UNIVERSITY OF CALIFORNIA
RIVERSIDE

Implementation of a Multi-Tiered Support System in Schools

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

Master of Arts

in

Education

by

Abigail Ann Hatch

September 2012

Thesis Committee:
Dr. Michael Vanderwood, Chairperson
Dr. Cathleen Geraghty-Jenkinson
Dr. Gregory Palardy

The Thesis of Abigail Ann Hatch is approved:

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Implementation of a Multi-Tiered Support System in Schools

by

Abigail Ann Hatch

Master of Arts, Graduate Program in Education
University of California, Riverside, September 2012
Dr. Michael Vanderwood, Chairperson

This study evaluated the properties of a 37-item survey designed to measure a school's implementation of a multi-tiered support system (MTSS). Twenty-four elementary schools from one urban school district participated in the study and were grouped based on assigned level of implementation. Survey responses were collected from 148 different staff members. Internal consistency was evaluated using Cronbach's alpha which equaled 0.981, 95% CI [0.975, 0.986]. Interrater agreement, as measured by average measure intraclass correlation (ICC), was 0.979, 95% CI [0.973, 0.985] across all schools and raters. An ANOVA comparing the mean survey score total across groups did not reveal any significant differences [$F(2, 96) = 0.41$, $p = 0.67$], nor were there any significant correlations between survey score totals and student outcomes as measured by AIMSweb. Reliability of the survey was strong, but further refinement is needed to improve the survey's ability to discriminate between implementation levels.

Table of Contents

List of Tables

Implementation of a Multi-Tiered Support System in Schools

Over the last twenty years, there has been a trend towards increased accountability and assessment in education (Barber, 2004). In the U.S., this trend became part of federal law with the passage of the No Child Left Behind Act (NCLB, 2002). Under this act, the National Assessment of Educational Progress (NAEP) is required to conduct national and state assessments at least once every two years in reading and mathematics, for grades 4 and 8. Data from these biennial assessments are published and known as the "Nation's Report Card". In the latest update on reading achievement (National Center for Education Statistics [NCES], 2011), the authors reported that eight percent of fourth graders scored in the Advanced range, 26 percent in Proficient, and 33 percent in the Basic score range. Although this represents some progress over the last 20 years, this still means that 33 percent (one-third) of fourth graders are not even able to demonstrate the "partial mastery of prerequisite knowledge and skills" (p. 6) that is required for the Basic achievement level. Similarly, among students receiving special education services, it is estimated that 40% were initially identified based on their difficulty with learning to read (U.S. Department of Education [USDE], Office of Special Education and Rehabilitative Services [OSERS], 2002). For many of these children, the reading difficulties they experienced may have been preventable (Torgesen, 2002; Vellutino et al., 1996).

Early difficulties with acquiring literacy skills are highly predictive of later reading problems. In a 1988 study, Juel found that 88% of poor readers in first grade will continue to be poor readers in fourth grade. Stanovich (1986) found a similar pattern, and coined the term, “Matthew Effect,” to describe how initial differences in literacy acquisition led to increasingly larger achievement gaps over time. Instruction plays a critical role in preventing the development of serious reading problems (Fletcher, Lyon, Fuchs, & Barnes, 2007; Scanlon, Vellutino, Small, Fanuele, & Sweeney, 2005).

Under NCLB (2002), schools are required to show adequate yearly progress (AYP) as measured by state standards-based exams and to meet certain targets each year. This has placed increased pressure on educators to improve students’ reading outcomes. Multi-tiered support systems (MTSSs) are one method schools are using to meet the need both for early identification of students who are at risk for poor reading outcomes and for provision of targeted instruction to correct those problems (Fuchs & Fuchs, 2004; Linan-Thompson & Vaughn, 2010; USDE, Institute of Education Sciences [IES], 2009).

Multi-Tiered Support Systems

As it is typically conceptualized, a MTSS, also commonly referred to as Response-to-Intervention (RTI), consists of a number of interconnected elements. The Office of Special Education and Rehabilitation Services (USDE, 2007) described the main components as (a) high quality, evidence-based core instruction, (b) universal screening, (c) progress monitoring, and (d) multiple levels (tiers) of instruction. The foundation of MTSSs is a problem-solving approach and reliance on data for making decisions (Fuchs, Fuchs, & Compton, 2010; Kovaleski, 2007; National Center on Response to Intervention [NCRTI], 2010). The current multi-tiered approach evolved over time as this method was scaled up from individual students to the entire school system and combined with an emphasis on prevention (Tilly, 2008; Vaughn & Klingner, 2007).

When a MTSS is applied to literacy development and all the components of the system are effectively working together, a number of positive outcomes can be expected (Linan-Thompson & Vaughn, 2010; NCRTI, 2010; Wanzek & Vaughn, 2007). At the student level, researchers have been able to identify improvements in reading skills using a wide variety of different measures (e.g., Gray Oral Reading Test, Woodcock-Johnson III, Test of Word Reading Efficiency; Denton, Fletcher, Simos, Papanicolaou, & Anthony, 2007; Kerins, Trotter, & Schoenbrodt, 2011). However, curriculum-based measurement is by far the most commonly used approach (e.g., Al Otaiba et al, 2011; Graves et al, 2011; Kerins et al, 2011). Indeed, collecting data on student progress and using that data to inform instruction and intervention is an integral part of a MTSS (USDE, 2009).

Researchers have looked at system-wide effects as well. In keeping with the focus on prevention of reading problems, frequently studied outcomes include (a) reduction in the number of referrals to special education and (b) an increase in the percentage of referrals who actually qualify for special education (i.e., increased accuracy; e.g., VanDerHeyden, Witt, & Gilbertson, 2007). In a meta-analysis of research on MTSSs, Burns, Appleton, and Stehouwer (2005) found a mean effect size of 1.53 (SD = 1.02) on systemic variables such as referrals to special education and number of students retained. In a three-year study of the effects of a MTSS implemented at 318 schools in Florida, Torgesen (2009) reported a dramatic reduction in the percentage of students diagnosed with a learning disability (LD) in kindergarten through third grade.

Similarly, VanDerHeyden, Witt and Gilbertson (2007) found that multi-year use of a MTSS led to a reduction in the number of evaluations conducted and an increase in the percentage of evaluated children who actually qualified. They also found that the disproportionality of identified children by gender (with males being overidentified at baseline) was reduced.

However, there is debate on the use of referrals as an outcome measure (Batsche, Kavale & Kovaleski, 2006; Vaughn & Fuchs, 2006). One key point in the debate is that, since MTSSs are implemented schoolwide, they should improve outcomes for *all* students. Shapiro and Clemens (2009) suggested ways to take the same curriculum-based measurements that are used to evaluate student progress and aggregate that information to determine the effectiveness of a school's MTSS at the grade and school levels. These proposed indicators included (a) change in percentage of students who score in the at-risk category, (b) students' rate of improvement overall, (c) movement between tiers, and (d) rate of improvement for students within Tiers 2 and 3. Each of these indicators can be calculated using screening data which allows schools to evaluate the effectiveness of their MTSS within 3-4 months.

Regardless of the outcome measures used, researchers and school staff also have to be aware that the effectiveness of any program can be hampered or even nullified by poor implementation. In their synthesis of research on intervention, Fixsen, Naoom, Blasé, Friedman, and Wallace (2005) pointed out that, "Desirable outcomes are achieved only when effective programs are implemented well" (p. 12). This problem has been observed across multiple fields of study (Fixsen et al., 2005) and may be one of the reasons for differences in program effectiveness that have been found in field-based vs. university-run intervention studies (e.g., Burns & Symington, 2002).

