UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Identifying Students for Intervention in Math Problem Solving: An Evaluation of Fluency-Based Word Problem Solving Measures

Permalink

https://escholarship.org/uc/item/75z661zm

Author

Sisco-Taylor, Dennis T.

Publication Date

2014

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Identifying Students for Intervention in Math Problem Solving: An Evaluation of Fluency-Based Word Problem Solving Measures

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Education

by

Dennis Trévaughn Sisco-Taylor

December 2014

Dissertation Committee:
Dr. H. Lee Swanson, Chairperson
Dr. Michael L. Vanderwood
Dr. Michael J. Orosco

Copyright by
Dennis Trévaughn Sisco-Taylor
2014

The Dissertation of Dennis Trévaughn Sisco-Taylor is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

Acknowledgements

Data used in this dissertation study were collected as part of a larger study funded by the U.S. Department of Education, Cognition and Student Learning in Special Education (USDE R324A090002), Institute of Education Sciences, awarded to H. Lee Swanson (dissertation chairperson).

The text of this dissertation, in part, is a reprint of the material as is appears in Do curriculum-based measures predict performance on word-problem-solving measures? [Sisco-Taylor, Fung, & Swanson, October, 2014]. The co-author, H. Lee Swanson, listed in that publication directed and supervised the research which forms the basis for this dissertation. The co-author, Wenson Fung, provided technical expertise and played an integral role in the editing process. Material presented in the manuscript was re-written, expanded, and is discussed in greater detail in the dissertation.

Dedication

This work is dedicated to the Sisco-Taylor family at large for being a remarkable support system for me throughout my entire life, especially my years in graduate school. Specific thanks go out to my parents, Tomasina and Dennis, for imparting your exceptional work ethics, and always providing me with your unconditional love. I also dedicate this work to the loving memory of my late grandmother, Tomasina Sisco.

Thank you to my dissertation chair, Dr. H. Lee Swanson, for your mentorship, and guidance. I am also especially grateful for the opportunity to work on your grant, and all of the opportunities it afforded me.

Additionally, special thanks to Dr. Mike Vanderwood, for your mentorship throughout my years in graduate school. You have helped me develop into a better school psychologist, both as a scholar and practitioner.

ABSTRACT OF THE DISSERTATION

Identifying Students for Intervention in Math Problem Solving:
An Evaluation of Fluency-Based Word Problem Solving Measures

by

Dennis Trévaughn Sisco-Taylor

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, December 2014
Dr. H. Lee Swanson, Chairperson

This study examined the extent to which initial performance, and growth on an experimental CBM word problem solving fluency measure (WPSF) were predictive of student performance on criterion measures of math problem solving. In addition, the extent to which WPSF could correctly classify students as a function of risk status was evaluated. Alternate forms of the WPSF measure were administered to 142 third grade students, along with multiple criterion measures of math problem solving. Results indicate that WPSF demonstrated moderate criterion validity, and was able to discriminate between students that were at-risk and not at-risk for problem solving difficulties. Implications for assessment practices in mathematics are discussed.

Table of Contents

Identifying Students for Intervention in Math Problem Solving:

An Evaluation of Fluency-Based Word Problem Solving Measures

The math achievement of America's youth has received an increased amount of attention in recent years. This is due, in large part, to the highly publicized and unfavorable comparisons between American students and their international counterparts in mathematics. Moreover, the struggle to bring students to a level of proficiency in mathematics in their primary years of schooling, and prepare them for advanced coursework at the secondary level has been well-documented. The National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 2011) revealed that only 40% of fourth grade students reached a level of proficiency in mathematics. Older students did not fare any better on the National Assessment of Educational Progress, as only 35% of eighth grade students reached a level of proficiency. Furthermore, survey data reflect that most secondary-level math teachers feel that their students' are unprepared for algebra coursework (National Mathematics Advisory Panel, 2008). These findings have troubling implications since algebra is seen by many as the "gate-keeper" to higher-level math courses, and groups such as the National Mathematics Advisory Panel (NMAP) have set the goal for all students to progress through a sequence of coursework that includes Algebra I, Geometry, and Algebra II (NMAP, 2008).

Ensuring that all students possess the skills necessary to progress through Algebra II courses at the secondary level satisfies the basal level of preparation imparted by the NMAP. However, education initiatives put forth by President Obama, such as "Educate to Innovate" (Whitehouse.gov, 2009), call for an even greater level of preparation in

mathematics for American students. Advancement in mathematics education is a core component of the President's plan to bring more Americans into science, technology, engineering, and mathematics (STEM) fields. This is because a lack of preparation in mathematics can serve as a barrier for entry into STEM fields. Students that lack the requisite skills to take on advanced coursework (i.e., trigonometry, calculus) in high school are less likely to go on to four year universities, and thus, unlikely to receive the training necessary to enter STEM fields (DeJarnette, 2012). Moreover, the sequence of math coursework in secondary education (algebra I, followed by geometry, algebra II, trigonometry, and then calculus) does not allow students to take on advanced courses if they have not mastered algebra by the time they enter high school. Thus, students who do not pass an Algebra I course by the eighth grade are unlikely to have the opportunity to undergo coursework that would prepare them for entry to STEM fields. Unfortunately, it seems as though many more students will soon face this reality since only 34% of eighth grade students were enrolled in Algebra I courses in 2011 (NCES, 2011).

The need to bring students to a level of proficiency in mathematics at a young age is clear. Beyond preparing students for college, or entry to STEM fields, bringing students to a level of proficiency in mathematics has implications for many basic life skills. For example, most students will face the need to purchase goods or services, perform household budgeting, or even complete technical tasks in the workplace that require basic math skills (Lembke, Hampton, & Beyers, 2012). Thus, echoing the sentiments of the NMAP, systemic changes are necessary to provide children with additional support in mathematics during their primary years of education.

One major point of emphasis in the final report of the NMAP (2008) was the need to utilize formative assessment practices in schools to improve math education. Formative assessment can be defined as the ongoing monitoring of student learning to inform instruction (NMAP, 2008). It has been shown to be effective in improving student outcomes in mathematics in multiple reviews of the literature (Gersten, Chard, Jayanthi, Baker, Morphy, & Flojo, 2009; NMAP, 2008). However, there has been disagreement among educators with regard to what the formative assessment process actually entails (Bennett, 2011). While some have utilized informal assessment methods that rely on observations, interviews, rubrics, etc. (e.g., Watson, 2006), others have used more systematic approaches that incorporate data from standardized performance tasks (Fuchs, Fuchs, Karns, et al., 1999). Put simply, all formative assessment approaches are not equally effective. The largest effects that have been reported for formative assessment practices have occurred when teachers used performance assessments to evaluate specific academic skills, and subsequently used the results from those assessments to make instructional changes (Gersten, Chard, et al., 2009). These effects were strengthened further when guidance was given to teachers on using the assessment data to make instructional changes. Given the importance of formative assessment to student learning, and the strong recommendations from the NMAP to include these processes in schools, it is important to present schools with an organized framework for integrating formative assessment practices and provide educators with instruments that can monitor student learning.

Response to Intervention Models for Mathematics

Response to intervention (RTI) models present an established framework for implementing formative assessment practices. They are described as early detection, prevention, and support systems that identify struggling students and provide them with assistance before they fall behind (Gersten, Beckmann, et al., 2009). As discussed in Riccomini and Witzel (2010), two types of formative assessment are utilized in RTI models: universal screening and progress monitoring. Lembke, Hampton, and Beyers (2012) described RTI programs as follows: First, students are given "generally effective" instruction by their classroom teacher (tier 1). As this occurs, students undergo a universal screening process where their skills are evaluated at three time points during the school year. Next, students that score below a pre-specified benchmark on the screening measure are placed in academic interventions that will provide supplemental support in the area of concern (tier 2). While receiving this supplemental instruction, students' progress is monitored. Additional levels of support are provided for students that continue to demonstrate a need while receiving the supplemental instruction (tier 3). This additional layer of support is generally a more intensive, and individualized form of intervention. Again, student progress is monitored closely to evaluate the effectiveness of the given intervention.

In the context of mathematics instruction: (1) All students would receive evidence-based instruction which targets conceptual understanding, computational fluency, and problem-solving skills; (2) All students undergo a universal screening process, where their skills are evaluated in these critical areas; (3) Students that do not

meet benchmark goals receive explicit supplemental instruction in the area of concern, and have their progress monitored continuously. This RTI framework has garnered empirical support (e.g., Denton, Fletcher, Anthony, & Francis, 2006; Speece, Case, & Molloy, 2003) with respect to reading applications, however, very little research has been conducted to examine its' efficacy in improving math outcomes. While RTI models in mathematics have been under-researched, some (e.g., Gersten, Beckmann, et al., 2009; Lembke, Hampton, & Beyers, 2012; Riccomini and Witzel 2010) have suggested that critical elements of RTI reading models could be applied to mathematics.

Fuchs, Compton, Fuchs, Paulsen, Bryant, and Hamlett (2005) reported a significant increase in the rates of improvement (ROI) of at-risk students that received supplemental math supports in comparison to at-risk students that were not administered the supports (*ES* = .40-.67). In addition to out-performing at-risk controls, the students that received the supplemental instruction demonstrated ROIs that were similar to, or better than not-at-risk peers on measures of calculation and applications (*ES* = .11-.45). Also, in a randomized control trial, Fuchs, Fuchs, Craddock, Hollenbeck, Hamlett, and Schatsneider (2008) demonstrated that the combination of evidence-based classroom instruction and supplemental intervention was more effective than typical classroom instruction (i.e., non-evidence-based curriculum) and small group intervention (*ES* = 1.34).

There is also an extensive body of literature on the effectiveness of math interventions that have been implemented outside of the RTI framework (see Gersten, Chard, et al., 2009 for review). Altogether, these findings suggest that students with math

difficulties benefit from the types of procedures that are evoked in RTI models. The formative assessment practices that are incorporated in RTI models, such as universal screening and progress monitoring, are critical components of this process since they help inform both general classroom instruction and supplemental instruction.

Using RTI Models to Target Critical Foundations of Algebra

K-8 math education should provide the basic foundation for algebra. The NMAP (2008) identified three clusters of concepts and skills that they referred to as the *Critical Foundations of Algebra.* These three broad clusters were: fluency with whole numbers, fluency with fractions, and particular aspects of geometry and measurement. Of these three critical foundations, fluency with whole numbers is the most pertinent to early elementary mathematics since students are expected to have mastered these skills by the end of grades 5 or 6 (NMAP, 2008). The NMAP (2008) gave a detailed description of what this critical foundation entails:

> It must clearly include a grasp of the meaning of the basic operations of addition, subtraction, multiplication, and division. It must also include use of the commutative, associative, and distributive properties; computational facility; and the knowledge of how to apply the operations to problem solving. (NMAP, 2008, pp.17)

Despite this exhaustive definition of fluency with whole numbers that includes aspects of applications and problem solving, American math education has often been criticized as disproportionately focusing on number facts and computation skills (e.g., Fletcher, Lyon, Fuchs, & Barnes, 2007). Application-based aspects of whole number fluency, such as

word problems have generally been ignored. When they have been studied, they have been limited to simple, one-step, arithmetic story problems with contrived narratives (Fuchs, Fuchs, & Prentice, 2004). Fuchs, Fuchs, Karns, Hamlett, and Katzaroff (1999) commented that "mathematics education is typified by shallow coverage of a large number of topics" (p. 610). Organizations such as the National Council of Teachers of Mathematics (NCTM) have taken similar positions, calling for an increased emphasis on the development of conceptual understanding and problem solving skills in elementary math education (NCTM, 2000). In a general sense, perspectives on math pedagogy have shifted from a vertical transfer perspective, where mastery of many simple skills facilitates acquisition of more complex skills, to a lateral transfer perspective, where children recognize patterns across numerous experiences in order to abstract generalized problem-solving principles or schemata (Fuchs et al., 1999).

The Importance of Word Problems. This paradigm shift in the approach to math education was likely due to the fact that students often have difficulty applying the isolated math skills they acquire. Such is the case with word problems. For example, only 31% of fourth grade students could correctly answer the following problem from the National Assessment of Educational Progress (NAEP):

> The early show and the late show for a movie last the same amount of time. The early show begins at 3:15 PM and ends at 4:27 PM. The late show begins at 7:30 PM. At what time does the late show end? Show your work. (NCES, 2011, pp. 32)

These types of word problem difficulties must be taken seriously since word problem skills have been shown to be predictive of pre-algebraic knowledge (Fuchs, Compton, et al., 2012).

Fuchs, Compton, and colleagues (2012) discussed the importance of word problems, arguing that they can be particularly useful for measuring pre-algebraic knowledge because said problems force students to transform the written narratives into algebraic equations (Fuchs, Compton, et al., 2012). Verschaffel and De Corte (1997) proposed that word problems serve as a vehicle for developing conceptual understanding and problem solving skills in mathematics. The reasoning behind their supposition lies within the construction of word problems. Instead of presenting students with number sentences or algorithms, as in computation tasks, word problems force students to use the text to identify missing information, construct the number sentence, and derive the appropriate calculation (Fuchs, Fuchs, Stuebing, et. al, 2008). Furthermore, word problem skills incorporate computation abilities, while also reflecting an understanding of relationships between known and unknown quantities (Fuchs, Compton, et al., 2012; Fuchs, Fuchs, Compton, et al., 2006). Fuchs et al. (1999) also found that teachers' use of word problem measures as performance assessments was beneficial to instructional planning. This is because the samples of work generated from the word problem tasks gave them more insight into how strong students' problem solving skills were, and the extent to which they were improving as a result of the instruction.

The importance of word problems in learning mathematics was also reflected in the recommendations made by Gersten, Beckmann, et al. (2009) in their RTI practice

guide for elementary and middle school mathematics. One of their recommendations, based on the best evidence of effective practices in mathematics, was to have math interventions that included instruction on solving word problems that is based on common underlying structures. Such interventions have acquired strong empirical support in the math literature through a series of well-designed studies (e.g., Fuchs, Fuchs, Prentice, et al., 2003; Fuchs, Fuchs, Prentice, Hamlett, Finelli, & Courey, 2004; Jitendra, Griffin, McGoey, Gardill, Bhat, & Riley, 1998; Xin, Jitendra, & Deatline-Buchman, 2005). These problem-solving interventions teach students about the semantic structure of various types of word problems, how to categorize these problems based on their structure, and how to determine the appropriate solutions for each problem type.

