MEASURING INTERVENTION IMPLEMENTATION:

EXAMINING THE RELATION OF DIMENSIONS OF IMPLEMENTATION TO STUDENT OUTCOMES IN A KINDERGARTEN MATHEMATICS INTERVENTION

by

TASIA L. BRAFFORD

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2021

DISSERTATION APPROVAL PAGE

Student: Tasia L. Brafford

Title: Measuring Intervention Implementation: Examining the Relation of Dimensions of Implementation to Student Outcomes in a Kindergarten Mathematics Intervention

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Elizabeth Harn, PhD	Chairperson
Benjamin Clarke, PhD	Core Member
Christian Doabler, PhD	Core Member
Derek Kosty, PhD	Core Member
Kathleen Scalise, PhD	Institutional Representative

and

Andy Karduna	Interim Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded September 2021.

DISSERTATION ABSTRACT

Tasia L. Brafford

Doctor of Philosophy

Department of Special Education and Clinical Sciences

September 2021

Title: Measuring Intervention Implementation: Examining the Relation of Dimensions of Implementation to Student Outcomes in a Kindergarten Mathematics Intervention

Progress monitoring in multi-tiered systems of support typically focus on student progress, but it is also important to consider information regarding implementation of instruction and evidence-based practices. Through the assessment of implementation, researchers and practitioners are better equipped to understand the true effects of an intervention and make changes to improve intervention instruction. Adherence data is typically collected in research, whereas implementation data in the school setting are collected inconsistently, if at all (Sanetti & Luh, 2019). Research suggests that different dimensions of implementation may relate to student outcome differentially based on content and context (Boardman et al., 2016; Odom et al., 2010). By investigating the different dimensions of implementation and their relation to each other as well as student outcomes within intervention settings, researchers and practitioners will gain knowledge regarding contextual factors that support improved intervention instruction.

The current study contributed to this research literature using data collected from the ROOTS efficacy trial (Clarke et al., 2012) to answer the following research questions: (1) What are the descriptive statistics of treatment adherence and *Ratings of Classroom Management and Instructional Support* (RCMIS)? Do these statistics differ by treatment group size? What are the underlying dimensions of instructional quality of measured by RCMIS?, (2) Is treatment adherence associated with factors of implementation quality?, and (3) Which measure of implementation (e.g., dimensions of instructional quality, treatment adherence) accounts for the most variance in student outcomes? Participants included 880 at-risk students assigned to ROOTS treatment intervention instruction, comprising 255 ROOTS groups. Descriptive statistics, exploratory factor analysis, regression analysis, and hierarchical linear modeling was used to answer the research questions.

Results indicated high levels of treatment adherence and RCMIS score across time, with no statistical differences across group sizes. A single factor of implementation quality was found when investigating the dimensions of the RCMIS. For the distal measure of student achievement, neither treatment adherence nor the RCMIS was found to relate to student outcomes, whereas both were similarly related to the proximal measure of student mathematics achievement. Results are discussed in the context of implementation in both research and practice.

CURRICULUM VITAE

NAME OF AUTHOR: Tasia L. Brafford

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon; Eugene, Oregon
Baylor University; Waco, Texas
The College of Saint Rose; Albany, New York

DEGREES AWARDED:

Doctor of Philosophy, Special Education, 2021, University of Oregon
Master of Science in Education, Educational Psychology, 2016, Baylor University
Bachelor of Science, Childhood and Special Education, 2009, The College of Saint Rose

AREAS OF SPECIAL INTEREST:

Mathematics intervention and assessment
Dually endorsed teacher preparation
Inclusive practices

PROFESSIONAL EXPERIENCE:

Graduate Research Assistant, Center on Teaching and Learning, Eugene, OR, 2019-2021
Special Education Teacher, Waco Independent School District, Waco, TX 2015-2016
ABA Therapist, Baylor Center for Assessment, Research, & Evaluation, Waco, TX, 2014-2016
Special Education Teacher, Killeen Independent School District, Killeen, TX, 2013-2015
Special Education Teacher, Florence Independent School District, Florence, TX, 2012-2013
Special Education Teacher, KIPP: Tech Valley Charter School, Albany, NY, 2011-2012

GRANTS, AWARDS, AND HONORS:

AREA Institute on Statistical Analysis Invited Participant, 2020
CLD Leadership Academy Selected Participant, 2019

TED Kaleidoscope Student Research Poster Award for Quantitative Research Recipient, Teacher Education Division of Council of Exceptional Children, 2018
CLD Leadership Institute Selected Participant, 2018
Weatherby Scholarship, Baylor University, 2015, 2016
Educational Psychology Department Scholarship, Baylor University, 2014, 2015, 2016
Teacher of the Week, Indian Spring Middle School, Waco Independent School District, 2016
Summa Cum Laude Graduate. The College of Saint Rose, 2009

PUBLICATIONS:

Doabler, C.T., Clarke, B., Kosty, D., Turtura, J., Firestone, A., Smolkowski, K., Jungjohann, K., Brafford, T. L., Nelson, N. J., Sutherland, M., Fien, H., & Maddox, S. (2019). Efficacy of a first-grade mathematics intervention on measurement and data analysis. *Exceptional Children, 86*, 77-94. https://doi.org/10.1177/0014402919857993

Doabler, C.T., Clarke, B., Firestone, A., Turtura, J., Jungjohann, K., Brafford, T., Sutherland, M., Nelson, N.J., & Fien, H. (2019). Applying the curriculum research framework in the design and development of a technology-based tier 2 mathematics intervention. *Journal of Special Education Technology, 34*, 176-189. https://doi.org/10.1177/0162643418812051

ACKNOWLEDGEMENTS

I would like to thank my advisor, Beth Harn, for her commitment to my success and belief in my abilities throughout this dissertation process and my graduate training. Her support and encouragement have been invaluable for my growth as a professional and a researcher. I would like to thank Ben Clarke for his mentorship throughout my time at the University of Oregon, even though I'll never be a school psychologist. I also would like to express my gratitude to Derek Kosty for his patience and guidance throughout this dissertation process, as well as to Kathleen Scalise who was integral in developing ideas that were meaningful to me as a teacher and researcher. I also wish to thank Chris Doabler for his willingness to serve on my committee and seeing the potential in me as a scholar. I also want to express my gratitude to my friends and family for their faith in me and support over the years. I am forever grateful to my best friend, my husband, Tyler for his support and love throughout the years. This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120304 to the Center on Teaching and Learning at the University of Oregon.

For Nicolas who taught me I am stronger than I think.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF TABLES

CHAPTER I: LITERATURE REVIEW

Measuring implementation in multi-tiered systems of support (MTSS) is multi-faceted and complex. The focus on student progress through regular progress monitoring often overlooks an important aspect of MTSS: the monitoring of implementation. Through investigations of implementation, researchers and practitioners can capture a better view of the intervention and the specific elements of Tier 2 instruction that are relevant to student outcomes can be examined. By assessing implementation, researchers are better able to describe the true effect of the intervention on student outcomes and educators are better able to make valid decisions about students' intervention responsiveness and support interventionists' instructional improvement (Sanetti & Luh, 2019). Evaluation of implementation is considered crucial to the investigation of program effects, but research indicates that implementation is often not perfect (Hill & Erickson, 2019), particularly in school settings.

Fidelity of Implementation

Sanetti and Luh (2019) provided an overview to the field of learning disabilities on the importance of implementation science within MTSS. The goal of implementation science is to encourage the successful implementation of evidence-based practices and evaluation of these practices to continually inform implementation in the real world setting (Nilsen, 2015; Smolkowski et al., 2019). Sanetti and Luh (2019), in an implementation science special issue of *Learning Disability Quarterly*, noted that implementation data within MTSS is often inconsistent, or worse, completely lacking. However, investigations of these data are critical because implementation can impact student achievement (Sanetti & Collier-Meek, 2019; Sanetti & Luh, 2019). In a national survey of school psychologists, Cochrane et al. (2019) asked if school-based problem-solving teams measured treatment integrity, implementation, to inform decisions about students' responsiveness to intervention and found that these data were measured "most of the time" by only 14% of respondents. These findings represent a need to investigate the important aspects of implementation relevant to student success.

To investigate implementation, one must first gain an understanding of what implementation entails. Different models and conceptualizations of implementation have been presented in the field, and will be described in chronological order in the following sections. Historically, implementation fidelity (e.g., fidelity of implementation, treatment fidelity, intervention implementation, program integrity, etc.) was defined as "the degree to which programs were implemented as planned" by Dane and Schneider (1998, p. 23) in their seminal article on the topic. Through this article, five dimensions of implementation fidelity were outlined: (a) adherence, (b) quality, (c) exposure (also referred to in the literature as duration or dosage), (d) participant responsiveness, and (e) program differentiation (Dane & Schneider, 1998). Implementation fidelity has been found to be a significant factor in student outcomes (National Research Council, 2004; O'Donnell, 2008). In an examination of the underlying factors of implementation fidelity, Fogarty et al. (2014) found that a single fidelity factor comprised of adherence, quality, dosage, student responsiveness, and program differentiation was statistically related to student literacy outcomes. Whereas others have focused solely on adherence and its relation to student outcomes (Biggs et al., 2019; Meza et al., 2020), understanding how and when specific elements of implementation relate to outcomes can inform intervention development and support schools in delivering evidence-based practices (Sanetti & Collier-Meek, 2019; Sanetti & Luh, 2019).

Differing Conceptualizations of Implementation

In their review of education and health literature to determine the best practice in developing and measuring fidelity, Mowbray et al. (2003) defined fidelity of implementation as "the extent to which delivery of an intervention adheres to the protocol or program model originally developed" (p. 315). They proposed specifying fidelity through aspects of both structure and process measures of the program or intervention. Structure measures of a program require less subjectivity than process measures and could include aspects such as checklists evaluating one's adherence to the program protocol. Process measures include interactions between the program staff and clients or across clients, treatment delivery, emotional climate, or program style and often are evaluated through rating scales. Mowbray et al. (2003) stated that evaluation of implementation should include a balance of process and structure measures to better capture the true state of program implementation.

Carroll et al. (2007) proposed a conceptual framework of implementation fidelity relating the dimensions outlined by Dane and Schneider (1998) and Mowbray et al. (2003). This conceptual model is based in the field of molecular sciences and focuses on the provision of an intervention to clients. The model included adherence and other structural aspects of fidelity of implementation (i.e., content, coverage, frequency, and duration) as being an intermediate influence from intervention to student outcomes. Potential moderators were also included in Carroll et al.'s (2007) model. These moderators were process in nature (i.e., policy, strategies to facilitate implementation, quality of delivery, and participant responsiveness) and are said to give stakeholders a better understanding of how and why an intervention works or does not. They argued that the degree to which full adherence is achieved may be moderated by factors affecting the delivery process. This model outlined the relation between the elements of implementation during an intervention and student outcomes, but has yet to be empirically tested within a large-scale intervention (Carroll et al., 2007).