Research on the implementation process has identified six key phases in the process (Fixsen, et al., 2005): (1) exploration and adoption, (2) program installation, (3) initial implementation, (4) full operation, (5) innovation, and (6) sustainability. School staff have to work through the first three stages of finding and deciding on an program, setting it up and preparing, and starting the implementation before they can reach full operation, where the program becomes routine and automatic. It is not until this stage, the fourth one in the process, that schools can expect to see the anticipated benefits of the system being implemented, and, when implementing a MTSS, it may take 3-5 years to reach full implementation (Shapiro & Clemens, 2009). For this reason, it is important for school personnel to evaluate how well they are implementing MTSSs at their sites even before trying to evaluate program effectiveness.

Implementation Guidelines and Outcomes

Due to the complexity of the process and the importance of effectively implementing a MTSS, a number of agencies have written specific guidelines for school personnel. A couple of notable sources are the Institute of Education Sciences (IES), and the National Association of State Directors of Special Education (NASDSE). What Works Clearinghouse (WWC), the authors of the IES publication titled, "Assisting Students Struggling with Reading: Response to Intervention (RtI) and Multi-Tier Intervention in the Primary Grades" (USDE, 2009), provided a summary of research related to the key components of a MTSS, and made specific recommendations on how to implement these key components. Similarly, NASDSE provided guidelines in their "Blueprints" publications for districts and individual schools (NASDSE, 2008a; NASDSE, 2008b). The NASDSE publications also contain self-evaluation ranking scales and resources that are linked with the guidelines given. Schools can use these documents for evaluating both their readiness for implementation and the extent to which it has occurred.

Other organizations have put together similar instruments. The National Center on Response to Intervention (NCRTI) offers a self-assessment that can be used to evaluate a state's current level of system implementation (NCRTI, 2009), and the Florida Problem Solving and Response to Intervention Project (FPSRTI) offers numerous measures to assess beliefs and perceptions, implementation of critical system components, personnel satisfaction with the process, etc. (e.g., FPSRTI, 2007). These instruments are intended to assist schools in evaluating the extent to which successful implementation has taken place, to identify whether there is improvement in implementation over time, and to provide important feedback to school staff on what is or is not being done well.

This information can be an important guide for professional development activities. Professional development and consultation are strategies frequently used to increase the extent of implementation (Chaparro, et al., 2012; Duhon, et al., 2008), and consultation can be used within a MTSS framework to guide school staff through the implementation process (Knotek, 2005; Powers, et al., 2008). The steps in the consultation process typically consist of (1) problem identification, (2) problem analysis,

(3) treatment implementation, and (4) treatment evaluation (Kratochwill & Bergan, 1990). These steps are similar to those of the implementation process and can occur in conjunction with it, where problem identification and problem analysis occur with exploration and adoption, treatment implementation occurs with program installation, initial implementation, and full operation, and treatment evaluation occurs throughout the process. The consultation process can also be used during implementation phases, e.g., during initial or full implementation, as a way to problem-solve and fine tune implementation. As a strategy, the consultative approach acknowledges that the essential outcome of the implementation process is change in adult behavior (Fixsen et al., 2005), and provides a process for facilitating that change in order to benefit the client (i.e., student). Since adult behavior changes are key outcomes, self-assessment instruments can be useful tools for determining whether or not such changes have occurred.

Purpose of Study

Although self-report tools are currently being used by schools, there is little research assessing the reliability or validity of these instruments, much less their connection with important student outcomes. The purpose of this study is to evaluate the psychometric properties of this MTSS survey and to examine the relationship between staff members' report of MTSS implementation and student outcomes at the school. Specifically, this study seeks to answer the following research questions:

1. Does the MTSS survey meet established standards for acceptable internal consistency and interrater agreement?

2. Are there differences in level of implementation across the Pilot, MTSS, and Screening groups?
3. To what extent are scores on the MTSS implementation survey correlated with student outcomes as measured by AIMSweb screening assessments?

Methods

Participants

Schools. The participating schools were all elementary schools from one urban school district in Southern California. The district decided to stagger their implementation of a MTSS approach and selected schools to participate at each stage. In Year 1, four schools started midway through the school year (Pilot group). Ten more schools started implementation in Year 2, at the beginning of the school year (MTSS group), and the remaining 10 schools were asked to implement universal screening only (Screening group).

Of the 24 schools in the district, 22 enrolled students in kindergarten through fifth grade. One school in the Pilot group had only kindergarten through second, and one school in the MTSS group enrolled only third through fifth grades. The schools in the Pilot group had significantly higher enrollment that those in the MTSS group. All schools had a majority of students eligible for free or reduced lunch (85% district-wide; Compton Unified School District, 2009), and the ethnic breakdown within the district was 76% Hispanic, 15% African-American, and less than 1% White, Asian, American-Indian or bi/multiracial (8% not reported; see Table 1).

Each school in the Pilot and MTSS groups received three days of training from one of two experts who had been hired to consult with the district and assist with the implementation process. These trainers had each worked with other school districts and assisted in their implementation of MTSSs. In this district, they conducted large-group trainings, typically with personnel from five schools at a time. School-selected teams of administrators, teachers, and support staff (e.g., reading specialists) attended the trainings, which consisted of lectures on the content with time allowed periodically for schools to discuss as a group how they planned to apply the information. The external consultants covered all facets of MTSS implementation – from universal screening to intervention, progress monitoring, and data-based decision-making. An evidence-based reading program, called Peer Assisted Learning Strategies (PALS; Fuchs, Fuchs, Simmons, & Mathes, 2008; Fuchs et al., 2011; Fuchs et al, 2008), was also included in the training and recommended as an enhancement to core instruction at Tier 1. A graduate student (the author) also worked with the MTSS schools as a consultant throughout Year 2 to provide additional, on-site training as requested and to monitor and support implementation efforts. The Screening schools received one day of AIMSweb training which covered the assessments used for universal screening. All schools also received additional training for their designated data management specialists on how to work with the AIMSweb site including how to download assessments, input data, and generate reports.

Staff. A total of 148 different staff members from 24 schools completed the survey (see Table 2). More than two-thirds of the respondents were female (117 out of

146). Most responses came from teachers (108 out of 148), but some were from resource specialists (16), principals or assistant principals (22), or other school personnel (2). The majority of respondents had been in their current position for more than six years (65 out of 147 for six to 10 years, 40 for more than 10 years). There was a fairly wide range of ages represented, but the majority of responses were from individuals in their thirties (61 out of 142) or forties (47). Of the three different implementation groups, the majority of the responses came from the MTSS group (81 out of 147) with only 24 from the Pilot group and 41 from the Screening group. However, all grade levels were fairly evenly represented (ranging from 17 to 21 out of 146) with 35 respondents answering for the school as a whole.

Assessments

Survey. For the purposes of this study, an existing MTSS survey was selected. This survey was developed by staff at a school district in northern California based upon training they had received on implementing MTSSs in schools. The survey questions were intentionally created to evaluate key aspects of the MTSS process. It had been used in the district for three years in order to evaluate the extent of implementation and to identify areas for further professional development. The study author revised the answer response choices to focus the survey more closely on the implementation aspect (see Appendix).

The survey itself was organized into three primary categories: Measurement, Curriculum and Instruction, and Problem-Solving Teams. These categories reflected the key components of a MTSS (Fuchs et al., 2010; Kovaleski, 2007; NCRTI, 2010; USDE, 2007). The Measurement strand asked about universal screening as well as the collection and use of data. The Curriculum and Instruction strand consisted of two areas – curriculum and instruction – and evaluated the quality of core (Tier 1) instruction. The third strand, Problem-Solving Teams, looked at five different areas: team characteristics, problem identification, plan development, plan implementation, and plan evaluation. Together, these areas evaluated the key MTSS components of progress monitoring and the use of multiple tiers. In addition, questions were included that assess the use of data for making decisions within a problem-solving approach.