The semantic properties of word problems have been a topic of interest for many years in the math literature. This was the result of research findings reflecting that children's solution strategies were tied specifically to specific semantic structures within word problems (e.g., Carpenter & Moser, 1984; Riley & Greeno, 1988). That is, children developed specific strategies to address certain types of problems. As reported in Riley and Greeno (1988), word problems can generally be placed in one of three categories with respect to their semantic structure: combine, change, or compare.

Combine problems involve two quantities, along with their combination (i.e., sum). Borrowing an example from Carpenter and Moser (1984), "Sara has 4 sugar donuts. She also has 9 plain donuts. How many donuts does Sara have altogether?" (p. 180). Subtraction problems can also fall within this combine category. Using the previous example, if the total number of doughnuts (i.e., sum) was initially given along with one

of the doughnut subtypes (i.e., addend), this would still represent a combine problem. In change problems, there is an initial quantity, a change in quantity, and a resulting quantity. For example, "Tim had 11 candies. He gave 7 candies to Martha. How many candies did Tim have left?" (Carpenter & Moser, 1984, p. 180). As with compare problems, change problems can take the form of addition or subtraction problems. The previous example would change to an addition problem if the resulting quantity was given along with the change in quantity. Finally, compare problems involve two quantities and their difference. To borrow an example from Jitendra, Sczesniak, and Deatline-Buchman (2005), "Joe has 24 CDs. He has 4 less CDs than Tom. How many CDs does Tom have?" (p. 371). When the difference is given, as in the example, these are addition problems. However, when the difference is unknown, these become subtraction problems.

The goal of word problem instruction is to get students to understand that they can apply the same strategies to problems with structural similarities, even when elements of the story within the problem change (Gersten, Beckmann, et al., 2009). In order to reach this level of generalization, students need multiple exposures to word problems with varied superficial features and cover stories (Gersten, Beckmann, et al., 2009). Further, as teachers make the proposed instructional adaptations which call for the inclusion of word problems in the math curriculum, they will also need to adopt assessment practices that help them gauge the extent to which students are developing their word problem competencies. Most modern-day math curricula include word problem exercises. However, for the most part, teachers have been left without assessment tools that can give

them low-inference data on children's problem solving abilities. Without measures that directly assess problem solving performance, teachers must make inferences on problem solving ability based on observation, rubrics, or other methods that may not be reliable (Kelley, Hosp, & Howell, 2008). Under these circumstances, it is possible that students with specific difficulties in math problem solving could go unnoticed.

As previously discussed, RTI models incorporate two types of formative assessment: universal screening and progress monitoring. To recap, universal screening process is carried out in two steps. First, students are administered brief assessments at the beginning, middle, and end of the school year (i.e., fall, winter, and spring). Second, students that fall below a pre-determined benchmark are provided with evidence-based supplemental instruction, and have their progress monitored. The main purpose of progress monitoring is to evaluate program effectiveness as measured by student growth (Riccomini & Witzel, 2010). Instructional decisions can be made by examining student growth. For example, flat or decreasing trajectories are indicative of ineffective programming, and thus, a need for instructional changes. When assessment tools used during the screening and progress monitoring processes are tied to specific academic skills, they are able to identify academic deficits in targeted areas. Thus, assessment tools that are capable of providing valid information with respect to achievement levels in academic areas of concern, and growth in those respective areas are extremely valuable. Curriculum-based measures (CBMs; Deno, 1985) have often been used to fill this role because of their utility as screeners, progress monitoring tools, and potential to highlight specific areas of weakness in a given domain (e.g., Fuchs, Fuchs, & Courey, 2005).

Curriculum-Based Measurement

Curriculum-based measurement refers to an assessment methodology that serves the purpose of indexing academic competence and measuring student growth. CBMs were the result of the efforts of Stanley Deno and his collegues at the University of Minnesota's Institute for Research on Learning Disabilities (IRLD); in their attempt to develop a simple and efficient, yet technically sound measurement system for assisting special education teachers (Stecker, Fuchs, & Fuchs, 2005). CBMs are standardized measures that invoke items that students are likely to see through their general coursework. Beyond their sound technical properties, CBMs possess the quality of providing low-inference information, and can be used in a repeated fashion (Kelley, Hosp, & Howell, 2008). The fact that they provide permanent products of student behavior eliminates the need for educators to draw inferences on students' skills in areas of interest.

While CBMs of mathematics were not originally examined at the IRLD, this line of research was eventually carried out by a couple of Stanley Deno's graduate students, Lynn Fuchs and Mark Shinn. Since the late 1980's, math CBMs targeting computation (e.g., Fuchs, Fuchs, Hamlett, & Stecker, 1990), concepts and applications (Fuchs, Fuchs, Hamlett, et al., 1994), and problem solving (e.g., Fuchs, Fuchs, Karns, et al., 1999) have emerged. As discussed in Kelley, Hosp, and Howell (2008), three types of CBMs are referenced in the literature: general outcome measures (GOMs), skills-based measures (SBMs), and mastery measures (MMs). GOMs are considered capstone tasks, and generally assess a variety of subskills in a broad area of achievement. SBMs tend to

incorporate an array of skills that are specific to one particular domain (i.e., computation or problem solving). Finally, MMs are direct measures of specific skills (i.e., two-digit subtraction); these measures tend to be particularly sensitive to change.

Most of the commercially-available math CBMs are either SBMs or MMs. Due to the multi-topic nature of mathematics (i.e., computation, problem solving), and lack of a true capstone task that exudes math proficiency, the development of a GOM for mathematics has been difficult (Kelley, Hosp, & Howell, 2008). Therefore, in mathematics we are left with SBMs that generally measure performance in the domains of computation and problem solving or applications. Research on word problem CBMs has been relatively limited, especially with regard to evaluating their technical properties (Foegen, Jiban, & Deno, 2007). Most researchers have utilized or evaluated experimental word problem measures; most of which had unique administration and scoring procedures. Given the degree of diversity, a thorough review is in order.

Word Problem Tasks. The most widely-used word problem measure within the math literature to date has been *Story Problems* (i.e., Jordan & Hanich, 2000), which was adapted from the earlier work of Carpenter and Moser (1984); Riley and Greeno (1988). The story problems measure includes a set of arithmetic word problems that require basic number combinations, with sums and minuends less than 9, and can be solved in one step. These word problems reflect the semantic structure (i.e., change, combine, compare) that was theorized by Carpenter and Moser (1984) and discussed previously herein. The dependent measure for story problems is the number of correct answers.

In Jordan and Hanich (2000), the investigators presented the word problems orally while also providing students with a written version to avoid the confounding of reading difficulties. Students were not timed on the task; so problems were only presented once the student completed the previous problem. Investigators allowed students to use physical referents such as coins to solve the problems. With a sample of (N = 49) second grade students, Jordan and Hanich demonstrated that student performance on the story problems task varied as a function of ability status, with average-achieving (AA) students outperforming reading-disabled (RD) students, RD students outperforming math-disabled (MD) students, and MD students outperforming students with combined math and reading disabilities (RD-MD; i.e., AA > RD > MD > RD-MD). This effect of achievement group status (between the four achievement groups) on the story problems measure was quite robust, as Jordan and Hanich reported an η^2 of .56 Internal consistency data were also provided for the story problems measure in this study; Cronbach's α = .84.

In the most recent peer-reviewed evaluation of the story problems measure, Fuchs, Compton, et al. (2012) administered the word-problem task to a sample of (N = 279) second grade students. The administration procedures were relatively consistent with those used in Jordan and Hanich (2000); the only differences were the absence of physical referents (i.e., manipulatives), and giving students 30 seconds to respond to each question. Second grade performance on the story problems measure was predictive of third grade performance on a measure of pre-algebraic knowledge, and the word problem

subtest from the Iowa Test of Basic Skills (*r* 's= .56 and .62, respectively). Internal consistency reported in this study was similar to that of Jordan and Hanich, α = .83.

Word Problem CBMs. To date, very few peer-reviewed investigations have examined the utility of word problem CBMs as predictors of math achievement; that is, measures that included grade-appropriate problems that were sampled from local math curricula. Fuchs, Fuchs, Karns, Hamlett, and Katzaroff (1999) investigated the use of story problem scenarios as performance assessments with a measure they would later call Real-Life Math CBM (CBM-RLM; Fuchs, Fuchs, & Courey, 2005). In this task, students were presented with multi-paragraph narratives that described a problem situation. Teachers read the narratives aloud to students while they followed along, and the assessments were not timed. Each problem was designed to: (1) force students to apply a core set of skills consistent with their grade level; (2) discriminate between relevant and irrelevant information; (3) generate information not present in the narrative; (4) explain their procedural math work; and (5) generate written communication related to the mathematics. Students were evaluated in each of these areas with a 6-point performance rubric.

In Fuchs, Fuchs, Karns, Hamlett, and Katzaroff (1999), the CBM-RLM performance assessment showed moderate criterion validity with the Comprehensive Test of Basic Skills (CTBS) operations and applications subtests (*r*'s = .62 and .67, respectively). Moreover, discriminant validity evidence was provided for the CBM-RLM slopes, as there were ordered significant differences between high- (.54), average- (.28),

and low-achieving students (.11). Fuchs, Fuchs, and Courey (2005) reported low alternate-form test-retest reliability coefficients (*r*'s = .66-.76) for the CBM-RLM.

Jitendra, Scezniak, and Deatline-Buchman (2005) evaluated a curriculum-based word problem-solving task that they called *Word Problem-Solving CBM*. Word problem-solving CBM included word problems that were sampled from third grade math textbooks. Word problems included 6 one-step, and 2 two-step addition and subtraction problems that varied in terms of semantic structure. The measures were group administered, and students were given ten minutes to complete eight word problems. In terms of scoring, students were assigned one point for a correct solution, and one point for a correct number model. The CBMs demonstrated moderate to strong correlations with the procedures and problem-solving subtests (*r*'s = .64 and .71, respectively) from the Stanford Achievement Test-Ninth Edition (SAT-9). The CBMs also accounted for substantive variance (R^2 = .46) in the TerraNova standardized achievement test's concepts and applications subtest, while a computation-fluency measure explained no additional significant variance.

Leh, Jitendra, Caskie, and Griffin (2007) expanded on this earlier work by investigating the utility of the word problem-solving CBMs in providing evidence of student growth in problem solving accuracy over time. Of particular interest, was whether such measures were sensitive to growth, and how these growth rates compared to computation fluency measures. Eight alternate forms of the measure were examined, and growth rates of low-achieving students were compared to those of average students. Their mixed effects growth model revealed significant growth rates (gains of .24 points

per week), but these growth rates were less than those demonstrated by the computation fluency (average gains of .36 points per week), and concepts and applications measures found in previous studies (.37 points per week reported in Shapiro, Edwards, & Zigmond, 2005). Further, the differences between low and average students in terms of growth rates were not significant.

Fuchs, Compton, Fuchs, et al., (2011) examined a word problem measure that they called *Algorithmic Word Problems* with a sample of (N = 122) third grade students. The measure included 10 word problems that required 1-4 steps. The algorithmic word problems measure was group administered, and problems were read aloud to students as they followed along on their own copies. Test administrators moved on to the next problem when it appeared that all but a few students were finished. The algorithmic word problem measure had a small correlation (r = .43) with the Iowa Test of Basic Skills (ITBS) problem solving subtest, and demonstrated good internal consistency (Cronbach's α = .85). Fuchs, Compton, Fuchs, et al. made an additional contribution to the word-problem literature base by examining the diagnostic efficiency (i.e., seeing how well it identified children that were at-risk for math difficulties) of algorithmic word problems, in addition to examining its' criterion validity. Using the 25th percentile on the ITBS as a cutoff for risk status, Fuchs and colleagues conducted a logistic regression analysis. While obtaining an area under the curve (AUC) of .83, the algorithmic word problems measure yielded poor specificity (.48) when using a cut score that maximized sensitivity (.88). Thus, when using a cut score that limited the number of false negatives (students that were identified as not at-risk but later scored below the 25th percentile), the number

of false positives increased (students that were identified as at-risk but later scored above the 25th percentile). This means that in order to identify the majority of students that need intervention, a large number of students would also be incorrectly identified as being in need of intervention.

Evaluating CBMs

L.S. Fuchs (2004) described a three-step program of research for establishing the tenability of CBMs. First, research is conducted to investigate the technical properties of the static score (i.e., student performance at one point in time). Next, the technical properties of the slope are evaluated in order to ensure that student growth is in fact associated with actual competence in the domain of interest. Finally, in the third step, research is conducted to evaluate the instructional utility of the measure. That is, the measure is able to provide information to practitioners that will assist them in making instructional decisions. The recommendations from L.S. Fuchs in establishing the tenability of CBMs address important issues relating to foundational validity evidence. However, like any other measures, CBMs must also be evaluated in terms of their reliability.

Reliability. Reliability refers to the consistency of measurements (Christ, Scullin, Tolbize, & Jiban, 2008). Within the classical test theory framework (CTT), reliability evidence can be provided through: internal consistency, test-retest, alternate-form, and inter-rater reliability (Christ et al., 2008). Internal consistency, often represented by Cronbach's α, is a measure of the degree to which a set of components (i.e., test items) are interrelated (Raykov & Marcoulides, 2011). On timed tests, where examinees are not

likely to respond to all of the items, internal consistency estimates lose their interpretability. Thus, they are often not appropriate for estimating reliability for CBMs (Christ et al., 2008). Test-retest methods assess the consistency of test scores across two test administrations. The correlation between the initial test score and the retest test score yields a reliability estimate, called a coefficient of stability (Raykov & Marcoulides, 2011). Since the intended purpose of CBMs is to measure student progress in specific academic domains, test-retest reliability estimates are not common in the literature. Alternate form or alternate form test-retest methods are when the alternate forms of a test are correlated to generate a coefficient of equivalence. These methods are better suited for CBM evaluation since they can be administered to students repeatedly without incurring the same level of practice effects inherent in basic test-retest methods. Coefficients of equivalence greater than .80 are generally desired, however, coefficients > .70 have been deemed acceptable by some (e.g., Salvia, Ysseldyke, & Bolt, 2006) for low-stakes situations such as progress monitoring (Raykov & Marcoulides, 2011). Inter-rater reliability is calculated when two separate raters score the same test; then divide the sum of agreements and disagreements by the number of agreements.