More recently, O'Donnell (2008) conducted a review of the literature in K-12 curriculum research and outlined fidelity to structure measures to include adherence and duration and fidelity to process measures to include program differentiation and quality of delivery. This conceptualization of implementation focuses on general education service provision, with all included articles providing core instruction to students in kindergarten through 12th grade. O'Donnell (2008) hypothesized that participant responsiveness incorporated aspects of both fidelity to structure and process measurement. Across the 23 empirical articles included in the review, there were at least six definitions of fidelity of implementation. These definitions all addressed adherence or integrity of the intervention, such as focusing on the extent to which a program was implemented as proposed (i.e., adherence) or the 'faithful' implementation of an intervention in practice (i.e., integrity), though the wording for the definitions differed slightly across articles. The author concluded that the inconsistencies in defining implementation as a construct make the measurement of implementation challenging (O'Donnell, 2008). Finding only five out of the 23 included articles investigating the relation between implementation and student outcomes, O'Donnell (2008) called for future research that thoroughly investigates how implementation can impact student achievement in intervention settings.

In a review of high impact general and special education journals, Swanson et al. (2013) also found inconsistent rates of reporting fidelity, with less than 70% of their 76 reviewed articles reporting on some aspect of fidelity. Observations were most often used across both general and special education journals to evaluate fidelity, with self-administered checklists and intervention dosage the next most common avenues to evaluate fidelity (Swanson et al., 2013). Quality of implementation was reported in less than 10% of the articles, indicating that this aspect of fidelity of implementation is often neglected in the research. Of the ten articles reporting on mathematics only interventions, only one reported on the quality of implementation. The authors called for future research to include quality of implementation data, as they and others (Gersten et al., 2005) contend that quality of implementation data may provide important insights on the relation between the intervention and student outcomes that is not captured by adherence data alone (Swanson et al., 2013).

Capin et al. (2018) informed the field of special education on implementation in reading intervention by completing a review of the reading intervention literature from 1995 through 2015. This review investigated how treatment fidelity is measured in the intervention literature with at-risk student populations in kindergarten through 3rd grade and included 175 studies. They reported that less than half of the reading intervention studies reported treatment fidelity data either qualitatively or quantitatively; those that did report treatment fidelity data focused on adherence to the program, which was high on average (Capin et al., 2018). Most of the reviewed articles used observations to collect implementation data, which aligns with recommendations to the field (Gersten et al., 2005). Consistent with O'Donnell's (2008) findings from ten years prior, they found that few studies related aspects of implementation or treatment fidelity of any kind to student outcomes in their analyses. This recent article underscores the importance of considering dimensions of implementation when discussing the impact of an intervention on student outcomes. Dimensions of implementation are an important variable in the delivery of these interventions that is being captured by researchers reporting fidelity, but not being used in the analyses (Capin et al., 2018).

Also from the field of special education, DeFouw et al. (2019) recently conducted a review of treatment fidelity factors in mathematics intervention research from 2004 to 2015. Across the 66 articles reviewed, 43 reported some quantitative aspect of implementation fidelity with high treatment fidelity reported. As with previous reviews (Capin et al., 2018), direct observations were most often used to evaluate implementation and adherence was the most assessed aspect of implementation. Whereas adherence was assessed across 88% of studies, quality was assessed in only 17% of studies and participant responsiveness was never assessed. As unfortunately has become common in this research literature, DeFouw et al. (2019) did not investigate the relation of treatment adherence or implementation to student outcomes, begging the question how does implementation relate to student outcomes within mathematics interventions.

Dimensions of Implementation

Dane and Schneider (1998) outlined the dimensions of treatment fidelity that are thought to be intervention-specific measures. Some aspects of implementation, though, can be investigated through intervention-neutral measures that can be used to gain an understanding of the salient factors of intervention instruction. The dimensions of implementation, along with how each is measured (i.e., through intervention-specific or neutral measures, direct observation or self-report, etc.) and common tools used are described next. Additionally, the importance of each dimension when considering implementation as a construct that impacts student achievement is described.

Exposure or Dosage

Exposure to the intervention program is the most fundamental aspect of implementation, for without exposure to the intervention, change in outcomes could not be attributed to the intervention. Exposure or dosage has also been conceptualized as duration, meaning the frequency or length of the intervention that the student participated in (Dane & Schneider, 1998; O'Donnell, 2008). Though authors often report how long the intervention is meant to last, information on the amount of time students were present for the intervention can often be overlooked. Without participation in the intervention, the remaining aspects of implementation are irrelevant since students cannot access what they are not present for.

Participant Responsiveness

Participant responsiveness has been conceptualized as the extent to which a student engages in an intervention, such as through measurement of their participation or enthusiasm regarding the content or activities of the intervention (Dane & Schneider, 1998; O'Donnell, 2008). Participant responsiveness is often difficult to capture, but has been evaluated through surveys of student interest in the program or direct observation regarding the extent to which students participate in a given lesson. Participant responsiveness could include indicators such as participation and enthusiasm and may also be linked to other aspects of implementation, such as the quality of implementation delivery (Dane & Schneider, 1998).

Program Differentiation

Program differentiation focuses on the extent to which the intervention differs from the comparison condition, and is a safeguard against treatment diffusion (Dane & Schneider, 1998; O'Donnell, 2008). Program differentiation requires observation or data collection across both treatment and control conditions to evaluate the differences, or possible cross-over, between the two conditions. This aspect of implementation may be more important to researchers than practitioners, as students who are identified to need intervention in school settings are often provided that intervention rather than assigned to either treatment or control as is done in research studies on interventions.

Treatment Adherence

Treatment adherence is the conventional way that researchers measure fidelity of implementation. Adherence has been defined in the literature as a dichotomy on whether the components of the intervention have been delivered as intended (Dane & Schneider, 1998; O'Donnell, 2008). In a review of the literature, O'Donnell (2008) found that a majority of research articles cited treatment adherence, whereas other aspects of fidelity of implementation were more rarely provided to the reader.

Tools that measure adherence to a treatment are structural measures of fidelity, meaning that these measures evaluate the quantity or frequency of different discrete aspects of instruction (O'Donnell, 2008). Observations rating the frequency of opportunities to respond are one example of structural measures of instruction used in the literature (e.g., Stichter et al., 2006; Sutherland et al., 2006). Structural measures are often used in specialized instruction settings, such as in special education classrooms or within Tier 2 interventions. An example of this type of measure can be found in the work on the ROOTS intervention program, in which trained observers evaluated the extent to which interventionists completed what is perceived as critical steps of the intervention program. These critical steps included the extent to which the interventionists: (a) met the lesson's instructional objectives, (b) followed the teacher scripting from the program, (c) used the outlined mathematical models for the lesson, and (d) taught the number of prescribed activities. The first three of these items were rated on a 4-point scale with 4 representing complete or all aspects were provided to the students and 1 being that none of the aspects for that item were presented to the students. For the final item, the number of prescribed activities, a count of the activities was collected by the observers, with 5 total activities possible. Though fidelity was measured, there has yet to be investigations on how these data relate to student outcomes.

The development of this adherence measure aligns with the guidelines set forth by Sanetti and Collier-Meek (2019) outlining the tools to evaluate and sustain successful interventions in schools. The first recommended step in creating an adherence measure is to break the intervention down into discrete steps and define each step in an observable, measurable way. All of these steps combined should be representative of the entire intervention. Second, the evaluation team must select the type of measure to be used, including permanent products, self-report, or direct observations. Next, the team must determine how each item will be rated. The final step is a multi-step process to get the measure operational. This includes creating the measure, training those who will be collecting the data, and determining a schedule for when data will be collected.

In an investigation on the influence of different assessment methods (e.g., direct observation versus permanent product) and other decisions regarding data collection (e.g., number of intervention steps assessed, observation timing, and number of sessions) on fidelity treatment data, Collier-Meek and colleagues (2020) found that treatment fidelity estimates are dependent on these important decisions made by researchers or evaluators of implementation fidelity. Permanent product data collection procedures consistently indicated higher levels of treatment fidelity than data collected via direct observations (Collier-Meek et al., 2020). This differs from previous work in this area where mixed results were found (Sanetti & Collier-Meek, 2014), so additional research is warranted. These findings have important implications for both researchers and practitioners overseeing interventions within MTSS because their decisions on collecting implementation data is an important consideration for both groups of stakeholders. For example, stakeholders could rely solely on permanent product data collection procedures, though it is known that these results differ from direct observations by a trained observer. The mixed findings across studies indicates it may be inconsistent whether the direct observations ratings would be higher or lower. With this inconsistent information, stakeholders cannot make informed decisions on whether their data is a true reflection of what is happening in the instructional environment.

Quality of Implementation

Treatment adherence measures as checklists of procedures can overlook important aspects of implementation, such as how well a program was delivered or how the relationship between instructor and student may impact student progress. For instance, though an interventionist may check all of the items on a checklist, if he or she reads directly from the script throughout the lesson without enthusiasm or any excitement, the students are likely to become disengaged and may not respond to the intervention as strongly as if the content or instruction was engaging. This point has been emphasized in the literature as the debate between adherence or fidelity and adaption (Dane & Schneider, 1998), with some believing that a compromise between adherence and adaption results in the most successful intervention outcomes (Bauman et al., 1991; Harn et al., 2011). For this compromise to be successful, though, one must know the essential characteristics of an intervention and deliver these with integrity, while given the opportunity to adapt other nonessential components of the intervention. One way to account for these important aspects of implementation is by using intervention-neutral tools that measure the quality of implementation of an intervention or program. Intervention-neutral measures are measures that can be used across interventions or programs and are not tied to a specific intervention (Fritz et al., 2019).

Dimensions of Implementation Quality. One important factor when evaluating implementation quality is its underlying dimensions, as these underlying dimensions of quality can aid in better understanding the components (e.g., skills, behaviors) that constitute quality of implementation. This investigation is important as it allows for the examination of different dimensions of implementation quality as they relate to treatment adherence, as well as student achievement. Despite this need, few articles report of both intervention fidelity and quality of intervention as Swanson et al. (2013) reported in their review of the literature: only 9.8% of their reviewed studies reported both treatment fidelity and quality of intervention treatment instruction. The use of intervention-neutral measures presents a promising way to investigate the implementation of intervention instruction in a way that is informative, feasible, and practical for both practitioners and researchers.

Measures of Implementation Quality. Some of these quality of implementation measures have demonstrated relations to student achievement in Tier 2 settings, including the *Recognizing Effective Special Education Teachers-Explicit Instruction* (RESET-EI; Johnson et al., 2018), *Quality of Intervention Delivery and Receipt* (QIDR; Harn et al., 2012), and *Ratings of Classroom Management and Instructional Support* measures. Each of these measures, and the research and development of the measure, will be discussed next.

Recognizing Effective Special Education Teachers-Explicit Instruction. The Recognizing Effective Special Education Teachers (RESET) Observation system (Johnson et al., 2018) is a set of 21 different observation protocols that can be used to evaluate a special education teacher's instruction. Each of the protocols evaluate a teacher's use of an evidence-based practice for special education instruction. RESET was developed to be used with special education teachers of students with high incidence disabilities. One of the observation protocols, the RESET-EI, is aimed to evaluate the use of explicit instruction strategies.