AIMSweb. The AIMSweb benchmark and progress monitoring system (www.aimsweb.com) was used. At the kindergarten and first grade level, the AIMSweb Tests of Early Literacy (TEL; Shinn & Shinn, 2002a) were used. Reading Curriculum-Based Measurement (RCBM; Shinn & Shinn, 2002b) was administered for first grade through fifth grade, and the Maze (Shinn & Shinn, 2002c) assessment was given for second through fifth. Prior research has found these measures to be reliable and valid for use within a MTSS (Kame'enui, 2002).

Tests of Early Literacy. The TEL include Letter Naming Fluency (LNF), Letter Sound Fluency (LSF), Phonemic Segmentation Fluency (PSF) and Nonsense Word Fluency (NWF). These tests were designed to measure early indicators that predict later reading success. They were administered to kindergarten and first grade students according to AIMSweb's recommended screening schedule (Shinn & Shinn, 2002a).

Letter Naming Fluency. The LNF assessment is a 1-minute test that measures students' ability to correctly name visually presented letters. The developers of the

Dynamic Indicators of Basic Literacy Skills (DIBELS) researched and reported on the technical adequacy of the LNF measure (Good et al., 2004). They found that alternate-form reliability of LNF was equal to .89 when used with kindergartners and .86 with first graders. When compared with scores on the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R), the median criterion-related validity was .70 for kindergarteners and .53 for first graders. From kindergarten to first grade, LNF's predictive validity for the WJ-R Reading Cluster was .66, and .72 for oral reading fluency (ORF). From first grade to second grade, LNF's predictive validity for the WJ-R Reading Cluster was .62, and .77 for ORF.

Letter Sound Fluency. LSF is also a timed, 1-minute assessment and measures students' ability to give the correct sound for visually presented letters. Tests of letter-sound fluency have found strong reliability as well as good concurrent and predictive validity (Ritchey, 2008; Speece & Case, 2001). In a 2001 study, Elliott, Lee, and Tollefson found alternate-form reliability of .82, test-retest reliability of .83, and interrater reliability of .82 with a sample of 75 kindergartners. Criterion-related validity with the WJ-R Reading Cluster was .58.

Phonemic Segmentation Fluency. The PSF assessment is 1 minute long and measures students' ability to segment orally presented words into separate phonemes. Good et al. (2004) found alternate-form reliability of .79, and Elliott, Lee and Tollefson (2001) reported criterion validity ranging from .60 to .89. However, compared with other early literacy measures, PSF has relatively weak predictive validity with studies reporting coefficients ranging from .13 to .54 (Clemens, Shapiro, & Thoemmes, 2011).

Nonsense Word Fluency. The NWF assessment is also 1 minute long and measures students' ability to sound out non-real words presented visually. Ritchey (2008) found reliability of .86, concurrent validity of .79, and predictive validity ranging from .65 to .76. Other studies have found predictive validity coefficients ranging from .37 to .74 (Clemens et al., 2011).

Reading. Two different AIMSweb measures were used to assess reading skill – R-CBM and Maze. Both of these measures were designed as indicators of reading comprehension ability. They were used with second through fifth grades at each of the screening periods, and R-CBM was also administered to first graders at the spring screening.

Reading – Curriculum-Based Measurement. AIMSweb R-CBM passages were used to assess oral reading fluency. These passages were intentionally written to be of equivalent difficulty within each grade level as measured by the Fry readability formula (Howe & Shinn, 2002). Passages varied in length based on grade level: 250 words for first and second grade, 300 for third grade, and 350 for fourth and fifth grades. For the passages used at the fall, winter, and spring benchmarks, alternate-form reliability ranged from .80 to .91 (Howe & Shinn, 2002). Studies have found test-retest reliability coefficients ranging from .82 to .97, interrater reliability of .99, and criterion-related validity coefficients ranging from .26 to .91 (Shinn & Shinn, 2002b).

Maze. AIMSweb Maze passages were also designed to have equivalent difficulty levels within each grade. These passages use a multiple-choice cloze task. Starting with the second sentence of each story, every seventh word is replaced by three choices within parentheses – the correct response, a "near" distractor, and a "far" distractor. The near distractor is a word that is the same part of speech as the correct answer, but does not make sense within the sentence. The far distractor is a word randomly selected from the story that does not make sense within the sentence. Students have three minutes to read the passage silent and circle the best response from the available choices. Shin, Deno, and Espin (2000) found alternate-form reliability ranging from .69 to .91. Hale and colleagues (2010) found criterion-related validity coefficients of .86 with both R-CBM and the WJ-III Broad Reading Cluster.

Procedures

Survey. The MTSS survey was distributed to school principals through Survey Monkey (www.surveymonkey.com) in the fall following Year 2, the MTSS group's first full year of implementation. Principals were asked to have the person most knowledgeable about MTSS implementation at each grade-level complete the survey for that grade specifically. They were also asked to complete the survey themselves based on their school's implementation overall.

Follow-up emails were sent out by the research team and district staff. Additionally after two weeks, the author called each of the principals who had not yet responded and dropped off a hard copy of the survey at each school. Fifty-five surveys were collected this way.

However, for the intended methods of data analysis (specifically, ANOVA), power analysis indicated that 159 surveys were needed in order to detect a medium effect ($F = .25$). For this reason, the author expanded the pool of potential respondents. An email was sent to all school staff whose contact information was on file (526 people) and printed copies of the survey were distributed to each school in the district. Efforts were continued until 148 responses were collected.

AIMSweb. AIMSweb assessments were administered throughout the school year and the data collected and input locally by school personnel. As part of their training, schools were taught to use one of two approaches when administering assessments – they could either train teachers to each assess their own class, or train a team and have them assess all the different classrooms. Most schools chose one approach or the other, but some used a combination, e.g., had second through fifth grade teachers assess their own classes but used a team to assess kindergarten and first grade.

Interrater agreement was measured for a small sample of R-CBM assessments. Total percentage agreement was calculated on a word-by-word basis, taking the number of agreements divided by the number of agreements plus disagreements and then multiplying by 100. Average agreement was 97% (N = 11; range = 93-100%). These results are in line with those reported in prior research (Shinn & Shinn, 2002b).

Results

At the end of the survey completion window, 148 responses were obtained, which equaled a response rate of 30% (148 out of 526) with an average of 6 completed surveys per school. Response rate varied by group: 19% (24 out of 128) for the Pilot group, 40% (81 out of 204) for the MTSS group, and 21% (41 out of 194) for the Screening group. Survey responses and AIMSweb assessment results were analyzed using IBM SPSS Statistics software (version 18). Survey results were analyzed for internal consistency using Cronbach's alpha and inter-rater agreement using a correlational approach (Salvia, Ysseldyke & Bolt, 2010). Responses were grouped by school, and intraclass correlations were calculated across grade-levels and raters. Survey totals were analyzed in SPSS using ANOVA to identify any differences in reported implementation between the Pilot, MTSS, and Screening groups. To examine the relationship between implementation and student outcomes, aggregate student growth scores were calculated by school, grade-level, and measure. Then, a correlation was run between level of implementation, as measured by the MTSS survey, and student growth from fall to spring screening, as measured by the AIMSweb assessments.

Of the completed surveys, 33% (49) had one or more questions that were left blank. To handle missing responses, mean imputation was considered but not used due to the effect that approach can have on correlations between variables (Tanguma, 2000). Rather than entirely deleting incomplete surveys, pairwise deletion was used when possible in order to retain the maximum amount of data (Pallant, 2007). In addition, five schools were missing student outcome data since they did not complete the spring screening. An independent samples t-test was conducted to compare survey totals and subscale scores for those with missing and non-missing student outcome data. There was no significant difference found between schools that did not complete the spring screening (M = 126.97, SD = 3.79) and those that did (M = 149.89, SD = 25.07; t (21) = -1.73, p = 0.10).