Limitations of Word Problem CBM Research

While previous studies have provided insight to the tenability of word problem solving measures as formative assessment tools in mathematics, research in this domain is still in its infancy. As a result, there are a number of limitations that have not been addressed in the word problem literature. First, most investigations of word problems have been limited to one-step arithmetic word problems with variations in semantic

structure (i.e., change, compare, combine). However, these types of problems are not reflective of those included in math curricula beyond children's first years in school (Fuchs, Fuchs, & Prentice, 2004). Furthermore, math reform efforts have called for the inclusion of more complex problem-solving tasks. That is, problems that include multiple steps and irrelevant information (Gersten, Beckmann, et al., 2009). Therefore, as instructional practices change to meet these demands, so must assessment practices.

Second, the word problem measures that appear in the literature have some limitations with regard to their technical properties. For example, the only reliability data that have been provided for fluency-based word problem measures has been evidence of internal consistency. While most of the internal consistency estimates have been adequate for low-stakes relative-type decisions (i.e., $\alpha > .80$), such as deciding which students are in need of intervention or which students are ready to be exited from interventions, internal consistency estimates are not appropriate for timed tests (Christ, Scullin, Tolbize, & Jiban, 2008). Only one study has provided test-retest reliability evidence (e.g., Fuchs, Fuchs, Karns, et al., 1999); in this instance, the reliability coefficients (.66-.76) did not reflect adequate stability. Therefore, there is not sufficient evidence at this point to suggest that word-problem measures could be used for universal screening or progress monitoring purposes.

While preliminary validity evidence has been presented in the form of concurrent or predictive correlations with various criterion measures, the usefulness of word-problem measures in identifying students that are at-risk in the area of math problem solving has only been examined in one previous investigation (e.g., Fuchs, Compton,

Fuchs, et al., 2011). In this investigation, the experimental word-problem screener had a significant but moderate correlation (r = .43) with the criterion measure. Therefore, there is reason to suspect that screeners that have stronger predictive relationships with criterion variables will demonstrate better classification accuracy. Several studies have shown that reading CBMs can be used to identify the students that are most at-risk for failing high stake assessments in reading (e.g., Compton, Fuchs, Fuchs, & Bryant, 2006; Deno et al., 2009), however, very little is known in regards to how well CBMs can fill this role in mathematics.

A third limitation of the word problem literature relates to administration procedures. Many of the studies have attempted to limit the contribution of reading skills by reading the problems to the students out loud, and giving students ample time to respond to the problems. However, these administration procedures are not in stride with procedures that are utilized in high-stakes testing situations. For example, students are expected to read and comprehend the math problems presented to them on state tests. Therefore, the inconsistency between these two administration formats provides serious limitations to the generalizability of results. Also in regard to administration practices, a true fluency-based word problem measure has yet to be evaluated. Although Jitendra and colleagues (2005) referred to their assessment as fluency-based, students were given ten minutes to complete eight problems. Whether or not such a long duration is really necessary to adequately predict later performance on criterion measures is unknown at this point.

A fourth limitation in the word problem literature is that there is a lack of information on the extent to which word problem measures can explain additional variance in math achievement beyond alternative measures that have been linked to math achievement in previous literature. For example, both computation and reading skills have both been linked to problem solving performance in previous literature (e.g., Fuchs, Fuchs, Compton, et al., 2006; Grimm, 2008). Despite the obvious relationships between these skills, however, most studies have not accounted for these variables as covariates in statistical analyses. I briefly review the literature on each of these covariates.

Computation. Algorithmic computation, which refers to adding, subtracting, multiplying, and dividing whole numbers, decimals, or fractions using algorithms or simple arithmetic, has been hypothesized as a facilitator of word problem skills (Fuchs, Fuchs, Compton, et al., 2006). Given the procedural nature of word problem tasks, it should be expected that students that struggle with procedural computation also struggle with word problems. Fuchs, Fuchs, Compton, et al. (2006) proposed a bottleneck hypothesis, which suggested that students that failed to master basic arithmetic or algorithms might not have the cognitive resources available to attend to procedural work.

Correlational studies (e.g., Thurber, Shinn, & Smolkowski, 2002) have also shown that computation and problem solving skills are highly related, even though they represent different constructs of mathematics. Moreover, the NMAP (2008) reported that conceptual understanding, computational fluency, and problem-solving skills were all mutually supportive; each supporting learning in the others. For these reasons, analyses of word problem skills should include computation skills.

Reading Comprehension. In addition to computation, reading skills play a crucial role in problem solving accuracy. Reading comprehension skills have been implicated in a number of studies as critical indicators of word problem solving skills. Jordan, Hanich, and Kaplan (2003) followed a cohort of students from 2nd grade through 3rd grade. In following these students through four waves of data collection, Jordan and colleagues found that students with deficits that were specific to mathematics outperformed students with combined deficits in reading and mathematics on word problem tasks. They also noted that this profile of combined difficulties remained constant across all of the time points.

In a longitudinal investigation that examined the role of reading comprehension on math skills, Grimm (2008) followed a cohort of students from 3rd grade through middle school. Grimm reported that students with higher reading comprehension scores in the 3rd grade showed more rapid growth in math problem solving and data interpretation skills across the study. Another notable finding was that reading comprehension made more of a contribution to student growth in math problem solving than any student demographic characteristics (i.e., SES, ethnicity), or prior math achievement.

Fuchs, Fuchs, Stuebing, Fletcher, Hamlett, and Lambert (2008) reported that word identification and language skills were critical variables in distinguishing between students that had combined deficits in reading and math, and students that had deficits that were specific to math. This finding has serious implications for the role of reading comprehension in word problem solving because the language composite used in the

study was formed through the combination of listening comprehension, expressive vocabulary, and grammatical closure tasks; all of which are highly related to reading comprehension.

Research Objectives

The main purpose of this study is to examine the utility of a fluency-based word problem measure in predicting math outcomes for third grade students. Of particular interest was the extent to which performance on the word problem solving measure was predictive of achievement outcomes on norm- and criterion-referenced measures of math problem solving ability. The performance of third grade students on these measures are of interest for three reasons. First, by third grade word problems have become a regular part of the math curriculum, and students have some familiarity with them. Second, by third grade, students generally possess the reading skills necessary to comprehend word problems. Finally, by third grade students are expected to be proficient with multi-digit, multi-step, addition and subtraction problems.

The proposed study will make a number of contributions to the math assessment literature. First, this investigation marks the beginning of a program of research that aspires to provide evidence on the technical features of a fluency-based measure of word problem solving. This study looks to address the first two steps of CBM evaluation research outlined in Fuchs (2004); examining the properties of the static score of a measure, and examining the properties of the slope. Second, the proposed study looks to provide consequential validity evidence for word problem measures, by examining their utility as screening measures. Third, this study will examine the role of covariates (i.e.,

computation, reading comprehension) in predicting math problem solving outcomes. Fourth, the suitability of two-minute samples of word problem solving performance will be evaluated. Finally, this study looks to examine the overall viability of a fluency-based word problem measure in a formative assessment process.

This study is guided by four research questions: (1) To what extent are scores from alternate forms of word problem CBM reliable and valid predictors of high stakes tests? (2) To what extent do word problem CBMs contribute significant variance to high stake test performance beyond the contribution of reading comprehension and calculation? (3) To what extent are word problem CBMs sensitive to differences in rate of improvement (ROI) between students at different levels of risk? (4) To what extent does word problem solving fluency performance and rate of improvement (ROI) discriminate between students at different levels of risk in math problem solving? Do measures of computation and reading comprehension improve discrimination?

Methods

Participants

The sample consisted of (N = 142) third grade students (72 males, 70 females) from southern California. Students were nested within 11 classrooms in two schools. The ethnic breakdown of the sample was: 61.97% White (n = 88), 11.97% Hispanic (n = 17), 8.45% African American (n = 12), 4.23% Asian (n = 6), and 13.38% that reported mixed ethnicity (n = 19). Of the 142 students, 6 (4.23%) were receiving special education services. The median percentage of students in classrooms that received free or reduced-price lunch (FRPL) in the study was 26%, indicating relatively low levels of poverty in

the student population. In the classification scheme used by Aud and colleagues (2010), schools with less than 25% of their student body receiving FRPL were classified as low-poverty schools. On the converse, schools with 75% or more children receiving FRPL were considered to be high-poverty (Aud et al., 2010). Participants were selected from a larger three-year longitudinal project that investigated the impact of cognitive strategy interventions on children with math difficulties. This study was funded by the U.S. Department of Education, Cognition and Student Learning in Special Education (USDE R324A090002), Institute of Education Sciences.

Selection for the larger study was based on returned informed parental consent, and fluid intelligence scores. Students were included in the current study if they had data recorded for the criterion measures and word problem solving fluency measures. A total of 251 third grade students were originally recruited for the purposes of this study. However, a total of 109 students were excluded from the study for the following reasons: 15 students were denied parental consent, 17 had fluid intelligence scores below the 16th percentile, and 77 had missing data.

Students with scores below the 16th percentile on the Raven Coloured Progressive Matrices test (Raven, 1976) were excluded from the study to prevent the confounding of specific math difficulties and more generalized intellectual deficits. Strong criterion validity between the Raven and WISC-R Full Scale score and California Achievement Test (*r*'s = .61 and .76, respectively) is documented in the technical manual. Also, Cotton, Kiely, Crewther, Thomson, Laycock, & Crewther (2005) reported internal consistency estimates ranging from .76-.88, and split-half reliability estimates ranging

from .81-.90, for children ages 6-11. Such levels of reliability have been reported to be sufficient for research purposes (Kamphaus, 2005).

An a priori power analysis determined that a sample of $N = 147$ would be necessary in order to uncover a medium effect (i.e., $R^2 = .10$) in a hierarchical regression model with four predictors at the recommended power level of .80. Furthermore, large effect sizes between word-problem measures and criterion measures of mathematics have been reported in past literature (e.g., $R^2 = .34$ in Jitendra, Scezniak, & Deatline-Buchman, 2005). Therefore, the current sample of $N = 142$ students was deemed sufficient to yield adequate power (.80) to detect the level of effect reported in the literature.

Procedures

As part of the larger investigation, students were administered a battery of assessments in the fall (pretest) and spring (posttest) of third grade in both group and individual formats by graduate student researchers. All pretest/posttests were counterbalanced for presentation order with alternate (form A or form B) versions randomly assigned. Within each classroom setting, students received twenty sessions of scripted intervention in small groups (i.e., 5 students or less); also from graduate student researchers or research assistants. During the intervention phase, the CBM measures were administered by classroom teachers every 6th school day.

Core classroom instruction. All of the study's participants interacted with their peers in their classrooms on tasks and activities related to the district-wide math curriculum. The core math instruction across conditions was the *enVisionMATH Learning Curriculum* (Pearson Publishers, 2009). The curriculum included visual

representations to show how quantities of a word problem were related, and general problem solving steps. The general problem solving steps in the teacher manual were to have children: (a) understand, (b) plan, (c) solve, and (d) look back. An independent evaluation (Resendez & Azin, 2009) following guidelines outlined in the What Works Clearing House Standards (U.S. Dept. of Education, 2008), indicated in random trials (teachers assigned randomly to treatment or control condition), that gains emerged in grades 2-4; effect sizes relative to control groups fell in the 0.20 range. A number of the curriculum's elements were also utilized in the study's treatments (e.g., find the key word). However, in contrast to the school district's required instruction, treatment conditions within the larger study directly focused on specific components of problem solving over consecutive sessions presented in a predetermined order. In addition, the lesson plans for the experimental condition focused directly on the propositional structure of word problems.

Predictor Variables

Word Problem Solving Fluency (WPSF). The word problem solving fluency (WPSF) CBMs were administered to groups of students by classroom teachers. Six alternate forms of the CBM were randomly assigned to student classrooms, and administered across six different time points (i.e., day 6, day 12, day 18, etc.). Students had two minutes to work on the CBM probes. Each of the measures had 12 word problems which included relevant and irrelevant propositions matched by complexity.

Similar to Jitendra, Scezniak, and Deatline-Buchman (2005), addition and subtraction word problems were sampled from commonly used third grade math

textbooks. Specifically, multi-step problems that had irrelevant information were included (see sample probes in Appendix A). This problem sampling approach was taken in order to obtain a collection of word problems that were more varied and complex than the arithmetic word problems that are frequently used in the literature. The word problems met the semantic criteria for change, compare, and combine problem types discussed in Carpenter & Moser (1984). The sentence structure, problem complexity, and format were the same across the six alternate forms, except for the substitution of names and numbers. To assist in controlling for possible order effects and unequal scaling (see Montague, Penfield, Ender, & Huang, 2010, for discussion of scale equivalence on these measures), one of three presentation orders was randomly assigned to each of the classrooms.

Based on the work of Leh and colleagues (2007), the six alternate forms were aggregated into three CBM measures to maximize reliability. Forms were combined in the following fashion: 1 with 2, 3 with 4, and 5 with 6. Combining the alternate forms this way ensured that the combined form would be representative of two consecutive WPSF forms, regardless of presentation order. Thus, a student's WPSF score will equal the sum of the number of problems correct from two consecutive WPSF alternate forms. Inter-scorer agreement was calculated for the WPSF measure by having two separate raters that received training in scoring the WPSF measure rate a sample of 20 randomly selected probes. These independent raters generated 100% agreement on raw scores.

As indicated previously, fluency-based measures are in line with high-stake testing situations where time impacts student performance. Thus, I will directly test

whether two-minute samples of word problem solving are adequate for the purposes of discriminating between students that are capable problem solvers, and those that are still building procedural fluency. As discussed in Christ, Johnson-Gros, and Hintze (2005), students with more developed skills are likely to complete more problems per unit of time, and thus earn higher scores. In the case of word-problem solving, students that must work out the entire algorithm for all problems will ultimately generate fewer correct solutions. Furthermore, 1-4 minute samples of math performance have been found to be sufficient for the purposes of guiding classroom instruction (Christ, Johnson-Gros, & Hintze, 2005).