The RESET-EI protocol consists of 25 items that detail elements of explicit instruction (Johnson et al., 2019). Each item is rated on a 3-point scale from 1 being *not implemented* to 3 being *implemented.* Research has demonstrated that the RESET-EI tool can be used to reliably evaluate teacher's instructional practice (Crawford et al., 2019; Johnson et al., 2019). In a recent study evaluating the relation between teachers' RESET-EI performance and their students' achievement, Johnson et al. (2021) found teacher performance on the full set of RESET-EI did not account for significant variance in student growth. The authors then investigated a subset of items that had the highest item difficulty, including items related to identifying and communicating goals (i.e., goals are clear, specific, and relevant), teaching procedures (i.e., reviewing prior skills and activating prior knowledge, clear modeling, and appropriate number of demonstrations), scaffolding, and pacing. Through a Rasch model investigation, an additional amount of variance (4.5%) above that explained by beginning year test scores was found using these subset of difficult items (Johnson et al., 2021). Reported reliability of the RESET-EI ranges from .93-.98, demonstrating high reliability across raters and consistency across teachers' lessons (Johnson et al., 2018). This demonstrates that specific aspects of instruction may be more related to student achievement, and therefore more important for teachers to implement with fidelity.

Quality of Intervention Delivery and Receipt. Another intervention-neutral measure developed is the *Quality of Intervention Delivery and Receipt* (QIDR; Harn et al., 2012), which incorporates multiple aspects of fidelity. The QIDR includes items related to quality of and adherence to the principals of explicit instruction (e.g., teacher models, signaling, etc.), as well as student responsiveness. The QIDR has been found to demonstrate reliable estimates of instructional delivery across session-long and shorter 10-minute observation periods (Forbes-Spear, 2014; Fritz et al., 2019; Harn et al., 2014). Forbes-Spear (2014) evaluated the extent to which differing observation measures (i.e., QIDR, opportunities to respond, and *Classroom Assessment Scoring System*; Pianta et al., 2008) related to one another and student outcomes. Forbes-Spear (2014) found that each of these measures were highly correlated with one another, but only the QIDR accounted for group differences in student achievement. Though each of the investigated measures were intervention-neutral, the QIDR, which incorporates behavioral components (as opportunities to respond measures) and quality of implementation components (as the CLASS measures), was the only measure found to account for group differences in student achievement. This demonstrates that using measures that include different aspects of fidelity is important when investigating the relation between instruction and student outcomes.

Ratings of Classroom Management and Instructional Support. One tool aimed at evaluating the educators' holistic approach to classroom practice and overall support is the *Ratings of Classroom Management and Instructional Support* (RCMIS; Doabler & Nelson-Walker, 2009). This observation tool incorporates aspects of comprehensive, overarching tools, while maintaining focus on components of explicit instruction (e.g., teacher modeling, checks for student understanding, opportunities for practice, etc.). In addition to these explicit instruction strategies, other items focus on the community learning environment, organization, management techniques, student engagement, and completion of instructional tasks. Each of the 14 items on the RCMIS provide information on the quality of instruction observed and are scored on a scale from 1 (*Not Present*) to 4 (*Highly Present*). The RCMIS provides a promising approach to observing Tier 2 intervention in a meaningful way through the use of a quality-based measure of explicit instructional techniques. Previous research demonstrated that the RCMIS can be used to establish instructional quality ratings (stability ICC = .35; Doabler et al., 2014), and can be predictive of student reading and mathematics achievement (Carlson et al., 2013). Seven components of the RCMIS specific to the quality of explicit instruction (i.e., group and individual practice opportunities, efficiency of instructional delivery, student participation, teaching modeling, academic feedback, and instructional scaffolding) were found to significantly relate to students' mathematics achievement in a structured, explicit instruction intervention program (Doabler et al., 2019).

Different subsets or components of the RCMIS have been investigated at different points in time, but there has not yet been a study investigating how the components or items relate to one another. This investigation will provide critical information to researchers and teachers, as it can identify the crucial components of the measure and provide insight on how these components relate to student outcomes.

Implementation Factors as Measures to Contextualize Student Outcomes

Implementation data can provide great information to researchers and practitioners on the true effects of an intervention, allowing for a better understanding of the active ingredients and the conditions necessary for an intervention to make an impact on student outcomes (Capin, 2018). Reviews of the literature reveal that researchers rarely use treatment fidelity data in their analysis of treatment effects (Capin et al., 2018), though those that did reported that treatment fidelity was predictive of student learning (Boardman et al., 2016; Odom et al., 2010). Fidelity data can help contextualize why an intervention may not have made the impact expected, such as providing information on if these outcomes are due to unsuccessful interventions or implementation failure (Swanson et al., 2013).

Implementation measures are often used as an explanation for null results within research investigations (Hill & Erickson, 2019). By investigating implementation, researchers can determine if null results are due to failure of proper implementation, flaws in program design, or lack of fit to the context (Dane & Schneider, 1998; Hill & Erickson, 2019). Hill and Erickson (2019) conducted a review of the literature investigating the extent to which program fidelity was reported in projects funded by Institute of Education Sciences and the National Science Foundation. Further, they examined if implementation fidelity was predictive of the success of the program, and the relation of implementation fidelity to study size, type of fidelity measured, and features of the intervention. These authors found that low implementation fidelity was significantly related to student outcomes, with interventions with low fidelity averaging about 24% fewer positive student outcomes than those with higher levels of fidelity of implementation (i.e., moderate or strong). In this review of the literature, studies with moderate and strong implementation fidelity were found to have the same likelihood to produce positive student outcome effects. Additionally, studies with larger teacher sample sizes were more likely to have fewer positive impacts on outcomes than average-sized teacher samples (i.e., 442 teachers). Further, researcher-developed measures were more likely to demonstrate positive effects compared to standardized assessments of implementation. Study design (i.e., use of classroom observations versus teacher self-report) and program features (i.e., provision of curricular materials) were also found to be predictive of implementation fidelity. This research highlights the need to further investigate the aspects of implementation that relate to student outcomes, and clearly understand how different levels of implementation can impact student achievement.

Relevance to Stakeholders

When investigating implementation measures to be used in classroom or Tier 2 settings, it is important to balance the needs of educators and researchers. The overlapping need is the same: to identify a valid and reliable measure(s) of implementation that is feasible to use. Different investigations in the literature can shed light on the most important components that should be included in an implementation measure.

Crawford et al. (2012) examined the relation between two types of fidelity of implementation measures (i.e., procedural and structural in nature) and student mathematics outcomes for middle schoolers participating in a supplemental web-based curriculum. The structural measures included total time participating in the intervention, concentrated time in intervention, observation of intervention fidelity (i.e., adherence), and pretest score. The procedural measures included decision-making, problem-solving, and communication. In their findings, the structural measures were found to be significantly related to student outcomes, whereas the procedural measures were not.

Boardman et al. (2016) investigated the implementation of Collaborative Strategic Reading (CSR), which has repeatedly been demonstrated an evidenced-based practice and measures of fidelity captured as part of their studies. Within this 2016 article, Boardman and colleagues not only reported that the fidelity was high, but also examined how it related to student outcomes and specifically comparing the role of implementation to students at more (i.e., special education status) or less risk. Two measures of implementation were used across two different studies, one study with four rounds of observation and three rounds of observation in the second study. The observations focused on two dimensions of fidelity: (a) fidelity to structure or dosage of CSR students received and (b) fidelity to process or the consistency of teachers implementing CSR strategies and the quality in which they did so. Across both studies, there was not a significant relation between dosage and student outcomes for either sample. However, across both studies there was a significant interaction between the quality of implementation and special education status, with students identified as receiving special education performing better in classrooms with higher quality CSR instruction.

Odom and colleagues' (2010) early childhood study found that an adherence measure of instructional implementation (i.e., the proportion of curricular content completed) demonstrated a better fit for mathematics outcomes than a qualitative measure. Whereas when they were examining social skill outcomes, they determined that the quality-based and multiplicative (i.e., the product of the quantity and quality measure scores) measures were more strongly related than the quantity-based measure, though each of these measure types were statistically significantly related to student social skills outcomes. These results demonstrated that different observational measures of implementation may be individually associated with different types of student outcomes (i.e., academic vs social emotional), and that the combined use of both types of measures may be key to understanding the construct of *effective instruction*. Understanding effective instruction is key to determining the components that should be present during instructional interactions, particularly for at-risk learners as they may benefit differentially from quality instruction (see special issue of *Exceptional Children*; Fuchs & Fuchs 2019).

Adherence is seen as the gold standard for researchers to document the effect of an intervention, but this may be different in schools. As previously discussed, Carroll et al. (2007) recommended different elements of implementation be collected, as these different elements can serve different purposes for different stakeholders. Though adherence may be the gold standard for researchers, the quality of instructional delivery may be the most important to school-based personnel for long-term implementation.

Purpose

The purpose of this dissertation reexamining results from an efficacy study of a mathematics intervention for kindergarteners that were at-risk for mathematics difficulty was three-fold. First, this investigation examined the underlying dimensions of a quality of implementation tool in the context of a kindergarten, Tier 2 intervention curriculum. The relation between the quality of implementation dimensions and treatment adherence was also investigated. Next, this dissertation examined which elements of implementation (e.g., quality of implementation, adherence, etc.) accounted for the most variance in student learning, therefore determining which approach was best in predicting student responsiveness. Results aimed to inform both researchers and practitioners on how to evaluate the implementation of evidence-based programs for at-risk learners in mathematics.

CHAPTER II: METHODS

Purpose and Research Questions

The purpose of this dissertation was to investigate the impact different domains of implementation (i.e., quality of instruction, treatment adherence) had on student outcomes within a randomized control trial evaluating the effectiveness of a kindergarten mathematics intervention program. This dissertation conducted secondary data analyses to answer the following research questions:

Research Question 1:

a. What are the descriptive statistics of treatment adherence and *Ratings of Classroom Management and Instructional Support*?
b. Do these statistics differ by treatment group size?
c. What are the underlying dimensions of instructional quality as measured by the *Ratings of Classroom Management and Instructional Support*?

Research Question 2:

Is treatment adherence associated with factors of implementation quality?

Research Question 3:

Which measure of implementation (e.g., dimensions of instructional quality, treatment adherence) accounts for the most variance in student outcomes?

Data for this study was derived from a randomized control trial (Clarke et al., 2012) that examined the effects of the ROOTS intervention curriculum on students’ mathematics achievement for kindergarten students in Tier 2, remedial, small group settings. Interventionists provided instruction to students randomly assigned to the ROOTS curriculum based on randomized block design; students found to be at risk using screening procedures within 60 classrooms were randomly assigned to one of three conditions: (a) ROOTS intervention with a 2:1 student to teacher ratio, (b) ROOTS intervention with a 5:1 student to teacher ratio; or (c) business as usual (no treatment control). For the purposes of the current investigation, only the data from the students assigned to the two treatment conditions was used to investigate the domains of instruction, the relation of implementation measures to one another, and the relation of implementation to student outcomes. This investigation aimed to better understand the instructional variables that are associated with student outcomes and responsiveness within a kindergarten intervention program. This chapter will introduce the methods of the larger randomized control trial and the specific procedures for the present investigation.