Internal Consistency of the Survey

The first research question concerned the reliability of the survey, and analysis was conducted to determine if it met established standards for internal consistency. The survey consisted of 6 questions about demographic information (e.g., gender, years of experience) and 37 questions about MTSSs organized into three different strands consisting of eight subcategories altogether. Since a MTSS is conceptualized as a system working together, all questions were considered to be evaluating the same construct – the implementation of a MTSS. For this reason, the survey as a whole was evaluated for internal consistency using Cronbach's alpha. The overall alpha coefficient was 0.981, 95% CI [0.975, 0.986] which is excellent (Fan & Thompson, 2001; Pallant, 2007). For the Measurement strand, which consisted of 4 items, it was 0.819. For the Curriculum and Instruction strand (10 items), it was 0.901, and for the Problem-Solving Teams strand (23 items), it was 0.980. The data shows that omitting any one question would not affect alpha significantly (see Table 3).

Interrater Agreement

The second part of research question 1 concerned interrater reliability. Of the 24 participating schools, multiple ratings were obtained for 21 schools. These data were used to calculate interrater agreement, which was done using intraclass correlation (ICC; Muschkin, Malone, & Conduct Problems Prevention Research Group, 2007). The ICC was used since it can estimate reliability when there are multiple (i.e., more than two) raters and when the data include a different number of raters for each target (school; Muschkin et al, 2007; Wuensch, 2007). The single measure intraclass correlation provides a measure of the reliability of the scores given by a single, typical rater. It is calculated using the following formula:

$$\frac{MS_{Subjects} - MS_{RxS}}{MS_{Subjects} + (df_{Raters})MS_{RxS} + \frac{n_{Raters}(MS_{Raters} - MS_{RxS})}{n_{Subjects}}}. \qquad (1)$$

In this study, however, multiple raters provided scores for the same school, so the average measure ICC was used since it provides a measure of the reliability of all raters averaged together. The average measure ICC is calculated using Equation 2 below.

$$\frac{r * icc}{1 + (r - 1)icc} \qquad (2)$$

where *r* is the number of raters and *icc* is the single measure intraclass correlation coefficient. The average measure ICC was 0.979, 95% CI [0.973, 0.985] but varied by school (see Table 4).

Level of Implementation – Group Comparison

The second research question was whether there were group differences in level of implementation. The survey responses were categorized by implementation group (i.e., Pilot, MTSS, and Screening), and a single-factor between-subjects analysis of variance (ANOVA) was run to identify any group differences in implementation level. ANOVA is a robust test, but can be more sensitive to violation of assumptions if the groups being compared do not have an equal number of subjects (Roberts & Russo, 1999). In this analysis, the three groups differed in size with the Pilot group containing 24 responses, 82 in the MTSS group, and 41 in the Screening group. For this reason, data were examined for any violation of assumptions that might impact the results and ability to detect significant group differences (Roberts & Russo, 1999).

The assumption of normal distribution was violated. Response data were skewed (skewness = -1.27). However, scores were skewed in the same way across groups, so mean comparison was still possible and appropriate. Levene's test of the equality of error variances was significant ($p < 0.05$), indicating that the assumption of homogeneity of variance was violated (on the total survey score, Group 1 variance = 412.48, Group 2 = 1543.45, and Group 3 = 1367.49). According to the guideline provided by Roberts and Russo (1999), it is still safe to analyze the data as long as the largest group variance is less than four times the smallest group variance. Although that was true in this case, the data were close to exceeding the guideline. For this reason, an F ratio was run to see if there was a significant difference between the largest group variance and the smallest group variance. This test was not significant [F (2, 96) = 3.74, p = 0.97].

After checking assumptions, the ANOVA was run using the 99 surveys that were complete listwise (see Table 5). There were no significant group mean differences on the overall survey score [F (2, 96) = 0.41, p = 0.67], or on the measurement [F (2, 130) = 0.88, p = 0.42], instruction [F (2, 124) = 0.37, p = 0.69], or problem-solving team [F (2, 98) = 0.64, p = 0.53] strands. Given the high p values, it is implausible that any violation of the assumptions underlying this statistical test caused otherwise significant differences to appear insignificant (Roberts & Russo, 1999).

In addition, a discriminant analysis was conducted to identify any individual items that could discriminate between groups. However, in testing the equality of group means, no significant differences between groups were identified for any of the individual test items. For this reason, no further analysis was conducted (Burns & Burns, 2008).

Finally, an analysis of the student outcome measures was done to determine if the two implementation groups (Pilot and MTSS) could be collapsed into one group which would improve power to detect group differences on the survey. The aggregate fall scores for each grade and outcome measure were compared across the two groups. The MTSS group had higher means on each measure, and of the 10 comparisons, 4 were significant. (See Table 6.) However, after applying the Bonferroni adjustment for the number of comparisons (n=10; $p = 0.005$), none of the differences between the Pilot and MTSS groups would be considered significant. Given that the Bonferroni adjustment is very conservative, the actual p-value for each test is somewhere between 0.005 and 0.05. Nonetheless, the two groups were subsequently combined into one Implementation group and compared with the Screening group.

T-tests were conducted to compare the total survey score, as well as the Measurement, Instruction, and Problem-Solving Teams strands for the two groups. (See Table 7.) In this case, Levene's test for equality of variances was not significant for the total score or any of the strands, so equal variance was assumed. There were no significant differences in scores with $t\ (97) = 0.80$, $p = 0.43$ (two-tailed); $t\ (131) = 1.31$, $p = 0.19$; $t\ (125) = 0.34$, $p = 0.74$; and $t\ (99) = 0.92$, $p = 0.36$, respectively for the total, Measurement, Instruction, and Problem-Solving Teams strands.

Correlation with Student Outcomes

The relationship between aggregate survey scores for each school and the student outcome measures at each grade level was investigated using Pearson product-moment correlation coefficients (see Table 8). After initially running the correlations, one significant relationship was found. There was a significant strong, negative correlation for the MAZE outcome measure at second grade, $r = -0.60$, $n = 16$, $p < 0.05$. The direction and strength of the relationship were unexpected, particularly given the lack of significant correlations between the survey score and the other outcome measures. Further examination revealed one growth score outlier that was impacting the correlation coefficient. With the outlying school removed, none of the correlation coefficients were significant.

Discussion

With the current focus on accountability and improving student outcomes, MTSSs offer a way to use data to improve student outcomes school-wide. However, it can be difficult to implement such a system effectively, and assessment tools are needed to help schools identify what they are or are not doing well. The purposes of this study were to evaluate the psychometric properties of an MTSS survey, to identify any group differences in level of implementation at different school sites, and to examine the relationship between level of implementation and student outcomes.

Internal Consistency of the Survey

The alpha coefficient was greater than 0.9, which meets even high reliability standards (Salvia, Ysseldyke & Bolt, 2010). Cronbach's alpha tends to increase as assessments become longer, so given the number of survey questions, the high alpha coefficient is unsurprising. However, there are drawbacks to longer surveys as well. The length of the survey may have contributed to the low response rate. Respondents were more likely to leave questions blank towards the end of the survey (from 7% up to 21%, see Table 3) which may indicate a fatigue effect. For future use by schools, the author recommends reducing the number of questions while still retaining a high reliability coefficient. Most questions were highly correlated with the total score, but given the number of questions, those with lower correlations could be omitted. For example, item 5, "Core curriculum is delivered with fidelity" only had a 0.46 correlation with the survey total, and if deleted, Cronbach's alpha would still be 0.98.