Computation. The numerical operations subtest from the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) was individually administered in the fall (pretest) and the spring (posttest) to assess computation. Two forms of the test were counterbalanced across participants at pretest and posttest. The subtest requires the solving of paper-and-pencil computation problems that increase in difficulty. The subtest yields raw scores that range from 0-40, and standard scores that have a mean of 100, and standard deviation of 15. The technical manual for the WIAT reported an average split-half reliability coefficient of .85 for the numerical operations subtest. Test-retest reliability coefficients for the numerical operations subtest were, on average .86, across grade levels. Evidence of criterion validity was also provided for the numerical operations subtest in the WIAT technical manual. The numerical operations subtest had correlations of .68 and .77, with the calculations subtest from the Woodcock-

Johnson-Revised, and the arithmetic subtest from the Wide Range Achievement Test-Revised.

Reading Comprehension. The text comprehension subtest from the Test of Reading Comprehension-Fourth Edition (TORC; Brown, Hammill, & Weiderholt, 2009) was used to assess reading comprehension. In this task, students are asked to silently read a short passage, and subsequently answer five multiple choice questions related to the passage. The raw score is the total number of correctly answered questions for the subtest. The test also yields scale scores that have a mean of 10, and a standard deviation of 3. The technical manual reported criterion validity evidence for the text comprehension subtest, with correlations of .55 and .61 with the broad reading cluster of the Woodcock Johnson Tests of Achievement-Third Edition and the verbal comprehension subtest from the Wechsler Intelligence Scale for Children-Fourth Edition, respectively. In terms of reliability, the average coefficient alpha for the text comprehension subtest was .95, and the test-retest coefficient was .83 for the entire sample.

Criterion Measures

Comprehensive Math Abilities Test (CMAT; Hresko, Schlieve, Herron, Swain, & Sherbenou, 2003). The problem solving subtest measures students' abilities to translate problems stated in the text to math problems in order to obtain the ultimate solution. Items require the manipulation operations, combinations of operations, use of formulas, etc. The test yields raw scores and scaled scores with a mean of 10 and standard deviation of 3. The dependent variable will be the total number of correctly

answered questions. Reliability estimates reported in the technical manual were as follows: internal consistency = .90, test-retest = .92. Evidence of criterion validity was provided in the technical manual through correlations between the CMAT problem solving subtest and the problem solving subtest of the Stanford Achievement Test-Ninth Edition (r = .41), and the applications subtest of the Woodcock Johnson-Revised (r = .39).

KeyMath-Revised (KeyMath-R; Connolly, 1988). The problem solving subtest from the KeyMath exposes students to routine and non-routine math problems in multiple domains. Students are asked to answer: (1) routine 'textbook-like' word problems that involve all four mathematical operations; (2) non-routine problems where the solutions are not immediately apparent, and the student must describe a strategy for arriving at the solution; (3) non-routine problems that ask students to answer questions using any strategy they have. The test yields raw scores, and scale scores that have a mean of 10 and a standard deviation of 3. The dependent variable is the total number of questions answered correctly. The technical manual reports criterion validity coefficients between the problem solving subtest and the Comprehensive Test of Basic Skills (CTBS) mathematics concepts and total mathematics subtests (r's = .60 and .60, respectively). Split-half reliability coefficients for the problem solving subtest were .88 across grades, on average.

Test of Mathematical Abilities-2nd Edition (TOMA-2; Brown, Cronin, & McEntire, 1994). The story problems subtest consists of 25 math problems presented in story format. Students are asked to read the story and solve the problems. Problems are

arranged in order of difficulty (from easiest to most difficult). The dependent variable is the number of correctly answered questions (0-25). The test yields scale scores (mean = 10, standard deviation = 3) in addition to raw scores. The average internal consistency estimate for the story problems subtest reported in the technical manual was .89. Test-retest reliability coefficients were .85, on average, across age groups. The correlation between the story problems subtest and the KeyMath scores were .51.

Problem solving-experimental (STAR) measure. An experimental problem solving measure was developed drawing upon sample questions from the California STAR (California's Standardized Testing and Reporting measure) test for Grade 3. The STAR test is typically administered at the end of the year to evaluate student and school progress. For this experimental measure, items were provided on the California Department of Education website (2009). In contrast to the standardized target measures (CMAT and KeyMath), this task required the children to read the story problem and select the best answer. Sixteen story problems from the sample items on the California Department of Education website (2009) for Grade 3 were administered for both pretest and post-test and involved the child silently reading the question and then circling the correct answer from four possible choices. Two forms of the measure were created that varied only in names and numbers, and were counterbalanced across presentation order during pretest and post-test. Pretest and post-tests were scored with 1 point for each correct answer, with a total 16 points possible. Previous investigations (e.g., Sisco-Taylor, Fung, & Swanson, submitted) have reported adequate reliability (α = .83), and

moderate predictive validity (r = .46-.49) coefficients with the problem solving subtests from the KeyMath and CMAT.

Risk Status. Students were grouped in terms of risk status in order to address two of the research questions. There were two levels of risk: at-risk, and not-at-risk. Students were considered at-risk if they scored below the 25th percentile on the KeyMath or CMAT problem solving subtests. While this is an arbitrary definition for risk, the 25th percentile cut-off score on standardized achievement measures has been commonly used to identify children at-risk (e.g., Clarke, Nese, et al., 2011; Fuchs, Compton, Fuchs, et al., 2011). Since both of the criterion measures of problem solving yield scale scores, a scale score of 8 (consistent with the 25th percentile) was used as the cutoff. Therefore, students with scale scores less than 8 were considered at-risk, and students with scales scores greater than or equal to 8 were considered not-at-risk.

Results

1. To what extent are scores from alternate forms of WPSF reliable and valid predictors of high stakes tests?

To provide reliability evidence for the WPSF measure, an alternate form test-retest method was utilized. Since the forms were counterbalanced across participants to control for the effect of presentation order, reliability estimates were generated from a subset (n = 65) of the total sample. This way, reliability estimates were produced from students that had the same presentation order, and were not impacted by prior exposure to the WPSF probes. The formula below was used to estimate the alternate form test-retest reliability of the WPSF probes.

$$\rho_{x,\,traf} = Corr(X_{tA}, X_{rB}) \quad (1)$$

Where:

$\rho_{x,\,traf}$ = Reliability coefficient for test-retest alternate form

Corr = correlation

X_{tA} = Score on Form A

X_{rB} = Score on Form B

Alternate form test-retest reliability estimates for the aggregated forms and single forms are displayed in Table 1. While reliability estimates for the single forms ranged from *r* = .50-.72, estimates for the aggregated forms ranged from *r* = .72-.74.

Pearson correlations between WPSF and the criterion measures are displayed in Table 2. Based on Cohen's (1988) conventions, the predictive correlations (from winter to spring) obtained between WPSF and the various criterion measures ranged medium to large in terms of magnitude (*r*'s = .46-.60). The smallest correlation was observed between WPSF and the STAR experimental measure (*r* = .46), while the strongest correlations were observed between WPSF and the CMAT problem solving and TOMA story problems subtests (*r*'s = .60). The correlation between WPSF and the KeyMath problem solving subtest (*r* = .54) was large in magnitude, similar to those generated from the TOMA and CMAT. Intercorrelations between the criterion variables also ranged from small to large in terms of strength (*r*'s = .29-.69).

2. To what extent does WPSF contribute to high stake test performance beyond the contribution of reading comprehension and calculation?

This research question was addressed by evaluating the WPSF measure as both an indicator of performance level, and measure of growth in math problem solving. A forced entry hierarchical regression method was utilized, where computation and reading comprehension scores were entered into the model first; followed by students' initial WPSF scores; followed by the WPSF ROI across the three time points (aggregated CBM forms A, B, and C). The individual slope estimates were generated by using the individual regression slope extraction method discussed in Pfister, Schwarz, Carson, and Jancyzk (2013), where regression coefficients are calculated for each student and subsequently extracted from Excel 2010. The Excel software calculates an ordinary least squares regression coefficient utilizing the following formula:

$$\beta = \frac{\sum x_i y_i - \frac{1}{n}\Sigma x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\Sigma x_i)^2} \quad (2)$$

Parameter estimates for the hierarchical regression models are presented in Table 3. Four separate hierarchical regression models were utilized in addressing this research question (i.e., one for each criterion measure). The calculation and reading comprehension scores were entered into the regression model first in order to gauge the extent to which these measures are predictive of math problem solving outcomes. The WPSF scores were introduced in the second step in order to demonstrate the extent to which these measures accounted for unique variance in word problem solving, above and beyond calculation and reading comprehension skills. Lastly, the WPSF ROIs were entered in the third step in order to see if improvement in word problem solving, as measured by WPSF, added

unique variance in predicting problem solving outcomes above and beyond calculation, reading comprehension, and initial performance on WPSF. In addition to the previously described models, alternate models were run where the WPSF scores were entered in the model first in order to provide a direct comparison between the WPSF measure, and the combination of the calculation and reading comprehension measures in predicting math problem solving outcomes. Results will be reported below for each of the respective criterion measures.

CMAT Problem Solving Subtest. As shown in Table 3, the initial model, containing measures of calculation and reading comprehension, accounted for approximately 30% of the variance in the CMAT (Adjusted $R^{2 =}$ .29), $F(2, 139) = 29.74$, $p < .001$. The second model, which included students' initial WPSF scores, produced an R^2 change of approximately 10% ($\Delta R^2 = .10$); a statistically significant contribution, $F(1, 138) = 23.52$, $p < .001$. The WPSF ROI also made a unique contribution ($\Delta R^2 = .05$, $F(1, 137) = 13.60$, $p < .001$) to the model, when all other variables were parceled in the analysis. However, once the WPSF ROI was introduced in step 3, calculation ($\beta = .14$, $p = .08$) and reading comprehension ($\beta = .09$, $p = .25$) were no longer significant predictors within the model. The final model, consisting of calculation, reading comprehension, WPSF initial score, and the WPSF ROI, accounted for 45% of the variance in the CMAT problem solving subtest (adjusted $R^2 = .44$), $F(4, 137) = 28.67$, $p < .001$. The alternative model, where the initial WPSF scores were entered into the model first, accounted for approximately 35% of the variance in the CMAT (Adjusted $R^2 = .35$), $F(1, 140) = 76.99$,

$p < .001$, in the first step. This shows that the WPSF score accounted for more variance in the CMAT than the combination of calculation and reading comprehension scores.

TOMA Story Problems Subtest. As shown in Table 3, the initial model of calculation and reading comprehension accounted for approximately 25% of the variance in the TOMA story problems raw score (Adjusted $R^{2} = .24$), $F(2, 139) = 22.88$, $p < .001$. The WPSF scores explained additional significant variance beyond the original model, $\Delta R^2 = .14$, $F(1, 138) = 32.42$, $p < .001$. However, once the WPSF scores were introduced in step 2, calculation ($\beta = -.01$, $p = .95$) was no longer a significant predictor. The WPSF ROI did not make a unique contribution ($\Delta R^2 = .00$, $F(1, 137) = .71$, $p = .40$) to the model, after accounting for all other variables. The final model, consisting of calculation, reading comprehension, WPSF initial score, and the WPSF ROI, accounted for 39% of the variance in the TOMA story problems subtest (adjusted $R^2 = .38$), $F(4, 137) = 22.26$, $p < .001$. The alternative model, where the initial WPSF scores were entered into the model first, accounted for approximately 36% of the variance in the TOMA (Adjusted $R^2 = .36$), $F(1, 140) = 78.71$, $p < .001$, in the first step. This shows that the WPSF score accounted for more variance in the TOMA than the combination of calculation and reading comprehension scores, which accounted for approximately 25% of the variance.

KeyMath Problem Solving Subtest. The initial model accounted for approximately 34% of the variance in the KeyMath problem solving raw score (Adjusted $R^{2} = .33$), $F(2, 139) = 35.10$, $p < .001$. The WPSF scores explained additional significant variance beyond the original model, $\Delta R^2 = .05$, $F(1, 138) = 11.16$, $p = .001$. Moreover, The WPSF ROI made a unique contribution ($\Delta R^2 = .12$, $F(1, 137) = 33.51$, $p < .001$) to

the model, after accounting for all other variables. However, once the WPSF ROI was introduced in the third step of the model, reading comprehension (β = .04, p = .63) was no longer significant as a predictor. The final model, consisting of calculation, reading comprehension, WPSF initial score, and the WPSF ROI, accounted for 51% of the variance in the KeyMath problem solving subtest (adjusted R^2 = .49), F(4, 137) = 35.09, p < .001. The alternative model, where the initial WPSF scores were entered into the model first, accounted for approximately 29% of the variance in the KeyMath (Adjusted R^2 = .29), F(1, 140) = 57.46, p < .001, in the first step. This shows that the WPSF score accounted for slightly less variance in the KeyMath than the combination of calculation and reading comprehension scores, which accounted for approximately 34% of the variance in the original model.

STAR Experimental Problem Solving Measure. The initial model accounted for approximately 14% of the variance in the STAR measure (Adjusted $R^{2 =}$ .13), F(2, 139) = 11.14, p < .001. Reading comprehension (β = .14, p = .10) did not emerge as a significant predictor in the original model, however. When added to the model, the WPSF initial score contributed additional unique variance in predicting the STAR raw score, ΔR^2 = .09, F(1, 138) = 15.30, p < .001. However, the calculation (β = .14, p = .14) measure was no longer a significant predictor of the STAR problem solving test once the WPSF initial score was introduced to the model. The WPSF ROI made a unique contribution (ΔR^2 = .03, F(1, 137) = 5.60, p < .05) to the model, when all other variables were parceled in the analysis. The final model, consisting of calculation, reading comprehension, WPSF initial score, and the WPSF ROI, accounted for 26% of the

variance in the STAR math measure (adjusted R^2 = .23), $F(4, 137) = 11.70$, $p < .001$. The alternative model, where the initial WPSF scores were entered into the model first, accounted for approximately 21% of the variance in the STAR (Adjusted R^2 = .21), $F(1, 140) = 37.43$, $p < .001$, in the first step. This shows that the WPSF score accounted for more variance in the STAR than the combination of calculation and reading comprehension scores, which accounted for approximately 14% of the variance in the original model.