Participants

A total of 23 schools participated in the ROOTS Efficacy Project (Clarke et al., 2012). Schools implemented the ROOTS intervention curriculum across three school years (2012-2015), and the study utilized a partially nested (within classrooms) randomized control trial (Baldwin et al., 2011). Approximately 10 students per classroom were identified as at-risk and assigned to one of three conditions: (a) ROOTS small group (2:1 student-teacher ratio); (b) ROOTS large group (5:1 student-teacher ratio); or (c) no treatment control group. A total of 880 students were randomly assigned to the ROOTS intervention; these students' data were analyzed in the current investigation. Across each participating school district, schools were targeted for participation based on their Title I status. Across all participating schools, most students were identified as either Hispanic or White. Within each school, 1-69% of students were English learners, 17-87% of students were eligible for free or reduced lunch, and 8-25% of students received special education services. Tier 1 mathematics instruction in the participating classrooms was provided 5 days a week and in English.

Students

Approximately 10 students from each participating classroom were identified as at-risk based on their initial performance on two standardized measures of early mathematics, the *Assessing Students Proficiency in Early Numeracy* (ASPENS; Clarke et al., 2011) and the *Number Sense Brief Screener* (NSB; Jordan et al., 2008). In late fall of their kindergarten year, a total of 3,066 students from 138 kindergarten classrooms were screened; students were found to be eligible for the intervention if they scored in the strategic or intensive ranges on the ASPENS Composite and a 20 or below on the NSB. A total of 1,251 students were identified as at-risk based on this screening procedure. Composite scores composed of the sum of norm-referenced standard ASPENS scores and raw NSB scores were then rank-ordered for each classroom and the 10 students with the lowest screening composite scores were then randomly assigned to one of the two intervention groups or the no-treatment control group. A total of 880 students participated in one of the 255 ROOTS intervention groups: 258 students participated in one of the small ROOTS intervention groups and 622 participated in the large ROOTS intervention groups. Demographic information for the ROOTS students is depicted in Table 1.

Table 1

Demographic Information for ROOTS Students

Characteristic	*N*	Percentage
Special education	70	8%
English language learner	201	24%
Female	425	51%
White	500	64%
Hispanic	185	24%

Note. Missing data ranged from $n = 4$ to $n = 97$; overall, $N = 880$.

Interventionists

All interventionists were either employed by the school district or hired specifically for the efficacy study. Interventionists had an average of 10.4 years of experience in education, and a majority identified as female (93.5%). Most interventionists had experience providing small group instruction (92.3%), held a bachelor's degree or higher (60.5%), and had taken a college-level algebra course (56.5%). Approximately 22% of interventionists held a current teaching license.

ROOTS Intervention

The ROOTS intervention program is a 50-lesson Tier 2 kindergarten intervention program designed to build students' proficiency in whole number concepts (Clarke et al., 2017). ROOTS was delivered outside of the Tier 1 core mathematics instruction for 20-minute sessions across 10 weeks beginning in late fall and ending in spring of students' kindergarten year. This intervention program is aligned with the Counting & Cardinality and Operations & Algebraic Thinking domains of the Common Core State Standards

(National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) and is meant to be delivered 5 days per week. ROOTS was developed using the principles of explicit and systematic mathematics instruction to include deliberate opportunities for explicit teacher models, intentional practice opportunities, visual representation of mathematics, academic feedback, and frequent opportunities for students to respond and discuss the mathematics content.

Components of ROOTS Instruction

There are four main activities within each ROOTS lesson, including: (a) “Nifty Fifty” warm-up, (b) instruction, (c) guided practice, (d) wrap-up through worksheet. Each activity, including the time frame for each, is further explained below (as described in Doabler et al, 2020).

Nifty Fifty Warm-up. The warm-up activity targeted numeral identification and the use of efficient counting strategies from 1-50 with the use of a number chart. For each day, the number focused on corresponded to the number of lessons completed in the intervention program. For instance, on the twentieth lesson of the ROOTS intervention, interventionists guided students on counting and identifying numbers up to 20. These three-minute warm-up activities supported students in identifying if one group of objects is greater than, less than, or equal to another group of objects. Additionally, the Nifty Fifty activity was aimed to support students’ one-to-one correspondence understanding.

Instruction. After the warm-up activity, interventionists provide explicit instruction to introduce new concepts or skills that align with the lessons’ objective. During this instruction, interventionists use concrete objects such as counting blocks to model and explain the targeted skill or concept. This instruction lasted approximately five minutes.

Guided Practice or Review. During the third activity of each lesson, interventionists either provided guided practice on the content or skill just introduced or provided review of previous content. This portion of the lesson lasted approximately seven minutes.

Wrap-up. In the final activity of the lesson, a worksheet activity was used to provide review of the lesson's content. Worksheets also included a note home in both English and Spanish on activities to provide students practice at home. The final activity took approximately five minutes.

Observational Measures

Two different observational measures were used for the present study: (a) a measure of treatment adherence and (b) a measure of instructional quality. Each of these measures are described below, along with data collection and coding procedures.

Data Collection

All observations were scheduled with the interventionists in advance. Throughout the length of the intervention, each ROOTS group was observed on three separate occasions. Twelve trained observers, who were blind to the research hypotheses, observed the ROOTS instruction for the duration of the ROOTS lesson for that day. Observers were former educators, doctoral students, faculty members, or experienced data collectors who received approximately 10 hours of training on direct observations, kindergarten mathematics, and the observational measures. All observers completed two reliability checks and met interobserver agreement of at least .85 on each check prior to conducting observations on their own. The average ROOTS group observation lasted 20.8 minutes (*SD* = 3.8 minutes). During observations, observers completed both the quality of instruction measure and treatment adherence checks. Of the 740 observations conducted, 139 included two observers for the evaluation of inter-observer agreement.

Treatment Adherence

Adherence to the critical components of ROOTS was measured via direct observations by research staff. During these observations, the observers rated the following components on a 4-point scale (4 = all, 3 = most, 2 = some, 1 = none): (a) lesson instruction met lesson's objectives, (b) interventionist followed the lesson's scripting, (c) interventionist used the mathematical models for the lesson, and (d) interventionist taught the number of the lesson's activities. The number of activities completed was also recorded during adherence observations. Stability estimates (ICC = .30) indicate a need for more than three observation occasions (Shoukri et al., 2004), though this was not feasible in the current study. Interobserver agreement ICCs were also calculated across observers for individual fidelity ratings, indicating moderate to nearly perfect agreement (.59-.92; Clarke et al., 2019), therefore, these data were used for the current analysis.

Process or Quality Measure

The *Ratings of Classroom Management and Instructional Support* (RCMIS) measure was used to evaluate quality of instruction. A total of 14 items were included on the RCMIS, each rated on a scale from 1 (*Not Present*) to 4 (*Highly Present*). Direct observations with trained raters were used to collect the RCMIS data. Table 2 outlines each item on the RCMIS. Stability estimates were moderate (ICC = .62 for summed RCMIS score) for three observation occasions (Shoukri et al., 2004).

Student Outcome Measures

Test of Early Mathematics Achievement, Third Edition (TEMA)

The Test of Early Mathematics Achievement, Third Edition (TEMA; Ginsburg & Baroody, 2003) was used to measure student mathematics achievement pre- and post-intervention. The TEMA is an individually administered assessment for children three to eight years old that takes 30-40 minutes to administer. It is a norm-referenced assessment that is intended to measure students informal and formal mathematics skills. The creators of the TEMA state that it has high internal reliability, with coefficient alphas ranging from .94-.96, and test-retest reliability, with coefficient alphas ranging from .82-.93 (Ginsburg & Baroody, 2003). For the purposes of the current study, TEMA posttest scores were used as one of the student outcome measure, with pretest scores nested within students in multi-level models.

ROOTS Assessment of Early Numeracy Skills (RAENS)

The ROOTS Assessment of Early Numeracy Skills (RAENS; Doabler et al., 2012) is a researcher-developed proximal measure that was untimed and individually administered at pretest at posttest. There was a total of 32 items, including items related to counting and cardinality, number operations, and base-10 system. During the assessment, students were asked to count and compare groups of objects and numbers, write and order numbers, and label 10-frames. Students also were asked to solve simple single-digit addition problems. The predictive validity of the RAENS ranges from .68 to

.83 for the TEMA and NSB; interrater scoring agreement was reported as 100% (Clarke et al., 2016).

Table 2

Items on the Ratings of Classroom Management and Instructional Support (RCMIS)

Item	Descriptors
Community of positive learning	Rapport, respect, positive attitude
Organization of instructional materials and learning tasks	Preparation, teacher-initiated transitions, accessibility
Effective small-group management techniques	Sets clear expectations, maximizes instructional time, addresses appropriate behavior
Support of students' emotional needs	Sensitivity, respect, support
Efficient delivery of instruction	Uses appropriate pacing, consistent language, minimizes student confusion
Student participation and engagement	Active involvement, compliance, competition of work
Effective teacher modeling and demonstrations	Models skills and concepts clearly, uses math representations effectively
High-quality opportunities for group practice	Offers frequency and rich opportunities for guided and independent practice
Checks of student understanding	Provides timely academic feedback, actively monitors practice opportunities
High-quality practice opportunities for individuals	Distributes individual practice opportunities, both guided and independent
Instructional scaffolding and support	Provides adequate think/response time and independent learning opportunities

Table 2 Continued

Items on the Ratings of Classroom Management and Instructional Support (RCMIS)

Item	Descriptors
Productive disposition of mathematical learning	Positive outlook on math, views math as important, confidence
Accomplishment of instructional tasks and activities	Completes tasks, uses time efficiently, student-initiated routines
Teaching for mathematical proficiency	States purpose of lesson, addresses big ideas, effective teaching examples, anticipates student misconceptions, frequent instructional interactions

Note. Items are listed in the order they appear on the RCMIS. Descriptors provide additional information regarding the behaviors observed for each item.

Analysis Plan

Research Question 1: What are the descriptive statistics of treatment adherence and Ratings of Classroom Management and Instructional Support? Do these statistics differ by treatment group size? What are the underlying dimensions of instructional quality as measured by the Ratings of Classroom Management and Instructional Support?

Descriptive Statistics. Descriptive statistics, including means, standard deviations, minimums, and maximums, were analyzed for the different implementation measures (i.e., treatment adherence and RCMIS score). The mean item ratings for each implementation variable were used to aid in interpretability. These data were run at the group-level and using SPSS 26.0.

Group Size Analysis. A *grouped* t-test was performed to test the differences in instructional quality and treatment adherence between the 2-student groups and the 5-

student groups for each variable and then analyzing if these differences could be due to chance using SPSS 26.0.. The Benjamini and Hochberg (1995) correction procedure was be used to account for multiple tests of significance.