Interrater Agreement

Interrater agreement as measured by average measure ICCs was satisfactory (> 0.5) to excellent (>0.9) for all schools, which is indicative of strong interrater reliability. However, across *all* schools and raters, the ICC was still high (0.97). This can be interpreted as lack of ability to discriminate between schools that fully implemented a MTSS and those that did not.

Although the average measure ICCs were high, the single measure ICCs were considerably lower (< 0.5 for 18 of 21 schools). This indicates that multiple ratings are needed in order to obtain reliable ratings of school implementation using this survey. In their evaluation of measurement strategies, Muschkin et al (2007) recommended a minimum of 3 raters when using ICCs, a recommendation consistent with this study's findings as well. However, it is possible that higher interrater agreement would have been obtained had there been enough data to evaluate implementation at the grade- rather than school-level. Multiple ratings were obtained for schools, but the instructions given were slightly different. Teachers were asked to rate implementation for a specific grade-level, while principals and other staff were asked to rate implementation for the school as a whole. This may well have lowered rates of agreement.

Level of Implementation – Group Comparison

Across the three groups, mean differences in survey score were not found between the initial implementers (Pilot group), those who were implementing a full MTSS (MTSS group), and those who were only implementing a portion of the MTSS approach (Screening group). Clearly, despite the training received, schools may not have been effectively implementing the key components of a MTSS. However, there are a number of factors that may have contributed to this finding.

To check whether the non-significant ANOVA results were due to a lack of statistical power, a power analysis was run using GPower (Faul & Erdfelder, 1992) with power ($1 - \beta$) set at 0.80 and $\alpha = 0.05$, two-tailed. For an ANOVA using three groups, this analysis indicated that a sample size of 159 would be needed to detect a medium (0.25) effect size and a sample of 969 to detect a small (0.10) one. With three groups, 0.05 alpha, and 0.80 power, the sample size of 100 surveys obtained in this study was only able to detect an effect of 0.32 or greater, which would be considered a large effect size. However, the power analysis does not account for the non-independence of the statistical tests, which may substantially reduce the effective *p*-values and thus also the power level.

Collapsing the two implementation groups (Pilot and MTSS) into one group (Implementation) improved power somewhat. In this case, the statistical analysis is a t-test rather than an ANOVA, so the effect size metric is different (*d* rather than *f*). With two groups, 0.05 alpha, and 0.80 power, power analysis indicated the sample (N = 100) was only large enough to detect an effect size of 0.57 or greater. This would still be considered a medium to large effect size, but, as with the prior analysis, this does not take into account the impact of non-independent statistical tests on the *p*-values and power level.

Response rates, response styles, and bias are known issues with the use of surveys. There was a difference between groups in response rates with staff from schools within the MTSS group responding at a higher rate (40%) than those in either the Pilot (19%) or Screening (20%) schools. Respondents' comments in combination with analysis of the raw data indicate that the survey results may have been impacted by specific response styles, i.e., consistent responding, omitting items (including respondent fatigue as discussed earlier), and social desirability (Cronbach, 1946; Hancock & Flowers, 2001; Mundia, 2011). While response sets can actually increase reliability, they reduce validity (Cronbach, 1946) since they impair efforts to measure accurately the construct of interest. In future revisions of this survey, these types of response styles could be dealt with in different ways. For example, though the current design using the same response options for each of the survey questions has advantages in ease of responding, it clearly could contribute to the problem of bias due to consistent responding. The survey could be revised to eliminate this. Social desirability may be more difficult to address, but strategies, such as reverse coding of questions, have been developed to address and prevent this potential source of bias (Mundia, 2011).

In addition, lack of significant group differences may have been caused, at least in part, by the "contamination effect." That is, the lines were blurred somewhat between schools that were implementing MTSS and those that were only screening their students. This was evident in survey responses from the Screening schools indicating that they completed more than just screening, and was reflected in the AIMSweb data collected. For example, one school that was only assigned to conduct screening assessments went further and identified specific students who were progress monitored. So, while lack of implementation at the MTSS schools would be expected to cause lack of group differences, group differences could also be reduced by schools from the Screening group implementing aspects of a MTSS that went beyond what they were assigned.

There also may have been differences in schools' interpretation of what an MTSS is, and in the amount of training and information key personnel received. Schools did not consistently send the same staff to the MTSS trainings, nor did they always send the same number of people. In order to obtain a higher number of completed surveys, the pool of potential respondents was broadened. However, this may have resulted in surveys being completed by teachers who were not very involved in the MTSS implementation process. There were no questions in the survey to assess the respondent's level of knowledge about MTSSs or the implementation process at his/her school. Instead, all the questions asked about the extent of implementation, so there is no clear way to identify responses that may be more accurate than others. Objective data should be collected and integrated with the survey responses or used as a separate assessment to determine the accuracy of staff members' judgment.

The survey itself was completed by school personnel more than four months after the end of Year 2. Since then, administrative teams (principal, assistant principal, and English/Language Arts specialist) at three schools were moved and two schools in the district had been closed. This staff movement created some confusion among respondents as to which school they were evaluating. Future research should examine the use of this survey both during the school year and after. After analyzing test-retest reliability, this would allow schools and district personnel to look for improvement in the level of implementation throughout the course of the school year. Feedback on level of implementation throughout the year would make this survey a useful part of a school's professional development and training process.

Correlation with Student Outcomes

The lack of correlation between survey responses and student growth was not particularly surprising given all the possible influences on student outcomes at both the school and individual levels. However, there are some specific issues that may have impacted the results of this study. For example, due to the number of surveys completed, responses were aggregated across the entire school rather than being analyzed specifically by grade. However, it is not unusual for MTSS to be implemented differently at different grade levels, even within the same school, so using the entire school as the unit of measurement may have weakened the correlation.

There may also have been a problem with restriction of range. Not all schools could be included in this analysis since some of the schools (six out of 24) did not conduct either the fall or the spring screening. This could also be taken as evidence that these schools did not do a particularly good job of implementing a MTSS, since this critical component was not completed consistently. The lack of a difference in survey scores between these schools and those that did complete the spring screening would also seem to indicate a weakness in the survey's ability to discriminate between schools that are implementing effectively and those that are not.

Limitations

In addition to concerns related to the survey itself, other limitations should be considered in interpreting the results of this study. The district's implementation plan created a natural experiment. However, schools were not randomly assigned. On average, the schools in the Pilot group had a significantly higher enrollment than those in the MTSS group. Research on implementation at multiple sites has found that larger school size predicts lower implementation fidelity (Zvoch, 2009), though the reasons for this relationship have not been clearly identified.

Also, insufficient data were collected to convincingly argue that group differences or a change in student outcomes were due to schools' implementation of a MTSS. For example, two of the schools in the Pilot group and five of those in the MTSS group received grant funding during Year 2. While some of the funding was allocated towards purchase of AIMSweb access and MTSS-related expenses, other portions were earmarked for purchase of technology and assisting English learners with mastering both math and English/language arts standards. These other programs may have impacted literacy outcomes and weakened the correlation between MTSS implementation and student outcomes. To more effectively demonstrate the effectiveness of MTSSs in the absence of random assignment, additional information on implementation would need to be collected.

Fidelity. Typically, fidelity checks are a critical part of efforts to evaluate implementation (e.g., treatment fidelity; Kovaleski, 2007; Mahdavi & Beebe-Frankenberger, 2009). To study MTSS implementation, all areas of the system would need to be evaluated. Core instruction (Tier 1) affects all students, and its quality can be expected to vary across schools and classrooms. This area needs to be observed and analyzed to take into account any group differences that might affect student outcomes regardless of implementation of the other aspects of a MTSS (e.g., Tier 2 intervention). Data would also need to be collected to rule out the possibility of unreliability in the student outcome measures (i.e., AIMSweb screening data). A very limited sample of interrater agreement data was collected, and different schools completed the screening at different times.