3. To what extent is WPSF sensitive to growth? Are there differences in rate of improvement (ROI) between students at different levels of risk?

The mean slope for the total sample (N = 142) of students across the three time points was M = .91 (SD = 1.22). Since the three time points covered a span of 12 weeks, with approximately four weeks between each time point, the weekly ROI was calculated by dividing the mean slope estimates by four. The weekly ROI for the total sample was M = .23; at-risk students had a weekly ROI of M = .17, while not at-risk students had a weekly ROI of M = .25.

In order to address this research question further, a mixed 2 (risk group) x 3 (time) ANOVA model with repeated measures on the last factor was utilized. Included within the model was the within-subjects factor *time*, which had three levels (times 1, 2, and 3); and the between-subjects factor *risk status*, which had two levels (at-risk, and not at-risk). Means and SD's for each of the respective risk groups are presented in Table 4.

Significant main effects were observed for risk group ($F(1, 140) = 340.22$, $p < .001$), and time (Wilks' Λ = .70, $F(2, 139) = 29.96$, $p < .001$). Effect size estimates for the

between-subjects factor risk group (partial η^2 = .31), and the within-subjects factor time (partial η^2 = .30) were both large in magnitude according to Cohen's (1988) conventions. As anticipated, a significant group effect emerged in favor of the not at-risk group (F(1, 140) = 340.22, p < .001). The mean initial WPSF score for the not at-risk group was M = 5.56, while the mean score for the at-risk group was M = 1.74. This pattern remained consistent across the three time points as there were also significant differences in WPSF levels at times 2 & 3.

The pooled mean raw scores grew across the three time points, increasing from M = 4.56 at time 1, to M = 6.30 at time 2, and M = 6.38 at time 3. A post-hoc Tukey test revealed that there were significant differences in the raw scores between times 1 & 2 (p < .001), and times 1 & 3 (p < .001); mean score differences between times 2 & 3 (p = .95) were not significant. Moreover, in addition to the observed linear change (F(1, 140) = 50.71, p < .001, partial η^2 = .27) across the three time points, evidence of a quadratic growth pattern (F(1, 140) = 18.89, p < .001, partial η^2 = .12) also emerged in the data. As shown in Figure 1, WPSF scores increased by an average of 1.67 points between times 1 & 2, and then leveled off between times 2 & 3, where they grew by an average of 0.12 points.

The time-by-group interaction tested within the model was not significant (Wilks' Λ = .99, F(2, 139) = 1.02, p = .36), indicating that students in the respective risk groups did not grow at significantly different rates in word problem solving during the study. Therefore, while significant growth on the WPSF measure was observed across the 12 weeks, these data do not suggest that the growth rates of at-risk students (M = .67, SD =

1.39) were significantly different from those of students that were not at-risk (M = .99, SD = 1.15).

4. **Does word problem solving fluency performance and rate of improvement (ROI) discriminate between students at different levels of risk in math problem solving? Do measures of computation and reading comprehension improve discrimination?**

This question was addressed through logistic regression, and receiver operating curve (ROC) analyses. In addition to addressing the research question, obtaining estimates of diagnostic efficiency for WPSF was also of interest. Therefore, in addition to calculating odds ratios for group membership through the logistic regression analysis, the following diagnostic efficiency statistics were considered: area under the curve (AUC), sensitivity, specificity, and classification accuracy.

The AUC serves as a simple summary of overall accuracy (Hanley & McNeil, 1982). It refers to the proportion of randomly chosen pairs of students for which the screening assessment (WPSF) accurately classifies as at/below, or above the 25th percentile cutoff for risk (Clarke, Nese, et al., 2011). While there are no standard conventions for interpreting the magnitude of AUC values, some guidelines or rules of thumb have been presented in past literature. For example, Dolan and Doyle (2000) reported that an AUC value greater than 0.75 was similar to a Cohen's *d* effect size greater than 0.5 (moderate effect). More recently, Fuchs, Fuchs, Compton, Bryant, Hamlett, and Seethaler (2007) presented the following categorical descriptors for interpreting AUC values in the math screening literature: AUC values less than .70

indicate a poor predictive model; .70-.79 indicate fair prediction; .80-.89 indicate good prediction; and values equal to or greater than .90 indicate excellent prediction. Sensitivity is the proportion of students that were correctly identified by WPSF as being at-risk. Specificity is the proportion of students accurately predicted by WPSF to be not at-risk. Classification accuracy is the number of accurate predictions (the sum of true positives and true negatives), over the total number of predictions (the sum of true positives, false positives, true negatives, and false negatives). In addition to the aforementioned diagnostic efficiency statistics, cut-points were generated from the ROC analysis that optimized levels of sensitivity and specificity. Results for the logistic regression, and ROC analyses will be reported separately below.

Logistic Regression. The dependent variable in the logistic regression analysis was risk status (at-risk, not at-risk). Recall that students were considered to be at-risk if they had a scale score of less than eight on either the CMAT or KeyMath problem solving subtests. Thus, students that were considered at-risk were coded "0", while students that were not at-risk were coded "1". The independent variables (WPSF initial score, WPSF ROI, WIAT, and TORC) were entered into the regression model in three blocks: First, a model was tested with no predictors to estimate the base rate for classification accuracy (the classification accuracy generated by assuming all students were not at-risk for math problem solving difficulties); next, the WPSF raw score and WPSF ROI were entered in the model; and finally, the raw scores from the WIAT numerical operations and TORC text comprehension subtests were entered in model.

The rationale behind entering the independent variables in the specified order was to gauge the extent to which WPSF data predicted risk status in math problem solving, independent of any other information (i.e., calculation skills, reading comprehension skills). The reading comprehension and calculation scores were entered into the model last in order to gauge the extent to which they contribute additional information to predicting risk status, beyond the WPSF data. This way, there will be a clear illustration of how much each additional source of information improves screening accuracy. Output from the logistic regression model are presented in Table 5.

Given no predictors, the logistic regression model was able to correctly classify approximately 74% of students (all students were classified as not at-risk). Once the WPSF raw scores and ROIs were entered in the model in the second step, classification accuracy increased to approximately 82% overall; 60% of the at-risk students were correctly classified (22 true positives, 15 false negatives), and 90% of not at-risk students were correctly classified (94 true negatives, 11 false positives). Furthermore, both the WPSF score (β = .74, p < .001) and WPSF ROI (β = .66, p < .01) were significant predictors in the model. The odds ratios for the WPSF score and ROI were 2.10 and 1.94, respectively. In order to provide a clear interpretation of the magnitude of the observed effect, the odds ratios were converted into effect sizes using the formula presented in Hasselblad and Hedges (1995, p. 170):

$$d_{HH} = L_{OR} \frac{\sqrt{3}}{\pi} \qquad (3)$$

Where L_{OR} is the natural logarithm of the odds ratio (OR), and $\pi = 3.142$. Using formula 3, the odds ratios for the WPSF score and ROI translated into significant small effect effects (*d*'s = .41, and .36, respectively), following Cohen's (1988) conventions.

The addition of calculation and reading comprehension scores in the third block did not improve the classification accuracy of the model. Moreover, neither the WIAT ($\beta = .02$, $p > .05$) nor the TORC ($\beta = .02$, $p > .05$) emerged as significant predictors in the model.

ROC Analysis. As illustrated in Figure 2, the WPSF raw scores generated an area under the curve (AUC) of .83. This means that when students were identified as members of the not at-risk group as a function of their initial WPSF level, they yielded scores that were greater than students that were identified as at-risk 83% of the time. Based on the conventions that have been outlined previously in the literature (e.g., Dolan & Doyle, 2000; Fuchs et al., 2007), an AUC value of this magnitude is consistent with a moderate effect (i.e., $d > .50$), and reflects a "good-predicting" model. In order to identify cut scores that optimized sensitivity and specificity, cut score decision rules outlined in Silberglitt and Hintze (2005) were utilized:

> (1) Determine the cut score(s) that yield at least 0.7 for sensitivity and specificity; (2) if possible, increase sensitivity from this point, continuing upward while still maintaining a specificity of 0.7, stopping if sensitivity exceeds 0.8; (3) if sensitivity exceeds 0.8 and specificity can still be increased, continue to maximize specificity (while maintaining sensitivity of 0.8); and (4) if both sensitivity and specificity exceed 0.8, repeat steps 2 and 3, using 0.9 as the next cutoff. (p. 316)

Cut scores and corresponding rates of sensitivity and specificity are presented in Table 6. As shown in Table 6, only one cut score (a raw score of 3.5) met the criteria. This score was consistent with a sensitivity rate of .81, and a specificity rate of .73. The next best cut score was 2.5, which increased sensitivity (.88) but sacrificed specificity (.54).

Discussion

This study endeavored to provide empirical support for the use of fluency-based word problem solving measures as formative assessment tools in early elementary education. In doing so, this investigation provided evidence on the technical features of a curriculum-based word problem solving fluency measure. Following the recommendations for evaluating CBM tools outlined in Fuchs (2004), this study looked to: (1) examine the technical properties of the WPSF static score; and (2) examine the technical properties of the slope. In addition to providing evidence on the technical properties of the WPSF measure, this study also attempted to provide consequential validity evidence on the WPSF measure by evaluating its' diagnostic efficiency as an academic screener. One final goal of this study was to examine the role of covariates commonly associated with math problem solving (i.e., calculation, and reading comprehension) in predicting student outcomes. The findings from this investigation will be discussed below as they relate to each of the respective research questions.

1. To what extent are scores from alternate forms of WPSF reliable and valid predictors of high stakes tests?

The majority of the reliability evidence reported for word problem measures has been on their internal consistency. These estimates have been reported from the .70-.80

range (e.g., Jitendra, Sczesniak, & Deatline-Buchman, 2005). However, the time-based nature of CBM tasks in general, and in this case WPSF, make estimates of internal consistency inadequate for establishing reliability evidence (Raykov & Marcoulides, 2011; Christ, Johnson-Gros, & Hintze, 2005). Instead, alternate form test-retest reliability estimates were generated in this study. As anticipated, the coefficients of equivalence for the single forms of WPSF were rather low, and inconsistent (*r*'s = .50-.72). This was consistent with the findings from Leh et al. (2007), where they opted to aggregate their CBM forms to enhance reliability estimates.

Similar to Leh and colleagues (2007), the aggregated forms in this study produced far more reliable estimates of word problem solving (*r*'s = .72-74). While these estimates from the aggregated forms were not above the recommended reliability threshold of $r = .80$ discussed by Gersten, Beckmann, and colleagues (2009), they did surpass the basal level of reliability evidence ($r \geq .70$) recommended for low-stakes, relative decision-making contexts (Christ, Johnson-Gros, & Hintze, 2005). Furthermore, these reliability estimates for WPSF are comparable, and in some instances superior, to contemporary CBM measures discussed in the literature, yet have the distinction of being obtained in a far more time-efficient manner. For example, Fuchs, Fuchs, and Courey (2005) reported alternate form test-retest reliability coefficients of $r = .66$-.76, and Jitendra and colleagues (2005) reported internal consistency estimates of $r = .76$-.83. While administration time was not reported for Real-Life Math CBM, Jitendra et al. (2005) allowed for 10 minutes of administration time.

The results from this investigation suggest that 10 minutes may not be necessary to garner a reliable estimate of word problem solving ability. While the two minute work sample of word problem solving ability obtained in this study did not produce reliability estimates that were consistent with the r = .80 reliability threshold, it is certainly possible that a reliable estimate could be attained in five minutes. This is a question that could be addressed in future research in this area.

The predictive correlations (winter to spring) between WPSF and the criterion measures ranged from r = .46-.60, small to moderate in terms of strength. These correlations are comparable to those reported in the literature for contemporary CBM measures. For example, Jitendra and colleagues (2005) reported concurrent validity coefficients ranging from .64-.71. They are also comparable to the correlation coefficients generated between the respective criterion measures used in this study (r's = .29-.69), which were produced in a concurrent fashion. These differences in correlation between the criterion measures were likely a product of differences in the administration and response formats for the respective tests. For example, while problems were read aloud to students for the CMAT and KeyMath subtests, students had to read the problems from the TOMA and STAR subtests on their own. Also, while the CMAT, KeyMath, and TOMA all had an open response format (i.e., students had to generate their own answers), the STAR test was in a multiple choice format, where the student had to select the best answer from a field of four possible answers. The criterion-referenced tests (i.e., CMAT, KeyMath, and TOMA) that required students to generate responses are likely more accurate representations of students' math abilities in the area of problem solving since

the likelihood of a student guessing the correct answer is greatly diminished in this format.

The correlation coefficients between WPSF and the TOMA story problems subtest, and the CMAT problem solving subtest were the highest (both *r*'s = .60). These criterion measures were also the most similar to the WPSF measure in response and administration format.

While the relationships between the CBM measures and the criterion measures are not as strong as those often observed in the CBM reading literature, they did meet the level of predictive validity evidence (r = .60 in a given school year) recommended by Gersten, Beckmann, et al. (2009) for math screeners. This finding has strong implications for practice, since it shows that the strength of association between word problem tasks and problem solving outcome measures was not compromised by the shorter administration time used in this study. Recall that students had eight minutes to work on the word problem CBM used in Jitendra et al. (2005), and students in this study had only two minutes to work on the measure.

2. To what extent does WPSF contribute to high stake test performance beyond the contribution of reading comprehension and calculation?

The hierarchical regression models revealed that the full battery of assessments, which included measures of calculation, reading comprehension, and word problem solving (level and slope), accounted for between 26-51% of the variance in criterion measures of problem solving at the end of the school year. The battery accounted for over one quarter of the variance in the STAR measure, and over one half of the variance in the

KeyMath problem solving subtest, all strong effects by Cohen's (1988) standards. In addition, the initial score, and rate of improvement (ROI) on the WPSF measures added unique variance in the prediction of math achievement at the end of the school year for all of the outcome measures, with the exception of the TOMA. Thus, even after accounting for students' calculation, and reading comprehension skills at the beginning of the school year, students' initial performances on WPSF, and their rates of improvement on the task provided useful information in predicting end of year outcomes in math problem solving.