Exploratory Factor Analysis. An exploratory factor analysis (EFA) was utilized to answer Research Question 1 to examine the underlying structure of the RCMIS. This measure was constructed to evaluate the quality of instructional interactions present during observations of Tier 2 instruction. The average score across the three observation timepoints was used for this analysis. Principal axis factoring with promax oblique rotation was used during the EFA. An oblique rotation was used in estimation as it was expected that any of the dimensions or factors describing the structure of the RCMIS would be intercorrelated. Kaiser's rule and the inspection of the scree plot was used to determine the number of factors to retain. Specifically, factors with eigenvalues greater than 1 will be retained (Kaiser's rule; Kaiser, 1970) and the scree plot was analyzed for the 'elbow' or sharp break to determine the number of factors to retain (Cattell, 1966). All factors with eigenvalues greater than 1 were evaluated for interpretability, and factors aligning with Kaiser's rule were retained.

Model Assumptions. Prior to running the EFA, tests for statistical model assumptions were completed. The Kaiser-Meyer-Olkin (KMO) test was used as a measure of sampling adequacy, which was used to measure the degree of intercorrelation between the items on the RCMIS; Barlett's test of sphericity was used to evaluate whether the population correlation matrix is uncorrelated (Dziuban & Shirkey, 1974). Additionally, visual inspection of correlations was used to test model assumptions prior to running the EFA (Preacher & MacCallum, 2003).

Research Question 2: Is treatment adherence associated with factors of implementation quality?

To evaluate the extent to which treatment adherence is associated with factors of implementation, a set of linear regression models was used. For all models, the independent variable was the measure of treatment adherence. Each of the models included the factor of instructional quality found in Research Question 1 as the dependent variable. The following equation was tested for this research question

$$Y = b_0 + b_1(\mathit{Treatment\ Adherence}) + e$$

where Y was the different factors of instructional quality, b_0 was the regression constant, b_1 was the regression coefficient for the treatment adherence measure (X), and e was the residual. The Benjamini and Hochberg (1995) correction procedure was used to account for multiple tests of significance.

Model Assumptions. Prior to running the linear regression models, model assumptions were tested, including multivariate normality and homoscedasticity. Multivariate normality was tested using visual inspection of histogram and probability plots of the dependent variable. Multicollinearity was evaluated using tolerance statistics, with expected values greater than 0.1 (Cohen et al., 2003). Next, homoscedasticity was evaluated through visual inspection of scatterplots of the independent variable with the dependent variable of factors of implementation quality.

Research Question 3: Which measure of implementation (e.g., dimensions of instructional quality, treatment adherence) accounts for the most variance in student outcomes?

For Research Question 3, 3-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002) were used. For all models, time (i) was a level-1 predictor, with students (j) nested within level-2, and group (k) association as a level-3 predictor to account for variance in implementation at the group level and controlling for differences in student outcomes related to group membership. The dependent variable was the measures of student outcomes (i.e., TEMA or RAENS). Each of the models included one of the implementation measures (i.e., treatment adherence or implementation quality) as the dependent variable. The repeated measures of each observation measure were averaged and entered into the HLM. All analyses were conducted using HLM 8.0 (Raudenbush et al., 2019). These models were used to examine the amount of variance in student outcomes explained by each observation measure. The models are specified by the following equations:

Level 1 Model: $TEST_{ijk} = \pi_{0jk} + \pi_{1jk}*(Time) + e_{ijk}$

Level 2 Model: $\pi_{0jk} = \beta_{00k} + r_{0jk}$

$\pi_{1jk} = \beta_{10k} + r_{1jk}$

Level 3 Model: $\beta_{00k} = \gamma_{000} + \gamma_{001}*(ImplementationMeasure) + u_{00k}$

$\beta_{10k} = \gamma_{100} + \gamma_{101}*(ImplementationMeasure) + u_{10k}$

Mixed Model: $Y_{ijk} = \gamma_{000} + \gamma_{001}(ImplementationMeasure) +$

$\gamma_{100}(Time) +$

$$\gamma_{101}(ImplementationMeasure*Time) + e_{ijk} +$$

$$r_{0jk} + r_{1jk}(Time) + u_{10k}(Time)$$

Models were run for each of the implementation measures separately, first with treatment adherence and then followed by instructional quality found in Research Question 1. Full maximum likelihood estimation was used for all analyses. Benjamini & Hochberg (1995) correction procedure was used to account for multiple tests of significance. r^2-equilvant (Rosnow & Rosenthal, 2003) were calculated to determine the amount of variance in student outcomes accounted for by each model.

Model Assumptions. Prior to running the HLM model, ICCs were calculated to determine the proportion of total variance in students' mathematics performance at posttest occurring between instructional groups. Histograms and box plots of each mathematics outcome were visually inspected for normality. Further, Cook's Distance was determined when running the HLM analyses. Homogeneity of variance was analyzed using a visual inspection of the Level 1 intercept residual plots and Level 2 intercept residuals. Finally, Durbin-Watson statistic was generated to determine the independence of residuals.

CHAPTER III: RESULTS

Research Question 1: What are the descriptive statistics of treatment adherence and Ratings of Classroom Management and Instructional Support? Do these statistics differ by treatment group size? What are the underlying dimensions of instructional quality as measured by the Ratings of Classroom Management and Instructional Support?

Descriptive Statistics

Item-level descriptive statistics for the treatment adherence and RCMIS measures are outlined in Table 3. Implementation measure scores were similar across the different group sizes, with high scores across both treatment adherence and RCMIS implementation measures.

Group Size Analysis

Differences in implementation based on group size were investigated using independent samples *t*-tests in SPSS 26.0 software. There were no significant differences between groups with 2 and groups with 5 students on the treatment adherence measure, $t(253) = 0.48$, $p = .635$ after Benjamini and Hochberg procedure was used. Similarly, there were not significant differences on the RCMIS item average score between the group sizes, $t(253) = -.37$, $p = .711$ after correction procedures.

Model Assumptions

The KMO test, a measure of sampling adequacy, was used to measure the degree of intercorrelation between the items on the RCMIS. This indicated adequate sampling with a KMO statistic of .97 (Kaiser, 1974). Barlett's test of sphericity, $\chi^2(91) = 401.37$, $p < .001$, was used to evaluate whether the population correlation matrix is uncorrelated and indicated that variables are stable for structure detection (Bartlett, 1954). Correlations were evaluated and determined that model assumptions were tenable.

Table 3

Descriptive Statistics for Group-Level Implementation Measures

Implementation measure	*N*	Mean	SD	Minimum	Maximum
Full sample					
Treatment adherence	255	3.64	0.40	2	4
RCMIS	255	3.14	0.48	2	4
2-student groups					
Treatment adherence	84	3.66	0.38	3	4
RCMIS	84	3.13	0.47	2	4
5-student groups					
Treatment adherence	171	3.63	0.41	2	4
RCMIS	171	3.15	0.49	2	4

Note. All data represent item level means for each implementation measure. RCMIS = *Ratings of Classroom Management and Instructional Support*. For the treatment adherence measure, 1 = none, 2 = some, 3 = most, and 4 = all of the aspects of the item were present. For the RCMIS, each item was rated from 1 = not present to 4 = highly present.

Exploratory Factor Analysis

An EFA was used to examine the underlying structure of the RCMIS items. This observational tool was constructed to measure quality of instructional interactions in Tier 2 intervention settings. The EFA was estimated using principal axis factoring with a promax oblique rotation. An oblique rotation was used in the estimation as it was

Table 4

Descriptive Statistics & Factor Loadings for Group-Level RCMIS Items

Item	Mean	SD	Minimum	Maximum	Factor Loading
Community of positive learning	3.32	0.51	2	4	.85
Organization of instructional materials & tasks	3.31	0.54	2	4	.79
Effective small-group management	3.09	0.60	1	4	.85
Support of students' emotional needs	3.25	0.50	2	4	.82
Efficient delivery of instruction	3.09	0.62	2	4	.89
Student participation & engagement	3.09	0.53	2	4	.79
Effective teacher models & demonstrations	3.08	0.59	1	4	.86
High quality group practice	3.03	0.60	1	4	.80
Checks of student understanding	3.12	0.57	1	4	.84
High quality individual practice	3.12	0.55	1	4	.89
Instructional scaffolding & support	3.04	0.55	1	4	.88
Productive disposition of math learning	3.16	0.56	2	4	.89

Table 4 Continued

Descriptive Statistics & Factor Loadings for Group-Level RCMIS Items

Item	Mean	SD	Minimum	Maximum	Factor Loading
Accomplishment of instructional tasks & activities	3.05	0.63	2	4	.88
Teaching for math proficiency	3.08	0.54	2	4	.93

Note. N = 254.

expected that the hypothesized dimensions or factors describing the structure would be intercorrelated. Using Kaiser's (1974) rule, the analysis extracted one factor (eigenvalue = 10.46) accounting for approximately 75% of the variance of the 14 items. Item means across all groups and observations are depicted in Table 4. Item commonalities were generally high, with all commonalities $h^2 > .66$ (Beavers, 2013).

Therefore, all items were retained in the analysis. Inspection of the pattern matrix revealed high loadings for items on the one factor, with all loadings ≥ .79. Factor loadings are also presented in Table 4. There was only one factor with an eigenvalue greater than one, and visual inspection of the scree plot also confirmed that a one-factor solution was appropriate. This one factor was defined as "implementation quality." The mean item level scores across all 14 items of the RCMIS will be used for later analysis.

Research Question 2: Is treatment adherence associated implementation quality?

Model Assumptions

Prior to running the linear regression model, model assumptions were tested including multivariate normality and homoscedasticity. Multivariate normality was tested using visual inspection of histogram and probability plots of the independent variable. Results indicated that the distribution of treatment adherence scores was slightly positively skewed with three outliers scoring above the 95th percentile of the distribution. Multicollinearity was evaluated using tolerance statistics, resulting in acceptable tolerance statistic of 1 (Cohen et al., 2003). Homoscedasticity was also evaluated through visual inspection of scatterplots of the independent variable with the residuals.

Regression Analysis

A regression equation with mean-centered treatment adherence as the independent variable and mean-centered quality of implementation was run using SPSS 26.0. The regression model was statistically significant as a predictor of implementation quality, $R^2 = .60$, $F(1, 253) = 380.62$, $MSR = 0.09$, $p < .001$. The intercept was not statistically significant, $t(1, 254) = .00$, $SE = .02$, $p = 1.00$. Treatment adherence was a statistically significant predictor of quality of implementation, $t(1, 253) = 19.51$, $SE = .05$, $p < .001$.

Research Question 3: Which measure of implementation (e.g., implementation quality, treatment adherence) accounts for the most variance in student outcomes?

TEMA

Model assumptions. Assumptions were tested throughout the generation of the HLMs. ICCs were calculated to determine the proportion of total variance in students' mathematics performance at posttest occurring between instructional groups. Using the null model without implementation measures as a predictor, about 3% of the proportion of variance occurred between timepoints and 26.8% of the variance in students' TEMA scores was between groups. This demonstrated that HLM was appropriate method for modeling the gains in student TEMA score across time. Histograms and box plots of posttest TEMA scores were visually inspected for normality; the distribution of scores approximated normality with a skewness of -0.24 (*SE* = 0.09) and kurtosis of -0.38 (*SE* = 0.17). Cook's distance was determined to be less than 1 and therefore within acceptable range; this demonstrated that there were no influential cases that should be investigated (Stevens, 1984). Homogeneity of variance was assessed using a visual inspection of the Level 1 intercept residual plots and demonstrated that this assumption is tenable because the Q-Q plot approximated the diagonal. Finally, Durbin-Watson statistic was generated and determined to be in acceptable range with scores between 1.5 and 2.5, so the assumption of independence of residuals is tenable (Durbin & Watson, 1971).