There is a need for further research in this area and for the use of instruments that can aid schools in evaluating their implementation efforts. Though the survey used in this study demonstrated good reliability, no external measures were available for evaluating criterion validity, and no significant group differences were identified using this tool. In future studies, this instrument should be streamlined to improve response and completion rates, questions should be included based on their ability to differentiate, and the survey's relationship with implementation outcomes, such as collection of screening data, occurrence of problem-solving team meetings, or use of progress monitoring data should be evaluated.

References

Al Otaiba, S., Folsom, J. S., Schatschneider, C., Wanzek, J., Greulich, L., Meadows, J., … Connor, C. M. (2011). Predicting first-grade reading performance from kindergarten response to Tier 1 instruction. *Exceptional Children, 77*(4), 453-47.

Al Otaiba, S., Puranik, D., Rouby, A., Greulich, L., Sidler, J. F., & Lee, J. (2010). Predicting kindergarteners' end-of-year spelling ability based on their reading, alphabetic, vocabulary, and phonological awareness skills, as well as prior literacy experiences. *Learning Disability Quarterly, 33*(3), 171-183.

Barber, M. (2004). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education, 185*(1), 7-38.

Batsche, G. M., Kavale, K. A., & Kovaleski, J. F. (2006). Competing views: A dialogue on Response to Intervention. *Assessment for Effective Intervention, 32*(1), 6-19. doi: 1.1177/15345084060320010301

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of Responsiveness-to-Intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23*, 381-394. doi: 10.1177/073428290502300406

Burns, M. K., & Symington, T. (2002). A meta-analysis of prereferral intervention teams: Student and systemic outcomes. *Journal of School Psychology, 40*(5), 437-447.

Burns, R. & Burns, R. (2008). *Business research methods and statistics using SPSS.* Thousand Oaks, CA: Sage Publications Inc. Retrieved from http://www.uk.sagepub.com/burns/chapters.htm

Chaparro, E. A., Smolkowski, K., Baker, S. K., Hanson, N., & Ryan-Jackson, K. (2012). A model for system-wide collaboration to support integrated social behavior and literacy evidence-based practices. *Psychology in the Schools, 49*(5), 465-482. doi: 10.1002/pits.21607

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Compton Unified School District. (2009). *Report to our community.* Retrieved from http://www.compton.k12.ca.us/

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494. doi: 10.1177/001316444600600405

Denton, C. A., Fletcher, J. M., Simos, P. G., Papanicolaou, A. C., & Anthony, J. L. (2007). An implementation of a tiered intervention model: Reading outcomes and neural correlates. In D. Haager, J. Klingner, & S. Vaughn (Eds.), *Evidence-based reading practices for Response to Intervention* (pp.107-138). Baltimore, MD: Paul H. Brooks Publishing Co.

Duhon, G. J., Mesmer, E. M., Gregerson, L., & Witt, J. C. (2009). Effects of public feedback during RTI team meetings on teacher implementation integrity and student academic performance. *Journal of School Psychology, 47*, 19–37. doi:10.1016/j.jsp.2008.09.002

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes. *Educational and Psychological Measurement, 61*(4), 517-531. doi: 1.1177/0013164401614001

Faul, F. & Erdfelder, E. (1992). *GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS* [Computer program]. Bonn, FRG: Bonn University, Dep. of Psychology.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Intervention research: A synthesis of the literature.* Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network (FMHI Publication #231).

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

Florida Problem-Solving/Response to Intervention Statewide Project. (2007). *Beliefs survey.* Retrieved from http://floridarti.usf.edu/

Fuchs, D., Fuchs, L., Simmons, D., & Mathes, P. (2008). *Peer-Assisted Learning Strategies reading methods for grades 2-6.* Nashville, TN: Vanderbilt University.

Fuchs, D., Fuchs, L., Svenson, E., Yen, L., Thompson, A., McMaster, K., … Braun, M. (2011). *Peer-Assisted Learning Strategies first grade reading teacher manual.* Nashville, TN: Vanderbilt University.

Fuchs, D., Fuchs, L., Thompson, A., Al Otaiba, S., Yen, L., McMaster, K., … Braun, M. (2008). *Peer-Assisted Learning Strategies kindergarten reading teacher manual.* Nashville, TN: Vanderbilt University.

Fuchs, L. S., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. *Assessment for Effective Intervention, 29*(4), 25-37.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking Response to Intervention at middle and high school. *School Psychology Review, 39*(1), 22-28.

Graves, A. W., Brandon, R., Duesbery, L., McIntosh, A., & Pyle, N. B. (2011). The effects of Tier 2 literacy instruction in sixth grade: Toward the development of a Response-to-Intervention model in middle school. *Learning Disability Quarterly, 34*(1), 73-86.

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, L., . . . Flindt, N. (2004). *Technical adequacy and decision making utility of DIBELS* (Tech. Rep. No. 7). Eugene, OR: University of Oregon.

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., & Martin, D. A. (2011). An investigation of silent versus aloud reading comprehension of elementary students using maze assessment procedures. *Psychology in the Schools, 48*(1), 4-13. doi: 10.1002/pits.20543

Hancock, D. R. & Flowers, C. P. (2001). Comparing social desirability responding on World Wide Web and paper-administered surveys. *Educational Technology Research and Development, 49*(1), 5-13.

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features.* Retrieved from https://aimsweb.pearson.com/downloads/PassagesTechnicalManual08252002.pdf

Individuals with Disabilities Education Improvement Act of 2004, P.L. 108-446 (2004). Retrieved from http://idea.ed.gov/download/statute.html

Juel., C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*(4), 437-447.

Kame'enui, E. J. (2002). *Executive summary of final report on Reading First reading assessment analysis.* Retrieved from AIMSweb: http://www.aimsweb.com

Kerins, M. R., Trotter, D., & Schoenbrodt, L. (2011). Effects of a Tier 2 intervention on literacy measures: Lessons learned. *Child Language Teaching and Therapy, 26*(3), 287-302. doi: 1.1177/0265659009349985

Knotek, S. E. (2005). Sustaining RTI through consultee-centered consultation. *The California School Psychologist, 10,* 93-104.

Kovaleski, J. F. (2007). Response to Intervention: Considerations for research and systems change. *School Psychology Review, 36*(4), 638-646.

Kratochwill, T. R., & Bergan, J. R. (1990). *Behavioral consultation in applied settings: An individual guide.* New York, NY: Plenum Press.

Linan-Thompson, S., & Vaughn, S. (2010). Evidence-based reading instruction: Developing and implementing reading programs at the core, supplemental, and intervention levels. In G. G. Peacock, R. A. Ervin, E. J. Daly, & K. W. Merrell (Eds.), *Practical handbook of school psychology: Effective practices for the 21st century* (pp. 274-286). New York, NY: Guilford.

Mahdavi, J. N., & Beebe-Frankenberger, M. E. (2009). Pioneering RTI systems that work: Social validity, collaboration, and context. *Teaching Exceptional Children, 42*(2), 64-72.

Mundia, L. (2011). Social desirability, non-response bias and reliability in a long self-report measure: Illustrations from the MMPI-2 administered to Brunei student teachers. *Educational Psychology: An International Journal of Experimental Educational Psychology, 31*(2), 207-224.