One particularly interesting finding was the difference in the importance of calculation and reading comprehension in predicting the outcomes for the respective criterion measures. While calculation was an important predictor for the KeyMath (β = .26, $p < .001$), it was not a significant predictor for any of the other criterion variables when accounting for WPSF performance. Similarly, reading comprehension was an important predictor for the TOMA ($\beta = .19$, $p < .05$), but was not a significant predictor for any of the other criterion variables after accounting for WPSF performance. This is an important finding, as it suggests that WPSF is accounting for variance in problem solving ability that cannot be explained by reading comprehension, and/or calculation skills.

It also lends further support to recommendations from Gersten and colleagues (2009), which called for the implementation of universal screening, and other forms of formative assessment to identify students with problem solving difficulties. If students that have exclusive problem solving difficulties can be identified early on in the school year, they can then be placed in interventions that are specifically tailored to remediate those deficits. Interventions that have focused on teaching the underlying structures of

word problems have been shown to be particularly effective in improving problem solving competencies (e.g., Fuchs, Fuchs, Prentice, et al., 2003; Fuchs, Fuchs, Prentice, Hamlett, Finelli, & Courey, 2004; Jitendra, Griffin, McGoey, Gardill, Bhat, & Riley, 1998; Xin, Jitendra, & Deatline-Buchman, 2005).

3. **To what extent is WPSF sensitive to growth? Are there differences in rate of improvement (ROI) between students at different levels of risk?**

The repeated-measures mixed ANOVA model produced a significant, large effect (partial η^2 = .30) for time across the three time points in the study, indicating that the WPSF measure is indeed capable of providing evidence of growth. The weekly growth rate for the total sample was .23; .17 and .25 for students that were at-risk and not at-risk, respectfully. While lower than ROIs that are typically reported in the reading literature, this weekly ROI was comparable to contemporary CBMs of problem solving, and other math measures alike. For instance, Leh and colleagues (2007) reported a weekly growth rate of .24 for their total sample, and Fuchs, Fuchs, and Courey (2005) reported ROIs of .11, .28, and .54 for low-, average-, and high-achieving students in their sample.

While the WPSF rate of improvement across the three time points emerged as a significant predictor of performance on both criterion measures, and contributed unique variance in predicting future math scores on those respective measures, there were no statistically significant differences between the slopes of at-risk students and students that were not at-risk in this study. Attempts to establish discriminant validity for problem solving CBM slopes have had mixed results in the literature. For example, Leh and colleagues (2007) did not find significant differences in weekly ROIs between low- and

average-achieving students. However, Fuchs, Fuchs, and Courey (2005) were able to show evidence of discriminant validity in CBM slopes between high-, average-, and low-achieving students.

One explanation for not finding significant differences in slope between the two risk groups is that there was a lack of precision in the slope estimates. The standard deviations for the slope estimates of the respective groups were rather large in comparison to the means (*SD*s = 0.29-0.35), indicating substantial variance associated with these slope estimates. This outcome is likely because only three data points were used to estimate slope in this study. Christ (2006) demonstrated that the standard error of the estimates (SEEs) decreased continuously when more data points were considered in calculating rates of improvement.

Another factor that may have impacted the WPSF measure's ability to measure growth in problem solving was the methodology used for scoring it. The total number of correct responses was used as the outcome variable for WPSF in this study. However, others (e.g., Foegen, Olson, & Impecoven-Lind, 2008) have used alternative scoring methodologies that reward students for accurately completing parts of the problem solving process (i.e., identifying correct numbers, choosing the correct algorithmic, etc.). Rewarding students for properly executing parts of the problem solving process would likely increase the range of possible scores, and be more sensitive to growth. Future research in this area should examine this possibility further.

Despite these issues with measurement error, however, the WPSF slope did emerge as a significant predictor in the various regression analyses conducted in this

study. This is an important finding for two reasons. First, it demonstrates that growth in problem solving skills can in fact be measured in an efficient manner. Second, growth rates can yield valuable information as it pertains to end of year success in math. Thus, as teachers make the shift to incorporating more activities related to math problem solving in their lesson plans, WPSF and measures alike hold promise for providing a means to evaluate the effectiveness of the instruction.

4. **Does WPSF performance and rate of improvement (ROI) discriminate between students at different levels of risk in math problem solving? Do measures of computation and reading comprehension improve discrimination?**

The WPSF measure demonstrated the ability to distinguish between students at different levels of risk, producing an AUC of .83. To provide a source of comparison, an AUC of .83 is consistent with a moderate effect size (i.e., $d > .50$). Moreover, when assigning students to risk groups based on WPSF scores, the classification accuracy was as high as .82, which is far greater than chance, and greater than the rate produced by the baseline model which assigned all students to the not at-risk group given the low frequency was at-risk students in the student population. The diagnostic efficiency statistics that were obtained by using optimal cut scores generated in the ROC analysis were also very promising. When selecting a cut score that holds the level of sensitivity at .81 (limiting the number of students that are misidentified as not at-risk), a respectable level of specificity was still attainable. This means that WPSF holds the capacity to correctly identify at-risk students without greatly over-identifying students as at-risk. The AUC, and other markers of diagnostic efficiency were also comparable to those

generated by other math CBMs in past research. For example, Clarke, Nese, et al. (2011) reported an AUC of .83, and corresponding levels of sensitivity and specificity of .70 and .83, respectively, for their easyCBM number sense measure. No other studies that evaluated the screening properties of word problem solving CBMs were identified in the literature review, so these findings are believed to be the first source of empirical data on the technical adequacy of word problem CBMs as screening tools for students in early elementary school.

One surprising finding that emerged in this study was that the calculation, and reading comprehension measures did not enhance the screening accuracy beyond WPSF performance and ROI. This speaks to the potency of WPSF as a screener, and highlights the implications that the measure could have for informing math instruction. If a measure can correctly distinguish between students who will meet an end of year criterion, and those that will not, it has the capacity to serve as an agent for selecting students for intervention. The diagnostic efficiency statistics that were generated in this study were actually comparable to those reported in the reading CBM literature (e.g., Compton, Fuchs, Fuchs, & Bryant, 2006; Deno et al., 2009), and these measures are already used on a fairly consistent basis throughout the country to inform reading instruction. Findings from this investigation provide preliminary evidence that math CBMs, and in particular, WPSF, have the potential to be used in a similar fashion as reading CBMs; they can be used for the purposes of identifying the students that present the greatest need for math intervention.

In this study, the 25th percentile was used as a cutoff for risk to identify students scoring in the lowest quartile on the criterion measures. This is an arbitrary cutoff, and the precision of the CBM measures can be evaluated using a number of different cutoffs. For example, Clarke, Nese, and colleagues (2011) experimented with cut points at the 10^{th}, 25^{th}, and 40^{th} percentile. From a practical aspect, the selection of students for math intervention in a given school will depend largely on the resources available at that school. For example, it may be of benefit to identify students scoring below the 15th percentile since the resources available may only allow one to serve a very small number of students.

Limitations

This study has two main limitations. First, only word problem solving accuracy was assessed by the CBM measure. The addition of other problems such as calculation (e.g., addition, subtraction) may be more representative of what is taught in schools and on high-stakes tests. However, because of the limited research on word problem solving CBM measures, one of the purposes of this study was to try to determine whether word problem solving CBM predicted word problem solving performance and the California STAR high-stakes test beyond that of calculation skills.

A second limitation is the 2-minute time limit that students had to complete the CBM measure. This may have assessed other areas that are related to word problem solving, including reading fluency, processing speed, and working memory (e.g., Andersson, 2007; Swanson & Beebe-Frankenberger, 2004; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008) rather than word problem solving ability. That is, students who have

better reading comprehension, faster phonological processing speed, and/or more working memory capacity may be able to answer the questions at a quicker pace, and thus answer more questions. However, one advantage of the time limit is that it can be easily implemented by teachers because it takes very little time to administer.

Conclusions and Implications

Large-scale reviews of the math literature (e.g., Gersten, Chard, et al., 2009; NMAP, 2008) have highlighted the need to utilize formative assessment practices in schools to improve math education. Formative assessment practices have been most effective when teachers use performance assessments to evaluate specific academic skills, and subsequently use those data to make instructional changes; effects are strengthened further when guidance is given to teachers on using assessment data to make instructional changes (Gersten, Chard, et al., 2009). Problem solving CBMs hold the promise of assisting in this process because they can provide low-inference information to teachers on students' problem solving skills, and be used in a repeated fashion.

This study examined the extent to which problem solving CBMs predict math achievement, specifically criterion-referenced measures of problem solving, and the California STAR standardized test, beyond that of traditional measures such as calculation, and reading comprehension. This investigation uncovered that WPSF accounted for approximately one-quarter of the variance in the STAR test, and up to one-half of the variance in the problem solving criterion measures. Predictive correlations with the STAR math test, and problem solving composite ranged from .46-.60 (similar to

correlations reported in other studies investigating word problem tasks (e.g., Fuchs, Compton, et al. 2012; Fuchs, Compton, Fuchs, et al., 2011; Jitendra, Scezniak, & Deatline-Buchman, 2005), providing early evidence of predictive validity for problem solving CBM. It must be noted that this is an initial step in the validation process, and that further research will be necessary to provide more empirical support for the psychometric properties of WPSF, and other word problem solving CBMs.

Findings from this study also show that problem solving CBMs hold the potential to be used as screeners for math difficulties. Results from the logistic regression and ROC analyses have strong implications for educational decision-making with regard to math instruction. The overall AUC of .83, and levels of sensitivity and specificity that were generated when using optimal cut-points, suggest that WPSF can aid educators in determining which students are in need of intervention. Since Gersten, Beckmann, et al. (2009), and others have made the recommendation of providing interventions targeting word problem solving during the early elementary years, it would benefit teachers to know which students need those interventions most.

Despite the limitations of the 2-minute administration time limit addressed in the previous section, findings from the current study suggest that 8-10 minute samples of word problem solving may not be necessary for the purposes of screening. The problem solving CBM was able to distinguish between students that were at-risk for problem solving difficulties, and students that were at relatively low-risk of having problem solving difficulties; all within a two-minute timeframe. Future research in this area might examine how much time is necessary to obtain an adequate sample of word problem

solving skills for the purposes screening, and progress monitoring. Measures that require less administration time also take less instructional time away from teachers; they are therefore more likely to be accepted by teachers. This aspect of social validity should also be considered when designing formative assessment tools.

While the WPSF ROI was a useful indicator in the regression models predicting scores on the problem solving outcome measures, the weekly ROIs were very small, and thus not very sensitive to change. This lack of sensitivity to change does not lend itself well to weekly or bi-weekly progress monitoring. Further, the absence of significant differences in ROI between the respective risk groups did not lend itself to establishing construct validity for the WPSF ROI. Future research should be directed toward evaluating the technical properties of the slope for WPSF while using different scoring methodologies. In the current study, the number of correct solutions was used as the outcome variable for WPSF. However, others (e.g., Foegan, Olson, & Impecoven-Lind, 2010) have used alternative scoring methodologies that reward students for correctly executing various steps within the problem solving process. Using this type of approach would likely increase the range of possible scores, and thus make WPSF more sensitive to changes in students' problem solving abilities.

Finally, the findings suggest that problem solving CBMs have additional predictive power, beyond that of traditional and often-used measures of math achievement. As discussed earlier, MMs target only isolated math skills, and thus do not provide an accurate assessment of students' overall progress in mathematics. As SBMs, problem solving CBMs tap into a broader range of math skills, and thus provide a better

indication of student progress. Unfortunately, a true general outcome measure (GOM) of math achievement, such as those boasted in the reading CBM literature, has yet to emerge in the math CBM literature. Therefore, at this juncture, the the use of multiple skill-based measures is recommended when assessing student progress in mathematics. This could be done by using WPSF or another type of word problem CBM, along with a computation, and/or concept and application measure.

References

Andersson, U. (2007). The contribution of working memory to children's mathematical word problem solving. *Applied Cognitive Psychology, 21*, 1201-1216. doi:10.1002/acp.1317

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M., …& Drake, L. (2010). *The Condition of Education 2010* (NCES 2010-028). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Brown, V. L., Cronin, M. E., & McEntire, E. (1994). *Test of Mathematical Abilities*. Austin, TX: PRO-ED.

Brown, V. L., Hammill, D., & Weiderholt, L. (1995). *Test of Reading Comprehension.* Austin, TX, PRO-ED.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education, 15*(3), 179-202. Retrieved from: http://www.jstor.org/stable/748348

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128-133.

Christ, T.J., Johnson-Gros, K.N., & Hintze, J.M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools, 42*(6), 615-622.

Christ, T.J., Scullin, S., Tolbize, A., & Jiban, C.L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention, 33*(4), 198-205. doi: 10.1177/1534508407313480

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal, G., Kame'enui, E. J., & Baker, S. K. (2011). Classification accuracy of easyCBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention, 36*(4), 243-255. doi: 10.1177/1534508411414153

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Elbraum.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394-409. doi: 10.1037/0022-0663.98.2.394

Connolly, A. J. (1998). *KeyMath (revised/normative update).* Circle Pines, MN: American Guidance.

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Coloured Progressive Matrices for primary school aged children from Victoria, Australia. *Personality and individual differences, 39*(3), 647-659.

DeJarnette, N.K. (2012). America's children: Providing early exposure to STEM (science, technology, engineering and math) intitiatives. *Education, 133*(1), 77-84.

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

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*(1), 44-55. doi: 10.1002/pits.20353

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities, 39*(5), 447-466.

Dolan, M., & Doyle, M. (2000). Violence risk prediction: Clinical and actuarial measures and the role of the psychopathy checklist. British Journal of Psychiatry, 177, 303-318. doi: 10.1192/bjp.177.4.303

Fletcher, J.M., Lyon, G.R., Fuchs, L.S., & Barnes, M.A. (2007). *Learning disabilities: From identification to intervention*. New York: The Guilford Press.

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. doi: 10.1177/00224669070410020101

Foegen, A., Olson, J. R., & Impecoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assessment for Effective Intervention, 33*(4), 240-249. doi: 10.1177/1534508407313489.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.

Fuchs, L.S., Compton, D.L., Fuchs, D., Hollenbeck, C.L., Hamlett, C.L., & Seethaler, P.M. (2011). Two-stage screening for math problem-solving difficulty using dynamic assessment of algebraic learning. *Journal of Learning Disabilities, 44*(4), 372-380. doi: 10.1177/0022219411407867

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*(3), 493.