RCMIS. The first HLMs run included the TEMA as the outcome measures, initially with the RCMIS as a level-3 predictor variable and then with treatment fidelity as a level-3 predictor variable. Table 5 presents the results of the HLMs regressing student gains on the TEMA student outcome measures across the intervention on the RCMIS and treatment adherence measures. For the first HLM, the Predictor x Time variable represents the difference in change in TEMA score from pretest to posttest due to a unit increase in RCMIS score. Results demonstrate that gains in mathematics achievement were not significantly associated with RCMIS score (p = .28, $r^2_{equivalent}$ = .006), meaning there was not a statistically significant difference in change in TEMA score from pretest to posttest due to a unit increase in RCMIS. Further, the association between RCMIS score and pretest mathematics performance was also not statistically significant, $p = .77$. The average change in outcome from pretest to posttest among the groups given the average score on the RCMIS was 9.57, $p = .001$, meaning that there was an increase of about 10 points from pretest to posttest for students in groups receiving average instructional quality, as measured by the RCMIS.

Variance components. Inspection of the model variance components showed that average TEMA test scores differed significantly across time, χ^2 (538, N = 1,662) = 13,166.96, $p < .001$, and relation between pretest and posttest scores differed significantly from one student to another, χ^2 (538, N = 880) = 4,374.96, $p < .001$. Additionally, gains in scores also differed significantly across groups after accounting for RCMIS score, posttest score, and the interaction between these two predictors, χ^2 (251, N = 255) = 366.48, $p < .001$.

Treatment Adherence. Using treatment adherence as a predictor, similar patterns emerged. The Predictor x Time variable indicated that the predicted gains in TEMA score from pretest to posttest were not significantly associated with treatment adherence score ($p = .19$, $r^2_{equivalent} = .009$). The association between treatment adherence and TEMA pretest mathematics performance was also not statistically significant, $p = .77$. Though the average change in TEMA score from pretest to posttest given the average score on the treatment adherence measure was 9.57, $p = .001$, indicating there was an increase of about 10 points from TEMA pretest to posttest for students in groups with average treatment adherence.

Table 5

Coefficients Analysis of the RCMIS & Treatment Adherence Measures Predicting Student Outcomes

	TEMA		RAENS	
Model parameters	RCMIS	TA	RCMIS	TA
Fixed effects				
Intercept	17.08** (0.31)	17.08** (0.31)	11.49** (0.25)	11.50** (0.25)
Predictor	0.18 (0.61)	0.23 (0.73)	0.33 (0.53)	0.30 (0.67)
Time	9.57** (0.21)	9.57** (0.21)	12.41** (0.21)	12.42** (0.21)
Predictor x time	0.57 (0.46)	0.84 (0.55)	1.08* (0.46)	1.37** (0.49)
Variance components				
Intercept	34.60*** (5.88)	34.60*** (5.88)	22.91*** (4.79)	22.91*** (4.79)
Student gains	22.14*** (4.71)	22.08*** (4.70)	22.94*** (4.79)	22.89*** (4.78)
Group intercept	34.60*** (5.88)	13.25*** (3.64)	8.01*** (2.83)	8.02*** (2.83)
Group gains	22.14*** (4.71)	3.18*** (1.78)	2.60*** (1.61)	2.56*** (1.60)
p values				
Intercept	.001	.001	.001	.001
Predictor	.767	.767	.645	.704
Time	.001	.001	.001	.001

Table 5 Continued

Coefficients Analysis of the RCMIS & Treatment Adherence Measures Predicting Student Outcomes

	TEMA		RAENS	
Model parameters	RCMIS	TA	RCMIS	TA
p values				
Predictor x time	.284	.189	.029	.010
$r^2_{equivalent}$				
Predictor x time	.006	.009	.022	.030

Note. Table cells show parameter estimates with standard errors in parentheses. $df = 253$. RCMIS = Ratings of Classroom Management and Instructional Support, TA = treatment adherence, TEMA = Test of Early Mathematics Achievement, RAENS = ROOTS Assessment of Early Numeracy Skills.
*$p < .05$, **$p < .01$, *** $p < .001$.

Variance Components. Model variance components demonstrated that average TEMA test scores differed significantly across time, χ^2 (538, N = 1,662) = 13,178.87, p < .001. Further, the relation between pretest and posttest scores differed significantly from one student to another, χ^2 (538, N = 880) = 4,378.86, p < .001, and gains in scores differed significantly across groups after accounting for RCMIS score, posttest score, and the interaction between these two predictors χ^2 (251, N = 255) = 366.29, p < .001.

RAENS

Model Assumptions. ICCs were again calculated to determine the proportion of total variance in students' RAENS performance at posttest occurring between instructional groups. Specifically, approximately 3% of the variance occurred across time and 25.9% of the variance in student RAENS score occurred at the instructional group level. This again demonstrated HLM was appropriate method to model the relation of gains across time for ROOTS groups. Histograms and box plots of RAENS posttest scores were positively skewed, with a skewness of 1.16 (*SE* = 0.09) and kurtosis of 1.14 (*SE* = 0.17). Cook's distance was also evaluated and determined to be less than 1, therefore within acceptable range. (Stevens, 1984). Visual inspection of the Level 1 intercept residual plots revealed that the homogeneity of variance assumption is tenable because the Q-Q plot approximated the diagonal. Durbin-Watson statistic was generated and determined to be in acceptable range (i.e., scores between 1.5 and 2.5), therefore the assumption of independence of residuals is tenable (Durbin & Watson, 1971).

RCMIS. Results using the RAENS as the outcome measures, first run with the RCMIS as a level-3 predictor variable and then with treatment fidelity as a level-3 predictor variable are presented in the final two columns of Table 5. Time of testing, dichotomously coded (0 = pretest, 1 = posttest) was again a level 1 predictor. In these models, the Predictor x Time variable indicates the predicted gains in RAENS score from pretest to posttest based on implementation measure score. For the HLM with RCMIS as a level 3 predictor, results demonstrate that gains in mathematics achievement were significantly associated with RCMIS score (p = .03, $r^2_{equivalent}$ = .022), meaning there was a statistically significant difference in change in RAENS score from pretest to posttest due to a unit increase in RCMIS. Specifically, for every unit increase in RCMIS, the predicted RAENS score would be expected to increase by approximately 1.08 points. Additionally, the association between RCMIS score and pretest mathematics performance was not statistically significant, p = .65, and the average change in outcome from pretest to posttest among the groups given the average score on the RCMIS was 12.41, p = .001, meaning that there was an increase of about 12 points from pretest to posttest for students in groups receiving average instructional quality, as measured by the RCMIS.

Variance Components. Variance components results showed that average RAENS scores differed significantly across time, χ^2 (540, N = 1,664) = 12,748.55, $p <$ .001. The relation between pretest and posttest scores differed significantly from one student to another, χ^2 (540, N = 880) = 6,576.76, $p < .001$, and gains in scores differed significantly across groups after accounting for RCMIS score, posttest score, and the interaction between these two predictors χ^2 (251, N = 255) = 339.44, $p < .001$.

Treatment Adherence. As with the TEMA, similar patterns were present when using treatment adherence as a level 3 predictor. The Predictor x Time variable indicated that the predicted gains in RAENS score from pretest to posttest were significantly associated with treatment adherence score ($p = .01$, $r^2_{equivalent} = .030$). The association between treatment adherence and TEMA pretest mathematics performance was also not statistically significant, $p = .70$. Though the average change in RAENS score from pretest to posttest given the average score on the RCMIS was 12.42, $p = .001$, indicating there was an increase of about 12 points from RAENS pretest to posttest for students in groups with average treatment adherence.

Variance Components. Inspecting the variance components results indicated that average RAENS scores differed significantly across time, χ^2 (540, N = 1,664) = 12,787.23, $p < .001$. The relation between pretest and posttest scores differed significantly from one student to another, χ^2 (540, N = 880) = 6,596.41, $p < .001$, and gains in scores differed significantly across groups after accounting for RCMIS score, posttest score, and the interaction between these two predictors χ^2 (251, N = 255) = 337.71, p < .001.

CHAPTER IV: DISCUSSION

The purpose of this dissertation was to reexamine results from an efficacy study of mathematics interventions for at-risk kindergarteners through the examination of (a) underlying dimensions of a quality of implementation measure, (b) the relation between quality of implementation and treatment adherence, and (c) the extent to which each measure accounts for variance in student achievement.

Constructs of Implementation Quality

Implementation is conceptualized as consisting of five different dimensions: quality of implementation, treatment adherence, dosage, participant responsiveness, and program differentiation (Dane & Schneider, 1998). Treatment adherence is commonly measured in research studies, but, in school settings, data collection on any aspect of implementation is often not or inconsistently collected (Sanetti & Luh, 2019). Therefore, critical information that could guide decision-making in schools is lacking. Information on implementation could provide insight into how instruction at Tier 2 is functioning, particularly how instruction is impacting student outcomes, such as achievement and responsiveness. Lack of consistency and agreement on what defines implementation in the literature also leads to confusion on what should be measured and how (O'Donnell, 2008). The current study aimed to provide increased knowledge on what specifically constitutes the dimension of quality of implementation by evaluating the constructs that underlie the quality of implementation measure in the ROOTS efficacy study, the *Ratings of Classroom Management and Instructional Support* (RCMIS).

The factor analysis results demonstrated that only one construct was found to be measured using the RCMIS (eigenvalue = 10.46), which was therefore referred to as “quality of implementation” throughout the rest of the current study. By identifying one construct as measured by the RCMIS, it was demonstrated that quality of implementation measures can include both underlying conceptual framework (i.e., components of explicit instruction) and content-specific aspects of implementation. By including both of these aspects of quality of implementation, educators and researchers may gain a more holistic view of what is occurring during instruction.

Relevance to Researchers

Future research should investigate the underlying constructs of other measures said to evaluate quality of implementation to determine if they also measure one construct or multiple. This will aid in creating a more common definition of what quality of implementation is, especially across theoretical orientations and content areas. Without consensus on what quality of implementation truly is and other implementation issues, more students may be found to meet criteria for more intensive intervention than may be necessary, resulting in school resources being inefficiently used (Fuchs & Fuchs, 2017; Sanetti & Luh, 2019). Researchers can aid in preventing these practical problems through scientific approaches to defining implementation of evidence-based practices and translating these practices to real-world practice and policy (Eccles & Mittman, 2006; Forman et al., 2013). Further, program-independent measures such as the RCMIS should be further investigated as these measures could be useful for schools to improve their practice in a feasible manner. Due to its evidence base, measures based on explicit instruction should also be examined to determine how these measures can be used to evaluate quality of implementation within the Tier 2 setting (Fritz et al., 2019).