Muschkin, C. G., Malone, P. S., & Conduct Problems Prevention Research Group. (2007): Multiple teacher ratings: An evaluation of measurement strategies. *Educational Research and Evaluation: An International Journal on Theory and Practice, 13*(1), 71-86. doi:1.1080/13803610601058215

National Association of State Directors of Special Education. (2008a). *Response to Intervention blueprints for implementation: District level edition.* Retrieved from http://http://www.nasdse.org/

National Association of State Directors of Special Education. (2008b). *Response to Intervention blueprints for implementation: School building level edition.* Retrieved from http://http://www.nasdse.org/

National Center for Education Statistics (2011). *The Nation's Report Card: Reading 2011* (NCES 2012–457). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Center on Response to Intervention. (2009). *RTI classification tool and resource locator self- assessment.* Retrieved from http://www.rtictrl.org/selfassessment/

National Center on Response to Intervention. (2010). *Essential components of RTI: A closer look at Response to Intervention.* Retrieved from http://www.rti4success.org/pdf/rtiessentialcomponents_04271.pdf

No Child Left Behind Act of 2001, P.L. 107-110 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-11.pdf

Pallant, J. (2007). *SPSS Survival Manual* (3rd ed.). New York, NY: McGraw-Hill Education.

Pearson. (2012). *AIMSweb technical manual (R-CBM and TEL).* Prepublication version, Eden Prairie, MN: NCS Pearson, Inc.

Powers, K. Hagans, K., & Busse, R. T. (2008). School psychologists as instructional consultants in a Response-to-Intervention model. *The California School Psychologist, 13*, 41-53.

Ritchey, K. D. (2008). Assessing letter sound knowledge: A comparison of Letter Sound Fluency and Nonsense Word Fluency. *Exceptional Children, 74*(4), 487-506.

Roberts, M. J., & Russo, R. (1999). *A student's guide to analysis of variance.* New York, NY: Routledge.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuele, D. P., & Sweeney, J. M. (2005). Severe reading difficulties—can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality, 13*(4), 209-227. doi:1.1207/s15327035ex1304_3

Shapiro, E. S., & Clemens, N. H. (2009). A conceptual model for evaluating system effects of Response to Intervention. *Assessment for Effective Intervention, 35*(1), 3-16. doi: 1.1177/1534508408330080

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*(3), 164-172.

Shinn, M. M., & Shinn, M. R. (2002a). *AIMSweb training workbook: Administration and scoring of early literacy measures for use with AIMSweb.* Eden Prairie, MN: Edformation, Inc. Retrieved from https://aimsweb.pearson.com/downloads/AdminandScoringEarlyLiteracy04012003.pdf

Shinn, M. M., & Shinn, M. R. (2002b). *AIMSweb training workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement.* Retrieved from https://aimsweb.pearson.com/downloads/AdminandScoringR-CBM09292005.pdf

Shinn, M. R., & Shinn, M. M. (2002c). *AIMSweb training workbook: Administration and scoring of reading maze for use in general outcome measurement.* Retrieved from https://aimsweb.pearson.com/downloads/AdminandScoringReadingMAZE10292002.pdf

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*, 735-749.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Tanguma, J. (2000, November). *A review of the literature on missing data.* Paper presented at the Annual Meeting of the Midwest Educational Research Association, Bowling Green, KY.

Tilly, W. D. (2008). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology v* (pp. 17-36). Bethesda, MD: National Association of School Psychologists.

Torgesen, J. K. (2009). The Response to Intervention instructional model: Some outcomes from a large-scale implementation in Reading First schools. Child *Development Perspectives, 3*(1), 38-40.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40*(1), 7-26.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse (2009). *Assisting students struggling with reading: Response to intervention and multi-tier intervention in the primary grades* (NCEE 2009–4045). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/rti_reading_pg_021809.pdf

U.S. Department of Education, Office of Special Education and Rehabilitative Services, President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families.* Retrieved from http://www.ed.gov/inits/commissionsboards/whspecialeducation/

U.S. Department of Education, Office of Special Education and Rehabilitative Services (2007). *Questions and answers on Response to Intervention (RTI) and Early Intervening Services (EIS)*. Retrieved from http://idea.ed.gov/explore/view/p/%2Croot%2Cdynamic%2CQaCorner%2C8%2C

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*(2), 225-256. doi:1.1016/j.jsp.2006.11.004

Vaughn, S. & Klingner, J. (2007). Overview of the three-tier model of reading intervention. In D. Haager, J. Klingner, & S. Vaughn (Eds.), *Evidence-based reading practices for Response to Intervention* (pp. 3-10). Baltimore, MD: Paul H. Brooks Publishing Co.

Vaughn, S., & Fuchs, L. S. (2006). A response to "Competing views: A dialogue on Response to Intervention" : Why Response to Intervention is necessary but not sufficient for identifying students with learning disabilities. *Assessment for Effective Intervention. 32*(1), 58-61. doi: 1.1177/15345084060320010801

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R, & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601-638.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541-561.

Wuensch, K. L. (2007). *Inter-rater agreement.* Retrieved from East Carolina University: http://core.ecu.edu/psyc/wuenschk/statslessons.htm

Table 1

School Enrollment and Ethnicity

	Mean (SD)
Pilot	
Enrollment	747 (187)
% Hispanic	88 (3)
% African-American	6 (2)
% Not Reported	5 (2)
MTSS	
Enrollment	475 (186)
% Hispanic	75 (20)
% African-American	18 (15)
% Not Reported	7 (5)
Screening	
Enrollment	519 (152)
% Hispanic	72 (12)
% African-American	17 (8)
% Not Reported	10 (5)

Note. Each school had less than 1% White, Asian, American-Indian or bi/multiracial students.

Table 2

Staff Characteristics

	N	%
Position	148	100
Teacher	108	73
Resource Specialist	16	11
Principal/Assistant Principal	22	15
Other (Substitute, Instructional Aide)	2	1
Age	142	96
20-29	2	1
30-39	61	41
40-49	47	32
50-59	20	14
60 or older	12	8
Gender	146	99
Male	29	20
Female	117	79
Years in Current Role	147	99
Less than one year.	6	4
One to two years.	12	8
Three to five years.	24	16
Six to 10 years.	65	44
More than 10 years.	40	27
Group	147	99
Pilot	24	16
MTSS	81	56
Screening	41	28
Grade	146	99
Kindergarten	17	12
First	18	12
Second	17	12
Third	21	14
Fourth	21	14
Fifth	17	12
School as a whole	35	24

Table 3

Item-Total Statistics

Item	Valid	Missing	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	138	10	0.59	0.98
2	136	12	0.77	0.98
3	138	10	0.60	0.98
4	135	13	0.68	0.98
5	138	10	0.46	0.98
6	138	10	0.63	0.98
7	136	12	0.77	0.98
8	131	17	0.67	0.98
9	135	13	0.65	0.98
10	132	16	0.61	0.98
11	134	14	0.65	0.98
12	135	13	0.66	0.98
13	135	13	0.69	0.98
14	135	13	0.74	0.98
15	126	22	0.82	0.98
16	126	22	0.72	0.98
17	126	22	0.79	0.98
18	125	23	0.80	0.98
19	124	24	0.85	0.98
20	123	25	0.85	0.98
21	125	23	0.76	0.98
22	126	22	0.76	0.98
23	109	39	0.85	0.98
24	125	23	0.78	0.98
25	127	21	0.77	0.98
26	123	25	0.89	0.98
27	125	23	0.86	0.98
28	122	26	0.81	0.98
29	125	23	0.83	0.98
30	126	22	0.84	0.98
31	120	28	0.83	0.98
32	122	26	0.86	0.98
33	119	29	0.86	0.98
34	118	30	0.80	0.98
35	117	31	0.75	0.98
36	118	30	0.76	0.98
	118	30	0.84	0.98