Fuchs, L. S., Compton, D. L., Fuchs, D., Powell, S. R., Schumacher, R. F., Hamlett, C. L., ... & Vukovic, R. K. (2012). Contributions of domain-general cognitive resources and different forms of arithmetic development to pre-algebraic knowledge. *Developmental Psychology, 48*(5), 1315. doi: 10.1037/a0027475

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlet, C., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*, 311-330. doi: 10.1177/001440290707300303

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., ... & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*(1), 29.

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention, 30*(2), 33-46. doi: 10.1177/073724770503000204

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology*, *100*(3), 491.

Fuchs, L.S., Fuchs, D., Hamlett, C.L., & Stecker, P.M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*, 6–22.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kubek, P., & Stecker, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Assessment for Effective Intervention*, *19*(4), 23-49.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., & Katzaroff, M. (1999). Mathematics performance assessment in the classroom: Effects on teacher planning and student problem solving. *American Educational Research Journal*, *36*(3), 609-646.

Fuchs, L.S., Fuchs, D., & Prentice, K. (2004). Responsiveness to mathematical problem-solving instruction: Comparing students at risk of mathematics disability with and without risk of reading disability. *Journal of Learning Disabilities*, *37*(4), 293-306.

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., ... & Jancek, D. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of educational psychology*,*95*(2), 293-305.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*,*96*(4), 635-647.

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology, 100*(1), 30-47. doi: 10.1037/0022-0663.100.1.30

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools*. (Practice Guide Report: NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Gersten, R., Chard, D.J., Jayanthi, M., Baker, S.K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202-1242. doi: 10.3102/0034654309334431

Grimm, K.J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*(3), 410-426. doi: 10.1080/87565640801982486

Hanley, J. A., & McNeil, B. J. (1982). The meaning of and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167-178.

Hresko, W., Schlieve, P. L., Herron, S. R., Sawain, C., & Sherbenou, R. (2003). *Comprehensive Math Abilities Test*. Austin, TX: PRO-ED.

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). Validation of curriculum-based mathematical word problem solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review, 34*, 358-371. Retrieved from http://search.proquest.com/docview/620896484?accountid=14521

Jitendra, A. K., Griffin, C. C., McGoey, K., Gardill, M. C., Bhat, P., & Riley, T. (1998). Effects of mathematical word problem solving by students at risk or with mild disabilities. *The Journal of Educational Research*, *91*(6), 345-355.

Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of learning disabilities*, *33*(6), 567-578.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child development*, *74*(3), 834-850.

Kamphaus, R.W. (2005). *Clinical assessment of child and adolescent intelligence* (2nd Edition). New York: Springer Science + Business Media, Inc.

Kelley, B., Hosp, J.L., & Howell, K.W. (2008). Curriculum-based evaluation and math: An overview. *Assessment for Effective Intervention, 33*(4), 250-256. doi: 10.1177/1534508407313490

Leh, J. M., Jitendra, A. K., & Caskie, G. I. L., & Griffin, C. C. (2007). An evaluation of curriculum-based measurement of mathematics word problem-solving measures for monitoring third-grade students' mathematics competence. *Assessment for Effective Intervention, 32*(2), 90-99. doi: 10.1177/15345084070320020601

Lembke, E. S., Hampton, D., & Beyers, S. J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools, 49*(3), 257-272. doi: 10.1002/pits.21596

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology, 48*, 39-52. doi: 10.1016/j.jsp.2009.08.002

National Center for Education Statistics (2011). *The Nation's Report Card: Mathematics 2011* (NCES 2012-458). Institute of Educational Sciences, U.S. Department of Education, Washington, D.C.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*, U.S. Department of Education: Washington, DC.

Pearson Publishers (2009) *Scott Forsman-Addison Wesley EnVisionMath*. NY: Pearson, Inc.

Pfister, R., Schwarz, K., Carson, R., & Jancyzk, M. (2013). Easy methods for extracting individual regression slopes: Comparing SPSS, R, and Excel. *Tutorials in Quantitative Methods for Psychology, 9*(2), 72-78.

Psychological Corporation (1992). *Wechsler Individual Achievement Test*. San Antonio TX: Harcourt Brace & Co.

Raykov, T. & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Raven, J. C. (1976). *Colored Progressive Matrices Test.* London. England: H. K. Lewis & Co. Ltd.

Resendez, M, & Azin, M (2008). *A study of the effects of Pearson's 2009 enVision Math Program, Technical Report.* Press Associates, INC. http://assets.personschool.com/asset_mgr/egacy/200843/enVmath_Effi_studyfinal report.

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*(1), 49-101. Retrieved from: http://www.jstor.org/stable/3233609

Riccomini, P.J., & Witzel, B.S. (2010). *Response to intervention in math*. California: Corwin.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2006). *Assessment in special and inclusive education*. Wadsworth.

Shapiro, E. S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention, 30*(2), 15-32. doi: 10.1177/073724770503000203

Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304-325.

Sisco-Taylor, D.T., Fung, W., & Swanson, H. L. (2013). *Do curriculum-based measures of word-problem solving predict math competencies?.* Manuscript submitted for publication.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18*(3), 147-156.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795-819.

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96,* 471-491. doi:10.1037/0022-0663.96.3.471

Thurber, R.S., Shinn, M.R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498-513. Retrieved from http://search.proquest.com/docview/

U. S. Department of Education, Institute of Education Sciences. (2008). What Works Clearing House: Procedures and Standards handbook (version 2.0). Retrieved from http://ies.ed.gov/ ncee/wwc/documentsum.aspx?sid=19

Verschaffel, L. & De Corte, E. (1997). Word problems: A vehicle for promoting authentic mathematical understanding and problem solving in the primary school. In T. Nunes & P. Bryant (Eds.), *Learning and teaching mathematics: An international perspective* (pp. 69-97). England: Psychology Press.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, *28*(4), 409-426. doi:10.1080/01443410701708228

Watson, A. (2006). Some difficulties in informal assessment in mathematics. *Assessment in Education, 13*(3), 289-303. doi: 10.1080/09695940601035445

Whitehouse.gov. (2009). Office of the Press Secretary Statement release. November 23, 2009. Retrieved from http://www.whitehouse.gov/the-press-office/president-obama-launches-educate innovate-campaign-excellence-science-technology-en.

Xin, Y. P., Jitendra, A. K., & Deatline-Buchman, A. (2005). Effects of mathematical word- problem-solving instruction on middle school students with learning problems. *The Journal of Special Education*, *39*(3), 181-192.

Table 1

Alternate Form Test-Retest Reliability Coefficients for Aggregated and Non-Aggregated Forms

Form	A	B	C			
A	1.00	.	.	.	.	.
B	.72	1.00	.	.	.	.
C	.74	.74	1.00	.	.	.
M	6.45	5.58	5.91			
SD	3.68	3.26	3.62			
Form	1	2	3	4	5	6
1	1.00	.	.	.	.	.
2	.65	1.00	.	.	.	.
3	.72	.51	1.00	.	.	.
4	.55	.57	.60	1.00	.	.
5	.50	.62	.56	.67	1.00	.
6	.69	.60	.52	.62	.59	1.00
M	3.25	3.20	2.68	2.88	2.62	3.29
SD	2.04	2.01	1.91	1.72	1.98	2.07

Note. Intercorrelations for students (n = 65) that were assigned to the first presentation order (received form 1 first). Form A = Form 1 + Form 2; Form B = Form 3 + Form 4; Form C = Form 5 + Form 6. All correlations were significant at a level of $p < .001$.

Table 2

Pearson Correlations Among Variables of Interest

Measure	1	2	3	4	5	6	7
1. WPSF	1.00	.	.	.	.	.	.
2. STAR	.46	1.00	.	.	.	.	.
3. CMAT	.60	.44	1.00	.	.	.	.
4. KeyMath	.54	.46	.69	1.00	.	.	.
5. TOMA	.60	.29	.50	.47	1.00	.	.
6. WIAT	.55	.35	.46	.53	.35	1.00	.
7. TORC	.53	.27	.46	.44	.47	.42	1.00
M	4.56	3.81	8.92	6.48	4.51	16.39	15.61
SD	3.30	2.13	2.88	2.58	2.15	3.17	5.33

Note. Intercorrelations from the total sample (n = 142). All correlations were significant at $p < .001$; WPSF = Word Problem Solving Fluency CBM; STAR = experimental word problem solving outcome measure; CMAT = Problem Solving subtest from the CMAT; KeyMath = Problem Solving subtest from the KeyMath; TOMA = Story Problems subtest from the TOMA-2; WIAT = Arithmetic subtest from the WIAT; TORC = Passage Comprehension subtest from the TORC-4.

Table 3

Hierarchical Regression Models Predicting Norm-Referenced Math Problem Solving Measures

	Criterion Measure							
	KeyMath		CMAT		TOMA		STAR	
Predictor(s)	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β
Step 1	.34***		.30***		.25***		.14***	
Calculation		.42***		.33***		.19*		.29***
Reading		.26***		.32***		.39***		.14
Step 2	.05***		.10***		.14***		.09***	
Calculation		.30***		.16*		-.01		.14
Reading		.16*		.17*		.21**		.01
WPSF		.29***		.42***		.49***		.38***
Step 3	.12***		.05***		.00		.03*	
Calculation		.26***		.14		-.01		.12
Reading		.04		.09		.19*		-.06
WPSF		.53***		.58***		.53***		.50***
WPSF ROI		.40***		.27***		.06		.20*
Total R^2	.51***		.45***		.39***		.26***	
n	142		142		142		142	

Note. Calculation = Numerical Operations subtest from the Wechsler Individual Achievement Test; Reading = Text Comprehension subtest from the Test of Reading Comprehension; WPSF = Initial score on word problem solving fluency; WPSF = Rate of improvement on word problem solving fluency CBM across three time points. *$p < .05$, **$p < .01$, ***$p < .001$.

Table 4

Descriptive Statistics by Risk Group on Variables of Interest

Measure	Not At-Risk			At-Risk				
	N	*M*	*SD*	*N*	*M*	*SD*	*t*(141)	*p*
Calculation	105	17.05	2.97	37	14.51	3.00	4.46	< .001
WPSF	105	5.56	3.16	37	1.74	1.85	7.03	< .001
WPSF ROI	105	0.25	0.29	37	0.17	0.35	1.37	.17
Reading Comprehension	105	16.72	4.77	37	12.43	5.60	4.49	< .001
STAR	105	4.31	2.01	37	2.41	1.86	5.04	< .001
KeyMath	105	7.51	2.10	37	3.57	1.24	10.75	< .001
CMAT	105	10.34	1.47	37	4.89	1.93	17.85	< .001
TOMA	105	5.07	2.02	37	2.95	1.72	5.71	< .001

Note. Raw scores are reported for all measures. Calculation = Numerical operations subtest from the WIAT; WPSF = initial CBM raw score; WPSF ROI = weekly rate of improvement on the WPSF; Reading Comprehension = Text Comprehension subtest from the TORC; STAR Experimental Measure = California STAR math problem solving criterion measure; KeyMath = Problem solving subtest from the KeyMath; CMAT = Problem solving subtest from the CMAT; TOMA = Story problems subtest from the TOMA.

Figure 1

Growth on WPSF for At-Risk and Not At-Risk Students Across 12 Weeks

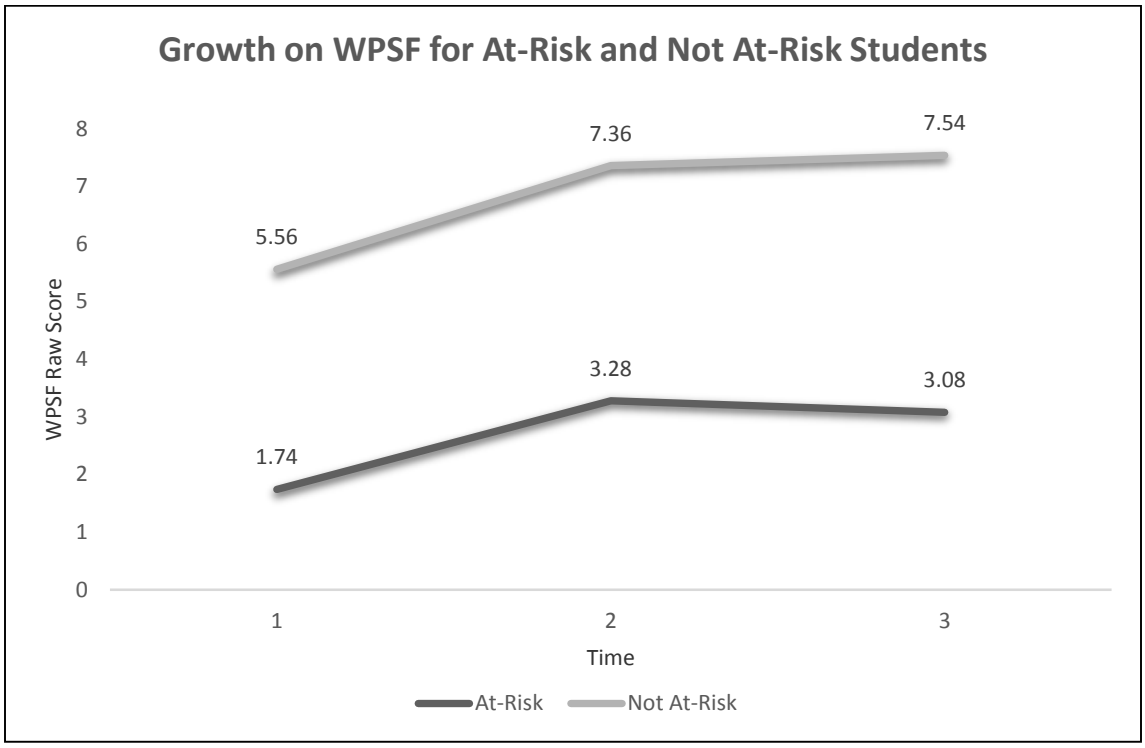

Table 5

Logistic Regression Predicting Risk Status from WPSF Level and Growth

				95% Confidence Interval for Odds Ratio		
	Parameters	*B* (SE)	Wald χ^2	LL	Odds Ratio	UL
Step 1	Constant	-1.95 (0.52)***	14.17	--	--	--
	WPSF	0.74 (0.14)***	28.95	1.60	2.10	2.75
	ROI	0.66 (0.22)**	9.25	1.27	1.94	2.97
Step 2	Constant	-2.35 (1.35)	3.01	--	--	--
	WPSF	0.71 (0.16)***	21.13	1.51	2.04	2.77
	ROI	0.64 (0.23)**	8.03	1.22	1.89	2.94
	Reading Comprehension	0.02 (0.05)	0.13	0.92	1.02	1.13
	Calculation	0.02(0.09)	0.03	0.84	1.02	1.22

