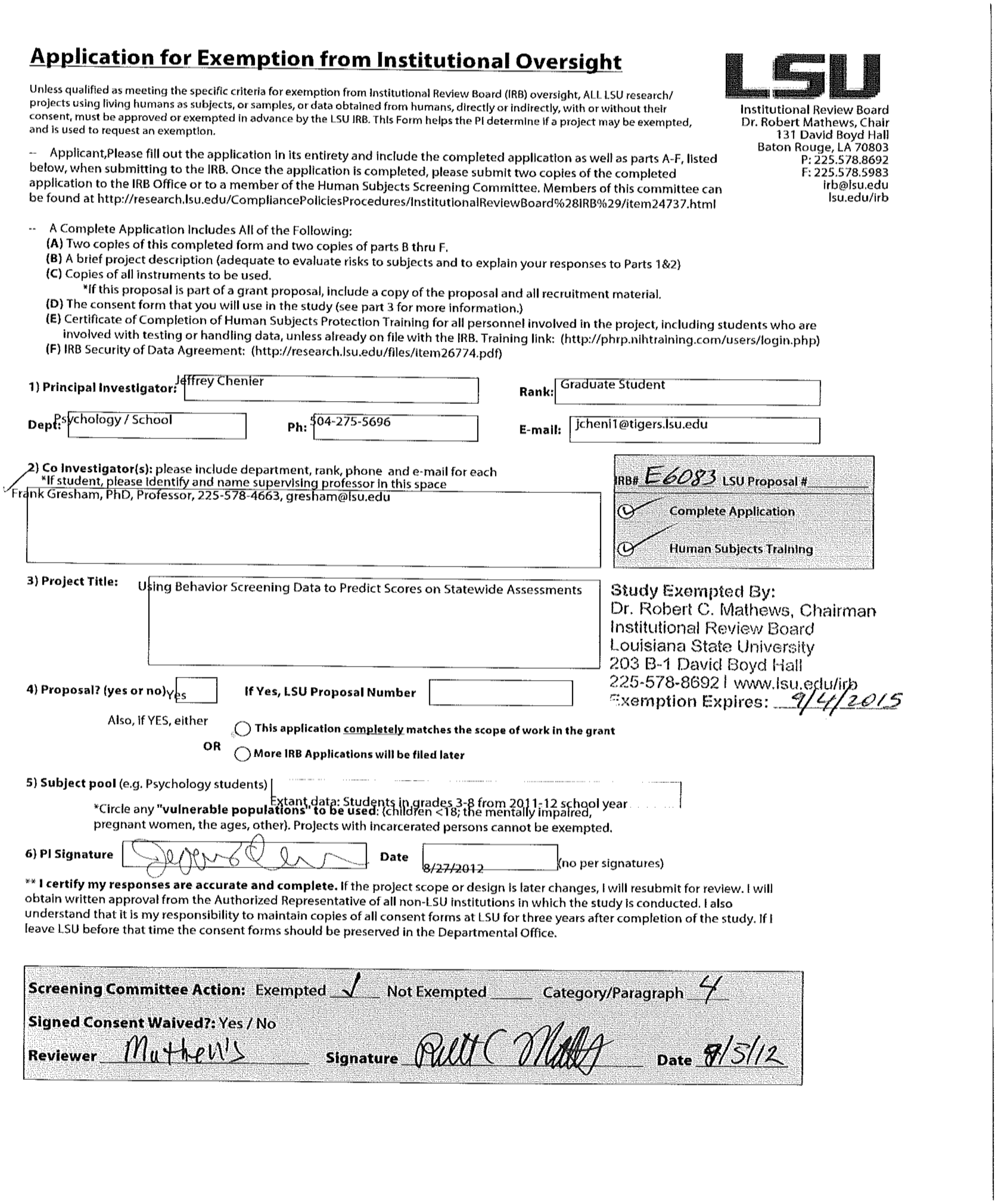
LEAP Grade 8

Achievement Level	Scaled Score Range	
	ELA	Math
Advanced	402-500	398-500
Mastery	356-401	376-397
Basic	315-355	321-375
Approaching Basic	269-314	296-320
Unsatisfactory	100-268	100-295

** Source: http://www.doe.state.la.us/testing/

APPENDIX C

Institutional Review Board Approval



Application for Exemption from Institutional Oversight

LSU
Institutional Review Board
Dr. Robert Mathews, Chair
131 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu
lsu.edu/irb

Unless qualified as meeting the specific criteria for exemption from Institutional Review Board (IRB) oversight, ALL LSU research/ projects using living humans as subjects, or samples, or data obtained from humans, directly or indirectly, with or without their consent, must be approved or exempted in advance by the LSU IRB. This Form helps the PI determine if a project may be exempted, and is used to request an exemption.

-- Applicant,Please fill out the application in its entirety and include the completed application as well as parts A-F, listed below, when submitting to the IRB. Once the application is completed, please submit two copies of the completed application to the IRB Office or to a member of the Human Subjects Screening Committee. Members of this committee can be found at http://research.lsu.edu/CompliancePoliciesProcedures/InstitutionalReviewBoard%28IRB%29/item24737.html

-- A Complete Application Includes All of the Following:
(A) Two copies of this completed form and two copies of parts B thru F.
(B) A brief project description (adequate to evaluate risks to subjects and to explain your responses to Parts 1&2)
(C) Copies of all instruments to be used.
*If this proposal is part of a grant proposal, include a copy of the proposal and all recruitment material.
(D) The consent form that you will use in the study (see part 3 for more information.)
(E) Certificate of Completion of Human Subjects Protection Training for all personnel involved in the project, including students who are involved with testing or handling data, unless already on file with the IRB. Training link: (http://phrp.nihtraining.com/users/login.php)
(F) IRB Security of Data Agreement: (http://research.lsu.edu/files/item26774.pdf)

1) Principal Investigator: Jeffrey Chenier **Rank:** Graduate Student

Dept: Psychology / School **Ph:** 504-275-5696 **E-mail:** jcheni1@tigers.lsu.edu

✓ **2) Co Investigator(s):** please include department, rank, phone and e-mail for each
*If student, please identify and name supervising professor in this space
Frank Gresham, PhD, Professor, 225-578-4663, gresham@lsu.edu

IRB# E6083 **LSU Proposal #**
✓ **Complete Application**
✓ **Human Subjects Training**

3) Project Title: Using Behavior Screening Data to Predict Scores on Statewide Assessments

Study Exempted By:
Dr. Robert C. Mathews, Chairman
Institutional Review Board
Louisiana State University
203 B-1 David Boyd Hall
225-578-8692 | www.lsu.edu/irb
Exemption Expires: 9/4/2015

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

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

5) Subject pool (e.g. Psychology students) Extant data: Students in grades 3-8 from 2011-12 school year
*Circle any **"vulnerable populations" to be used**: (children <18; the mentally impaired, pregnant women, the ages, other). Projects with incarcerated persons cannot be exempted.

6) PI Signature **Date** 8/27/2012 (no per signatures)

** **I certify my responses are accurate and complete.** If the project scope or design is later changes, 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 4
Signed Consent Waived?: Yes / No
Reviewer Mathews **Signature** **Date** 9/5/12

VITA

Jeffrey S. Chenier is a candidate for the Doctor of Philosophy degree in the school psychology program at Louisiana State University. He graduated with a Bachelor of Science degree in psychology in 2007 and received a Master of Arts degree in psychology in 2010 from the Louisiana State University. Jeff is currently working as a school psychologist on the pupil appraisal team in East Feliciana Parish School District. Jeff completed his graduate work under the supervision of Dr. Frank M. Gresham.