Relevance to Practitioners

By creating measures that could be relevant to improve instruction (i.e., used for feedback to practitioners), practitioners could be empowered with tools to increase the likelihood for student success. Measures that evaluate the quality of implementation in Tier 2 settings, such as the RCMIS, can guide practitioners in improving their practice through the monitoring of their instruction (Fritz et al., 2019; Harn, 2017). Educators need to understand what they are measuring through these tools, as well as if a tool is measuring more than one construct. The RCMIS allows educators to be confident they are measuring one construct – quality of implementation.

Relation of Implementation Measures to Each Other

Treatment adherence is often seen as the gold standard for researchers, while practitioners value measures that can facilitate feedback to interventionists. Treatment adherence measures are often shorter and more structural in nature, which can make them easier to administer in school settings. Quality of implementation measures can facilitate feedback to practitioners, but tend to take longer to administer and are more process in nature. For example, the RCMIS can provide interventionists a rating on the level of student participation and engagement as well as if the interventionist is using effective teacher modeling and demonstrations throughout a specific lesson. Some other aspects of the RCMIS, such as establishing community of positive learning, may require more in-depth discussion between the interventionist and the person providing feedback, and could include videos or demonstrations by the person providing feedback to ensure the interventionists understands what the item truly means. Another way to establish this common understanding could be by creating behavioral descriptors of each item.

Another consideration when deciding what type of implementation measure to use is the reliability of the measure (Sanetti & Luh, 2019). Treatment adherence measures and other structural-focused measures can produce high reliability and validity but lack reflective properties that can be utilized from quality of implementation and process-focused measures (Mowbray et al., 2003). It is important to understand not only the benefits of each type of measure, but also how these measures relate to one another in order to best select the measure to use in specific settings. The current study aimed to assist in understanding how a treatment adherence measure and a quality of implementation measure related to one another, as this could aid researchers and practitioners in determining which would be the best to use in a given situation.

Results indicated that treatment adherence was statistically related to quality of implementation; these two measures were highly related (R^2 = .60). Through this finding, it can be concluded that with high treatment adherence, one would expect high quality implementation and vice versa. This is important to note since it may not be feasible for researchers or practitioners to complete more than one implementation measure.

Relevance to Researchers

These two implementation measures (i.e., treatment adherence and implementation quality) were found to be related to one another and could be used in conjunction with one another or individually, depending on the needs of the researcher. Though treatment adherence is more often seen in research, quality of implementation measures can provide valuable information to the research community, such as contextual factors (i.e., behavioral expectations, explicit instruction components) that could influence implementation overall, and therefore student outcomes.

Relevance to Practitioners

Since implementation measures are highly related to one another, it may be useful for practitioners to interchange the measures based on the purpose of the observation. For example, if feedback is necessary to improve practice or the observation is to gain a qualitative understanding of the intervention instruction, quality of implementation measures may be necessary to use (Fritz et al., 2019; Harn et al., 2017). Conversely, if observation is occurring as a checkpoint between more in-depth observations, then a treatment adherence measure may be more appropriate. Though other measures of implementation (i.e., dosage, program differentiation, participant responsiveness) were not able to be investigated in this project, the two measures of implementation investigated demonstrate that the constructs of implementation are related and can provide critical information regarding Tier 2 instruction.

Implementation Relates to Student Outcomes

Implementation measures are rarely used in research to contextualize student outcomes or investigate the true effects of an intervention (Capin et al., 2018; O'Donnell, 2008). Some researchers have found that implementation measures are predictive of student learning (Boardman et al., 2016; Odom et al., 2010). Different measures or constructs of implementation have been found to relate to different types of student outcomes or content. More research needs to be conducted to evaluate patterns of implementation measures relating to specific student outcomes, as there are few of these studies. The current study aimed to add to this literature by investigating how two types of implementation measures (i.e., treatment adherence and quality of implementation measures) related to two different types of student mathematics outcomes (i.e., a distal and a proximal measure).

Both measures of implementation were related to student proximal outcomes, but not the distal outcome measures. Similar patterns emerged across both implementation measures, with the amount of variance accounted for by these measures corresponding to the amount of variance explained by other observational measures relating to student outcomes (i.e., Doabler et al., 2020; Varghese et al., 2021). These similar patterns are logical for the two implementation measures under study, as they were highly correlated (R^2 = .60). Though not statistically significant, the quality of implementation and treatment adherence measures were in the expected direction, with higher implementation scores corresponding to higher student outcomes. This trend was statistically significant with the proximal measure, though, with about one more correct answer on the RAENS for a one point increase in treatment adherence or quality of implementation score. Since the mean item score was used, both measures had a range of 1-4, but the sensitivity of each measure may differ. For instance, moving from a 3 to a 4 on an item on the RCMIS may be more difficult than moving from a 3 to 4 on a treatment adherence item. The low ICCs for each measure indicate that additional observation points are needed to establish a stable estimate of treatment adherence and quality of implementation (Shoukri et al.,, 2004) and may attenuate the associations between the observation measure and student outcomes. Though one extra point at the cut point for intervention may be meaningful, the practical significance of one additional point on a proximal measure may not be influential.

Relevance to Researchers

Implementation should be measured to truly evaluate the effectiveness of an intervention. Given that each of the constructs of implementations studied here related to student proximal outcomes, researchers should consider capturing measures of implementation during Tier 2 interventions to evaluate how implementation affected student outcomes in their work. These types of investigations can aid in determining under what conditions the intervention is most effective. Further investigations can also shed light on how to improve interventions under development and assist researchers in examining differential effects for different types of students or under different conditions (i.e., low or high quality or treatment adherence; Odom et al., 2010). Future research should evaluate how other measures of implementation (i.e., dosage, participant responsiveness, program differentiation) relate to student outcomes, and if these relations differ by content area or type of student outcome. Through the collection of implementation data in the control and treatment conditions, future research can provide valuable information about implementation dimensions, specifically program differentiation, and how these aspects of implementation affect student outcomes (Halle et al., in press).

Relevance to Practitioners

Educators should consider evaluating how implementation affects student outcomes and devote more time to not only progress monitoring student achievement but also implementation at the Tier 2 level (Harn et al., 2017). Implementation measures related differently to the different measures of student achievement, so schools should consider the progress monitoring measures they use to evaluate student progress in Tier 2.

Limitations

The findings of this dissertation should be considered with the corresponding methodological and measurement limitations.

Methodological Limitations

Carroll et al. (2007) and others (Doabler et al., 2020) have suggested that there could be a moderating or mediating relation affecting the delivery of an intervention, with Carroll et al. (2007) recommending that these moderators are process in nature. The original purpose of this dissertation was to evaluate this relation, but it could not be evaluated because of the cross-sectional modeling nature of the data. Full mediation requires longitudinal data, with repeated measures across time for the predictor, mediator, and outcome (Maxwell & Cole, 2007; Maxwell et al., 2011). With treatment adherence and quality of implementation measures being observed at the same time in the current study, bias could be introduced and can result in biased estimates if mediation were to have been used (Mitchell & Maxwell, 2013; Smolkowski, n.d.). Due to the nature of the data, the treatment adherence measure and quality of implementation measure were collected on the same day by the same observer. This results in observations that are simultaneous and, since they are conducted by the same person, ‘yoked’ in nature. Therefore, mediation models were not appropriate for use with the current data. More research needs to be conducted with measurement nets purposefully created to evaluate if process-natured measures, such as the quality of implementation measure studied here, influence the relation between treatment adherence and student outcomes. One way to allow for this future research is to have a randomized measurement schedule, where observers are not only randomly assigned to conduct a specific observation, but the measure used during that observation is also randomly assigned.

Measurement Limitations

Other measures of implementation could not be included in the current investigation due to the nature of the data. These types of measures, including measures of participant responsiveness, program differentiation, and dosage, should be included in future work on how constructs of implementation relate to student outcomes.

Future Directions

With the given limitations, future work in both research and practice can be guided by the current findings.

Research

In future work, consideration should be paid on how to collect data in a manner that would allow for testing of Carroll et al.'s (2007) theory. This would require empirically testing if treatment adherence predicts quality of implementation or the other way around. Further, future work should include analyses of how implementation impacted student outcomes in other content areas or with different types of student outcomes. This is especially important in Tier 2 settings where little oversight can occur in practice. In future work, researchers should consider using multiple measures of implementation to determine which relates most to targeted outcomes. Although similar results occurred here with treatment adherence and quality of implementation, dosage, participant responsiveness, and program differentiation were not investigated due to limitations in the data set. Different types of measures may serve different purposes at different stages of research, so researchers should carefully attend to the implementation measures they use and select measures based on the purpose of the research (Halle et al., in press).

With the measurement tools in mind, more research needs to be conducted on the underlying constructs of implementation measures. This would allow for a clearer understanding for the field on what exactly is being measured with these tools, leading to other work on how these measures relate to student outcomes. The quality of implementation measure relates to student outcomes in the current study, but not all measures are created with the same theoretical underpinnings as the RCMIS. Researchers must evaluate the tools being used to determine what constructs of implementation are being measured as well as how they are related to student outcomes.

Practice

Educators should be evaluating implementation in practice as it provides critical information regarding what is occurring with instruction in Tier 2 settings. These measures should be used to monitor instruction and provide feedback to practitioners to increase treatment adherence and quality of implementation, as both have been found here to relate to student outcomes. Specifically, higher treatment adherence and quality of implementation are related to higher student outcomes.

REFERENCES CITED

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*, 149–165. https://doi.org/10.1037/a0023464

Bartlett, M.S. (1954) A Note on the Multiplying Factors for Various Chi Square Approximations. *Journal of the Royal Statistical Society, 16*, 296-298.

Bauman, L. J., Stein, R. E. K., & Ireys, H. T. (1991). Reinventing fidelity: The transfer of social technology among settings. *American Journal of Community Psychology, 19*, 619-639.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation, 18*, Article 6. https://doi.org/10.7275/qv2q-rk76

Biggs, B. K., Vernberg, E. M., Twemlow, S. W., Fonagy, P., & Dill, E. J. (2019). Treatment adherence and its relation to teacher attitudes and student outcomes in elementary school-based violence prevention program. *School Psychology Review, 37*, 533-549. https://doi.org/10.1080/02796015.2008.12087866

Boardman, A. G., Buckley, P., Vaughn, S., Roberts, G., Scornavacco, K., & Klinger, J. (2016). Relationship between implementation of collaborative strategic reading and student outcomes with adolescents with disabilities. *Journal of Learning Disabilities, 49*, 644-657. https://doi.org/10.1177/0022219416640784

Capin, P. J. (2018). *Examining the role of treatment adherence and instructional quality in the context of content-area reading instructional treatment for eighth graders* [Doctoral dissertation, The University of Texas at Austin]. The University of Texas at Austin Texas ScholarWorks.