Note. * $\chi^2 < .05$. ** $\chi^2 < .01$. *** $\chi^2 < .001$. WPSF = initial raw score on Word Problem Solving Fluency; ROI = Word Problem Solving Fluency rate of improvement; Reading Comprehension = Text comprehension subtest from the Test of Reading Comprehension; Calculation = Numerical operations subtest from the WIAT

Figure 2

Prediction of Spring Risk Status Using Winter WPSF Scores

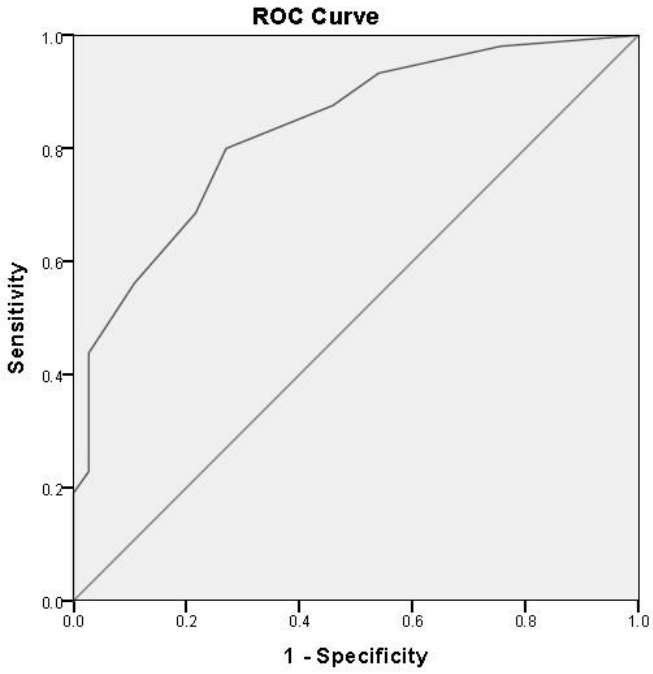

Table 6

Diagnostic Efficiency of WPSF Predicting Problem Solving Risk Status

Cut Scores	Sensitivity	Specificity	AUC	Classification Accuracy	TP	FP	TN	FN
--[a]	.00	.74	--	.74	0	0	105	37
--[b]	.59	.90	--	.82	22	15	94	11
3.5	.81	.73	.83	.75	30	7	77	28
2.5	.89	.54	--	.63	33	4	57	48

Note. TP = true positives; FP = false positives; TN = true negatives; FN = false negatives. [a]Diagnostic efficiency statistics generated from baseline logistic regression model with no predictors. [b]Diagnostic efficiency statistics generated from logistic regression model with WPSF initial raw scores, and ROIs.

Appendix A

Probe 1

1. Brandon has 4 red marbles and 5 green marbles. Brandon's brother Mike has 15 marbles. How many marbles does Brandon have in all?

2. Andy had some stamps. Then his dad bought him 5 more stamps. Now Andy has 35 stamps. How many stamps did he have to start with?

3. Steve scored 2 goals less than Greg in the soccer game. Greg scored 7 goals. How many goals did Steve score?

4. Anna found 27 seashells on the beach. Tanya found 16 seashells. How many fewer seashells did Tanya find?

5. Mia planted 10 daisies and 8 lilies in her garden. However, 3 of the flowers died. How many flowers are left?

6. Mom baked chocolate chip and sugar cookies. She baked 15 chocolate chip cookies. She baked 5 less sugar cookies than chocolate chip cookies. How many cookies did mom bake in all?

7. Ken, Rob, and Dylan were picking apples in the orchard. Ken picked 20 apples. Rob picked 25 apples. How many apples did Dylan pick if, altogether, the boys picked 75 apples?

8. Erin sees 4 beetles. Sam sees 6 beetles. Mary sees 5 beetles. They all see 3 butterflies, too. How many beetles do Erin, Sam, and Mary see altogether?

9. Sofie ate 6 strawberries. That is 4 less strawberries than what Emma ate. How many strawberries did Emma eat?

10. Caleb picked 5 ears of corn. Gabe picked 3 more ears of corn than Caleb. Jack picked 2 more ears of corn than Gabe. How many ears of corn did they pick in all?

11. Julie digs up 4 onions. Ben digs up 6 more onions than Julie. How many onions do they both pick together?

12. Mrs. Robinson has 19 students in her class. There are 9 boys in her class. Mr. Ross has 20 students in his class, and 12 of them are boys. How many more girls are in Mrs. Robinson's class than in Mr. Ross's class?

Probe 2

1. Ethan has 9 small trucks and 4 big trucks. Ethan's brother Josh has 11 trucks. How many trucks does Ethan have in all?

2. James had some baseball cards. Then his uncle gave him 7 more baseball cards. Now James has 27 cards. How many baseball cards did he have to start with?

3. Ryan scored 20 points less in the new video game than Zach. Zach scored 60 points. How many points did Ryan score?

4. Sofia picked 19 flowers in the park. Mia picked 15 flowers. How many fewer flowers did Mia pick?

5. Zoe had 6 ladybugs with five dots and 10 ladybugs with six dots. However, 5 of the ladybugs flew away. How many ladybugs are left?

6. Daniel and Jack scored many goals at their last soccer game. Daniel scored 9 goals. Jack scored 3 less goals than Daniel. How many goals did they both score in all?

7. Luke, Rob, and Noah collect coins from different countries. Luke has 11 coins. Rob has 16 coins. How many coins does Noah have, if altogether the boys have 39 coins?

8. Bill saw 7 hermit crabs. Peter saw 6 hermit crabs. Leo saw 4 hermit crabs. He also saw 5 starfish. How many hermit crabs did the three boys see in all?

9. There are 9 frogs on a log. That is 4 less frogs than the number of frogs in the pond. How many frogs are there in the pond?

10. Tom picked 5 berries. Rick picked 3 more berries than Tom. Freddy picked 4 more berries than Rick. How many berries did they pick in all?

11. Paul counted 20 striped fish. He also counted 8 more clownfish than striped fish. How many fish did Paul count in all?

12. A pet shop sold 12 hamsters and 8 kittens on Saturday. They sold another 22 pets on Sunday. 10 of those pets were hamsters and the rest were kittens. How many kittens did the pet store sell in all?

Probe 3

1. Alexis has 4 baby dolls and 7 Barbie dolls. Alexis's cousin Lily has 6 Barbie dolls. How many dolls does Alexis have in all?

2. Dylan read several pages of the book in the morning. In the afternoon, he read 10 more pages. So far, Dylan has read 24 pages in all. How many pages did Dylan read in the morning?

3. Michael spent $3 less on lunch than Connor. Connor spent $8 to buy his lunch. How much money did Michael spend on his lunch?

4. Lauren already colored 17 pages in her coloring book. Sarah colored 14 pages in her coloring book. How many fewer pages did Sarah color than Lauren?

5. Mom bought 8 yellow apples and 9 green apples. Her daughter, Mary ate 4 of the apples. How many apples are left?

6. A pet store had puppies and kittens for sale. They sold 9 kittens today. They sold 3 less puppies than kittens. How many pets did the store sell in all?

7. Christian, Dan, and Ted were selling boxes of popcorn to raise money for their school. Christian sold 6 boxes. Dan sold 8 boxes. How many boxes of popcorn did Ted sell, if altogether the boys sold 24 boxes?

8. There are 12 motorboats in the harbor. There are also 4 ferryboats and 10 fishing boats in the harbor. 4 fishermen are looking at the boats. How many boats are there in the harbor in all?

9. Zoe has 14 colored pencils, that is 5 less than what Kelly has. How many colored pencils does Kelly have?

10. Kit saw 7 squirrels in a tree. Abby saw 2 less squirrels than Kit. Brianna saw 3 more than Abby. How many squirrels do they see in all?

11. Zack and Mike like to go to the pool. Zack jumped in the pool 7 times. Mike jumped in 4 more times than Zack. How many times did they both jump in the pool altogether?

12. Amy picked 8 small flowers and 12 big flowers. Dana picked 24 flowers in all. 14 of them were small. How many more big flowers did Amy pick than Dana?

Probe 4

1. Isabella had 12 addition problems and 11 subtraction problems to solve. Isabella's brother Owen had 25 problems to do. How many problems did Isabella need to solve in all?

2. Carter had saved some money. His dad gave him $6 more dollars. Carter now has $26 in all. How much money did Carter have at the beginning?

3. Mom baked 8 less sugar cookies than chocolate chip cookies. She baked 19 chocolate chip cookies. How many sugar cookies did mom bake?

4. Alyssa practiced playing piano for 25 minutes today. Her sister Riley practiced for 15 minutes. How many fewer minutes did Riley practice playing piano than Alyssa?

5. Maya had 20 blue beads and 25 red beads in her bag. She used 15 beads to make a necklace for her sister. How many beads does Maya have left?

6. Mr. McDonald had a little farm. He had 25 goats on his farm. He had 15 less cows than goats on the farm. How many animals did Mr. McDonald have on his farm in all?

7. Mrs. Leopold has stickers that she likes to give to her students. Last week, she gave out 24 stickers. This week, she gave out 33 stickers. How many stickers are left, if at the beginning, Mrs. Leopold had 80 stickers in all?

8. There are 5 white roses blooming in grandma's garden. There are 4 yellow roses and 8 red roses. There are still 6 tulips blooming, too. How many roses are blooming in all?

9. There are 7 birds sitting on the first tree. That is 5 less than the birds sitting on the second tree. How many birds are sitting on the second tree?

10. Bob found 15 acorns in the park. His friend, Rick found 4 less. Maya found 2 more acorns than Rick. How many acorns did the three of them find in all?

11. Anna used 10 shells to make a bracelet. Her sister Jenna used 3 mores shells for her bracelet. How many shells did they both use in all?

12. Adam gives food to the otters. Yesterday, he gave them 13 mussels and 10 crabs. Today he gave the otters 26 food items. 12 of them were crabs, and the rest were mussels. How many mussels did Adam give to the otters altogether for both days?

Probe 5

1. Gabriel bought new things from a toy store. Gabriel bought a kite for $6 and a soccer ball for $12. His brother, Christian got a robotic dog for $22. How much money did Gabriel spend on his toys in all?

2. Sammy sold some Girl Scout cookies in the morning. Later in the afternoon, she sold 25 more boxes. If Sammy sold 55 boxes in all, how many boxes of cookies did she sell in the morning?

3. Jacob spent 20 minutes less doing his homework today than practicing soccer. He had soccer practice for 45 minutes. How many minutes did Jacob spend on his homework today?

4. Sydney is 44 inches tall. Her sister, Kylee is 57 inches tall. How many inches shorter is Sydney than her sister?

5. Debby baked 10 banana nut muffins and 8 lemon muffins. Alyssa ate 4 of the muffins. How many muffins were left?

6. William and Matt were fishing last Sunday. William caught 15 little fish. Matt caught 5 less fish than William. How many fish did they both catch in all?

7. There were 26 cloudy days in June, July, and August. There were 14 cloudy days in June. There were 7 cloudy days in July. How many cloudy days were there in August?

8. Eva picked 8 carrots. Sammy picked 6 carrots. Lola picked 4 carrots and 2 tomatoes. How many carrots did they pick in all?

9. There are 6 big fish in a tank. That is 4 less than the number of little fish in the tank. How many little fish are there in the tank?

10. Gina saw 7 leafy sea dragons. She saw 2 more weedy sea dragons than leafy sea dragons. She also saw 5 less zebra sea dragons than weedy sea dragons. How many sea dragons did Gina see in all?

11. Jose picked 5 green peppers. He picked 2 more red than green peppers. How many peppers did Jose pick in all?

12. There are 8 boys and 12 girls in Mr. Jackson's class. 4 boys have blue eyes, and the rest have brown. 7 girls have brown eyes, and the rest have blue. How many students in Mr. Jackson's class have blue eyes?

Probe 6

1. Olivia planted 7 red tulips and 5 yellow tulips. Her sister, Ella planted 8 tulips. How many tulips did Olivia plant in all?

2. Caleb likes to eat junk food. This month, he gained 4 pounds. Now Caleb weighs 68 pounds. How many pounds did Caleb weigh before he gained 4 pounds?

3. Hannah lost 5 teeth less than her older sister Taylor. Taylor lost 12 teeth. How many teeth did Hannah lose?

4. Andrew did 49 push-ups. Lucas did 37 push-ups. How many fewer push-ups did Lucas do compared to Andrew?

5. A high school administration hired 10 male teachers and 8 female teachers for this year. However, 3 of the teachers left. How many new teachers stayed?

6. There are a different number of boys and girls in Mrs. Carter's class. There are 9 girls in the class. There are 3 less girls than boys in the class. How many students are there in Mrs. Carter's class in all?

7. Alex wants to buy a bike that costs $65. Alex had already saved $40. His father gave him $10 more for washing cars. How much more money does Alex need to save to get the bike?

8. John cut 2 onions to make salsa. He cut 8 tomatoes and 4 peppers. It took him 2 hours to make salsa. How many vegetables did John use in all?

9. Rosa drew 8 flowers. She drew 3 less butterflies than flowers. How many butterflies did Rosa draw?

10. Tina read 5 books in June. She read 2 more books in July than in June. She read 3 more books in August than in July. How many books did Tina read in all?

11. Paul has 20 whale stickers. He has 7 more shark stickers than whale stickers. How many stickers does Paul have in all?

12. Jerry made 26 flags. He made 14 red flags. The rest of the flags were blue. Bob made 31 flags. 11 of them were red and the rest were blue. How many more blue flags did Bob make than Jerry?