Table 4

Intraclass Correlation (ICC) by School

School #	# of Raters	ICC	95% CI
1	3	0.94	[0.78, 1.00]
2	5	0.93	[0.80, 0.99]
3	3	0.60	[-0.54, 0.99]
4	1	n/a	
5	7	0.94	[0.85, 0.99]
6	10	0.87	[0.72, 0.96]
7	2	0.89	[0.43, 1.00]
8	7	1.00	[0.99, 1.00]
9	7	0.95	[0.88, 0.99]
10	1	n/a	
11	3	0.96	[0.84, 1.00]
12	2	0.95	[0.75, 1.00]
13	3	0.96	[0.86, 1.00]
14	0	n/a	
15	10	0.99	[0.98, 1.00]
16	3	1.00	[0.99 1.00]
17	3	0.97	[0.88, 1.00]
18	3	0.91	[0.65, 1.00]
19	3	0.94	[0.78, 1.00]
20	4	0.99	[0.97, 1.00]
21	2	0.97	[0.85, 1.00]
22	2	0.93	[0.64, 1.00]
23	11	0.96	[0.92, 0.99]
24	4	0.97	[0.90, 1.00]

Table 5

Group Differences on Survey

Survey Strand	*N*	Pilot	MTSS	Screening	*F*	*p*
Measurement	133	(n = 24) 17.50 (2.60)	(n = 72) 17.68 (3.47)	(n = 37) 16.78 (3.60)	$F(2, 130) = 0.88$	0.42
Instruction	127	(n = 23) 40.13 (6.68)	(n = 67) 41.70 (8.92)	(n = 37) 40.76 (7.82)	$F(2, 124) = 0.37$	0.69
Problem-Solving Team	101	(n = 19) 96.79 (14.92)	(n = 54) 92.31 (27.75)	(n = 28) 88.32 (25.55)	$F(2, 98) = 0.64$	0.53
Total	99	(n = 19) 154.58 (20.31)	(n = 53) 150.45 (39.29)	(n = 27) 145.11 (36.98)	$F(2, 96) = 0.41$	0.67

Table 6

Pre-Implementation Comparison between Pilot and MTSS Groups

Grade	Measure	Pilot Mean (SD)	MTSS Mean (SD)	DF	*t*	*p*
K	LNF	7.75 (5.91)	18.83 (5.60)	8	-3.00*	0.02
1	NWF	27.75 (9.32)	37.57 (5.86)	9	-2.18	0.06
2	MAZE	2.67 (1.16)	5.50 (2.17)	7	-2.07	0.08
	R-CBM	45.00 (10.07)	62.00 (7.95)	8	-2.99*	0.02
3	MAZE	9.00 (2.65)	10.29 (2.81)	8	-0.67	0.52
	R-CBM	70.00 (4.58)	81.57 (11.41)	8	-1.65	0.14
4	MAZE	8.67 (1.53)	10.29 (2.63)	8	-0.98	0.36
	R-CBM	89.33 (4.51)	97.43 (4.43)	8	-2.65*	0.03
5	MAZE	10.33 (1.16)	13.71 (3.25)	8	-1.71	0.13
	R-CBM	96.67 (6.81)	111.86 (6.04)	9	-3.53**	<0.01

* $p < 0.05$
** $p < 0.01$

Table 7

Group Differences on Survey – Two-Group Analysis

Survey Strand	*N*	Implementation	Screening	df	*t*	*p*
Measurement	133	(n = 96) 17.64 (3.26)	(n = 37) 16.78 (3.60)	131	1.31	0.19
Instruction	127	(n = 90) 41.30 (8.40)	(n = 37) 40.76 (7.82)	125	0.34	0.74
Problem-Solving Team	101	(n = 73) 93.48 (25.03)	(n = 28) 88.32 (25.55)	99	0.92	0.36
Total	99	(n = 72) 151.54 (35.19)	(n = 27) 145.11 (36.98)	97	0.80	0.43

Table 8

Correlation between Aggregate Student Growth Scores and Survey Scores

Grade	Measure	N	Scale Total
K	LNF	17	-0.22
1	NWF	18	0.02
2	MAZE	15	-0.01
	R-CBM	18	-0.11
3	MAZE	13	0.26
	R-CBM	16	0.06
4	MAZE	15	0.02
	R-CBM	18	-0.07
5	MAZE	16	0.22
	R-CBM	18	0.13

Appendix:

Survey Questions and Response Choices

Response options:

Answer choices

Yes, done consistently and in accordance with district guidelines. No further help needed.

Yes, but with some exceptions. A plan is in place to improve.

Yes, but with some exceptions. We need help creating a plan for improvement.

No, but a plan is in place to improve.

No, and we need help to improve implementation.

A**nswer choices from original survey**

Got it! No further help needed

Got part of it. No help needed at this time

Got part of it. Would like more support

Don't got it, but don't want support at this time

Don't got it. Would like more support

STRAND 1: MEASUREMENT

1. Screening data are taken 3 times per year.
2. Students receiving intervention are progress monitored, with data at least every 2-3 weeks.

3. Data collected in AIMSweb are provided to teachers in a timely manner.
4. Data are used to help in instructional planning

STRAND 2: CURRICULUM AND INSTRUCTION

Area 1: Curriculum

5. Core curriculum is delivered with fidelity.
6. Teachers are using universal access time to support “some risk” and “at risk” students in core curriculum.
7. The school has a system to evaluate the effectiveness of core (Tier I), supplement (Tier II), and intensive (Tier III) programs.

STRAND 2: CURRICULUM AND INSTRUCTION

Area 2: Instruction

8. Teachers understand the five components of reading and how they interrelate.
9. Teachers regularly use progress-monitoring data to inform their instructional practices and differentiate instruction.
10. Grade-level teachers meet at least monthly to review student progress, make decisions about resources and interventions.
11. Instructional groups are formed based on student need using flexible grouping options.
12. The school allows for a flexible reading scheduling to stagger reading instruction for students who need more time.

13. Reading coaches and/or other support staff meet with teachers with students who do not respond to the general curriculum.
14. The extra instructional time provided to struggling students is targeted to specific reading skills deficits.

STRAND 3: Problem Solving Teams

Area 1: Team Characteristics

15. The problem-solving team(s) have balanced representation of grade level, general and special education staff.
16. An administrator is a member of the problem-solving team.
17. There a regularly scheduled meeting time and place.
18. The team has forms used at the meeting to lead the team through the problem-solving process.
19. Data are regularly collected on team functioning (e.g., students served).

STRAND 3: Problem Solving Teams

Area 2:Problem Identification

20. When multiple problems are identified, the problem-solving team prioritizes them.
21. The team uses a general education database to indentify and define problems (e.g., AIMSweb, benchmarks).
22. The data collected during the problem identification stage are displayed in a graphic or summary format.

23. There are procedures for addressing severe problems in a timely manner.

STRAND 3: Problem Solving Teams

Area 3:Plan Development

24. The intervention plan(s) is supported by research.
25. The intervention is linked to the assessment.
26. The problem-solving team identifies the goal of an intervention plan (who, what, where, when) and it is provided to all team members.
27. There is a system in place to collect frequent on-going data to determine if the plan is working.
28. Data collected to evaluate the plan are displayed in a graphic format.
29. There a commitment to continue an intervention, as prescribed in the plan, until a team decision is made to discontinue it.
30. There is a system in place to communicate the on-going results of the intervention plan with teachers and parents.

STRAND 3: Problem Solving Teams

Area 4:Plan Implementation

31. Problem-solving team members commit to an evaluation of whether the intervention is being implemented as planned (fidelity checks).
32. There is a procedure for providing the teacher with support if the plan is not being implemented as described.

33. Student progress towards the identified goal is evaluated on a regular basis, as described.

STRAND 3: Problem Solving Teams

Area 5:Plan Evaluation

34. There is an agreed upon timeline for plan evaluation.
35. When a plan has not been successful, it recycles through the problem solving process.
36. When a plan is effective, decisions are made about fading the intervention.
37. There are criteria for determining when a child's needs exceed the resources of the problem-solving team and special education eligibility is considered.