Capin, P., Walker, M. A., Vaughn, S., & Wanzek, J. (2018). Examining how treatment fidelity is supported, measured, and reported in K-3 reading intervention research. *Educational Psychology Reivew, 30*, 885-919. https://doi.org/10.1007/s10648-017-9429-z

Carlson, S. E., Nelson-Walker, N. J., Kennedy, P. C., Biancarosa, G., Turtura, J., Crone, D. A., Baker, S. K., Fien, H. & Cohen, J. (2013, March). *Measuring the quantity and quality of instructional interactions for middle school literacy improvement* [Conference presentation]. Society for Research on Educational Effectiveness Conference.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, Article 40. https://doi.org/10.1186/1748-5908-2-40

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). *A randomized control trial of a Tier 2 kindergarten mathematics intervention (Project ROOTS).* U.S. Department of Education, Institute of Education Sciences, Special Education Research, CFDA No. 84.32A, 2012-2016, Funding No. R324A120304.

Clarke, B., Doabler, C. T., Kosty, D., Nelson, E. K., Smolkowski, K., Fien, H., & Turtura, J. (2017). Testing the efficacy of a kindergarten mathematics intervention by small group size. *AERA Open, 3*(1), 1-16. https://doi.org/10.1177/2332858417706899

Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness, 9*, 607–634. https://doi.org/10.1080/19345747.2015.1116034

Clarke, B., Doabler, C. T., Smolkowski, K., Turtura, J., Kosty, D., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2019). Exploring the relationship between initial mathematics skill and a kindergarten mathematics intervention. *Exceptional Children, 85*, 129-146. https://doi.org/10.1177/0014402918799503

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing Student Proficiency in Early Number Sense (ASPENS).* Cambium Learning Group.

Cochrane, W., Sanetti, L. M. H., & Minster, M. (2019). School psychologists' beliefs and practices about treatment integrity in 2008 and 2017. *Psychology in the Schools, 56*, 295-305. https://doi.org/10.1002/pits.22177

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2019). Variance and reliability in special education observation rubrics. *Assessment for Effective Intervention, 45*(1), 27-37. https://doi.org/10.1177/1534508418781010

Crawford, L., Carpenter, D. M., Wilson, M. T., Schmeister, M., & McDonald, M. (2012). Testing the relation between fidelity of implementation and student outcomes in math. *Assessment for Effective Intervention, 37*, 224-235. https://doi.org/10.1177/1534508411436111

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23-45.

DeFouw, E. R., Codding, R. S., Collier-Meek, M. A., & Gould, K. M. (2019). Examining dimensions of treatment intensity and treatment fidelity in mathematics intervention research for students at risk. *Remedial and Special Education, 40*, 298-312. https://doi.org/10.1177/0741932518774801

Doabler, C. T., Clarke, B., & Fien, H. (2012). *ROOTS assessment of early numeracy skills* [unpublished measure]. Center on Teaching and Learning, University of Oregon.

Doabler, C. T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. (2020). Measuring the quantity and quality of explicit instructional interactions in an empirically validated tier 2 mathematics intervention. *Learning Disability Quarterly, 44*, 50-62. https://doi.org/10.1177/0731948719884921

Doabler, C. T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. K. (2019). Measuring the quantity and quality of explicit instructional interactions in an empirically validated tier 2 kindergarten mathematics intervention. *Learning Disability Quarterly,* Advanced online article. https://doi.org/10.1177/0731948719884921

Doabler, C. T., Nelson, N. J., Kosty, D. B., Fien, H., Baker, S. K., Smolkowski, K., & Clarke, B. (2014). Examining teachers' use of evidence-based practices during core mathematics instruction. *Assessment for Effective Intervention, 39*, 99-111. https://doi.org/10.1177/1534508413511848

Doabler, C., & Nelson-Walker, N. J. (2009). *Ratings of classroom management and instructional support*. Center on Teaching and Learning at the University of Oregon.

Durbin, J., & Watson, G. S. (1971). Testing for serial correlation in least squares regression: III. *Biometrika*, *58*(1), 1-19.

Dziuban, C. D. & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? *Psychological Bulletin, 81*, 358-361. https://doi.org/10.1037/h0036316

Eccles, M. P., & Mittman, B. S. (2006). Welcome to implementation science. *Implementation Science, 1*, 1–3. https://doi.org/10.1186 /1748-5908-1-1

Fogarty, M., Oslung, E., Simmons, D., Davis, J., Simmons, L., Anderson, L., Clemens, N., & Roberts, G. (2014). Examining the effectiveness of a multicomponent reading comprehension intervention in middle schools: A focus on treatment fidelity. *Educational Psychology Review, 26*, 425-449. https://doi.org/10.1007/s10648-014-9270-6

Forbes-Spear, C. (2014). *Examining the relationship between implementation and student outcomes: The application of an implementation measurement framework* (Doctoral dissertation). Scholars Bank, University of Oregon.

Forman, S. G., Shapiro, E. S., Codding, R. S., Gonzales, J. E., Reddy, L. A., Rosenfield, S. A., Sanetti, L. M. H., & Stoiber, K. C. (2013). Implementation science and school psychology. *School Psychology Quarterly, 28*, 77–100. https://doi.org/10.1037/spq0000019

Fritz, R., Harn, B., Biancarosa, G., Lucero, A., & Flannery, K. B. (2019). How much is enough? Evaluating intervention implementation efficiently. *Assessment for Effective Intervention, 44*, 135-144. https://doi.org/10.1177/1534508418772909

Fuchs, D., & Fuchs, L. S. (2019). Moderation analysis as a method for understanding for whom validated interventions work. [Special issue]. *Exceptional Children, 85*(2).

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*, 149-164.

Ginsburg, H., & Broody, A. (2003). *Test of Early Mathematics Ability- Third Edition.* Pro-Ed.

Halle, T. G., Metz, A. J., & Martinez-Beck, I. (Eds.). (in press). Applying implementation science in childhood settings. Brookes.

Harn, B. (2017). Making RTI effective by coordinating the system of instructional supports. *Perspectives on Language and Literacy,* 15-18.

Harn, B. A., Chard, D. J., Kame'enui, E. J. (2011). Meeting societies' increased expectations through responsive instruction: The power and potential of systemwide approaches. *Preventing School Failure: Alternative Education for Children and Youth, 55*, 232-239. https://doi.org/10.1080/1045988X.2010.548416

Harn, B. A., Forbes-Spear, C., Fritz, R., & Berg, T. (2012). *Quality of Intervention Delivery and Receipt (QIDR) observation tool.* University of Oregon.

Harn, B. A., Parisi Damico, D., & Stoolmiller, M. (2017). Examining the variation of fidelity across an intervention: Implications for measuring and evaluating student learning. *Preventing School Failure: Alternative Education for Children and Youth*, *61*, 289-302. http://doi.org/10.1080/1045988X.2016.1275504

Harn, B. A., Spear, C. F., Fritz, R., Berg, T. R., & Basaraba, D. (2014, February). *Examining the relation of features of implementation to student outcomes* [Conference presentation]. Pacific Coast Research Conference.

Hill, H. C., Erickson, A. (2019). Using implementation fidelity to aid in interpreting program effects: A brief review. *Educational Researcher*, *48*, 590-598. https://doi.org/10.3102/0013189X19891436

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2018). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues and Practice, 37*, 35-44. https://doi.org/10.1111/emip.12182

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2019). Developing an explicit instruction special education teacher observation rubric. *The Journal of Special Education*, *53*(1), 28–40. https://doi.org/10.1177/0022466918796224

Johnson, E. S., Zheng, Y., Crawford, A., & Moylan, L. A. (2021). The relationship of special education teacher performance on observation instruments with student outcomes. *Journal of Learning Disabilities*. Advance online publication. https://doi.org/10.1177/0022219420908906

Jordan, N., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Ed.), *Mathematical difficulties: Psychology and intervention* (pp. 45–57). Academic Press.

Kaiser, H. F. (1970). A second-generation Little Jiffy. *Psychometrika, 35*, 401-415.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31–36.

Meza, J. I., Friedman, L. M., Dvorsky, M. R., Kass, P., Chabbra, D., & Pfiffner, L. J. (2020). Outcomes of school-home intervention for attention and behavior problems: Teacher adherence matters. *School Mental Health, 12,* 703-715. https://doi.org/10.1007/s12310-020-09378-7

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*(1), 23-44.

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, *46*(5), 816-841. https://doi.org/10.1080/00273171.2011.606716

Mitchell, M. A., & Maxwell, S. E. (2013). A comparison of the cross-sectional and sequential designs when assessing longitudinal mediation. *Multivariate Behavioral Research*, *48*(3), 301-339. https://doi.org/10.1080/00273171.2013.784696

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315-340.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards Mathematics.* National Governors Association Center for Best Practices, Council of Chief State School Officers. http://www.corestandards.org/Math/

National Research Council Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials, Mathematical Sciences Education Board, Center for Education, Division of Behavioral and Social Sciences and Education. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations.* National Academies Press.

Nilsen, P. (2015). Making sense of implementation theories, models, and frameworks. *Implementation Science, 10,* Article 53. https://doi.org/10.1186/s13012-015-0242-0

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., Horn, E., Palmer, S., & Marquis, J. (2010). Examining different forms of implementation and in early childhood curriculum research. *Early Childhood Research Quarterly*, *25*, 314-328. https://doi.org/10.1016/j.ecresq.2010.03.001

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom Assessment Schoring System.* Paul H. Brooks.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*, 13-43. https://doi.org/10.1207/S15328031US0201_02

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. & Congdon, R. (2019). *HLM 8 for Windows* [Computer software]. Scientific Software International, Inc.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*, 221-237.

Sanetti, L. H. H., & Collier-Meek, M. A. (2019). *Supporting successful interventions in schools: Tools to plan, evaluate, and sustain effective implementation.* The Guilford Press.

Sanetti, L. M. H., & Luh, H.-J. (2019). Fidelity of implementation in the field of learning disabilities. *Learning Disability Quarterly, 42*, 204-216. https://doi.org/10.1177/0731948719851514

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. Statistical Methods in Medical Research, 13, 251–271. https://doi.org/10.1191/0962280204sm365ra

Soper, D. (2006). *Free statistics calculators: Significance of the difference between two correlations calculators.* https://www.danielsoper.com/statcalc/calculator.aspx?id=104

Smolkowski, K. (n.d.). *Mediation.* https://homes.ori.org/keiths/bibliography/statistics-mediation.html

Smolkowski, K., Crawford, L., Seeley, J. R., & Rochelle, J. (2019). Implementation science for research on learning disabilities. *Learning Disability Quarterly, 42*, 192-203. https://doi.org/10.1177/0731948719851512

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334-344.

Stichter, J. P., Lewis, T. J., Whittaker, T. A., Richter, M., Johnson, N. W., & Trussell, R. P. (2008). Assessing teacher use of opportunities to respond and effective classroom management strategies: Comparisons among high- and low-risk elementary schools. *Journal of Positive Behavior Interventions, 11*, 68-81.

Sutherland, K. S., Alder, N., & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with EBD. *Journal of Emotional and Behavioral Disorders, 11*, 239-248.

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Euducation*, *47*, 3-13. https://doi.org/10.1177/0022466911419516

Varghese, C., Bratsch-Hines, M., Aiken, H., & Vernon-Feagans, L. (2021). Elementary teachers' intervention fidelity in relation to reading and vocabulary outcomes for students at risk for reading-related disabilities. *Journal of Learning Disabilities* [Advance online publication]. https://doi.org/10.1177/0022219421999844