Implementation Factors and Their Influence on Student Mathematics Outcomes

Tasia Brafford
The Meadows Center for Preventing Educational Risk at The University of Texas at Austin

Beth Harn
University of Oregon

Ben Clarke
Center on Teaching and LearningUniversity of Oregon

Christian T. Doabler
The Meadows Center for Preventing Educational Risk at The University of Texas at Austin

Derek Kosty
Oregon Research Institute, Eugene, United States

Kathleen Scalise
University of Oregon

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324B200012 to The University of Texas at Austin. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Citation:
Brafford, T., Harn, B., Clarke, B., Doabler, C. T., Kosty, D., & Scalise, K. (2023). Implementation factors and their influence on student mathematics outcomes. *Learning Disabilities Research & Practice, 38*, 5-14. https://doi.org/10.1111/ldrp.12298

Learning Disabilities Research & Practice, *0*(0), 1–10

DOI: 10.1111/ldrp.12298

Implementation Factors and Their Influence on Student Mathematics Outcomes

Tasia Brafford
The Meadows Center for Preventing Educational Risk at The University of Texas at Austin
Beth Harn
University of Oregon
Ben Clarke
Center on Teaching and LearningUniversity of Oregon
Christian T. Doabler
The Meadows Center for Preventing Educational Risk at The University of Texas at Austin
Derek Kosty
Oregon Research Institute, Eugene, United States
Kathleen Scalise
University of Oregon

Assessing implementation allows for a better understanding of an intervention's effects and the mechanisms that influence its impact. Two main areas of implementation are (a) the quality with which an intervention is delivered and (b) instructors' adherence to the programmed intervention. The current study used data from a kindergarten mathematics intervention program to (a) examine if and how treatment adherence was associated with implementation quality and (b) explore implementation measures' relation to student mathematics outcomes. Results indicated high implementation scores across time for both adherence and quality. Neither treatment adherence nor implementation quality was found to relate to a general outcome measure of student mathematics achievement; however, both were similarly related to the curricular-aligned measure.

LEARNING DISABILITIES RESEARCH & PRACTICE

Early mathematics proficiency lays the foundation for numerous more complex skills in a student's future, including later performance in mathematics and life skills (e.g., budgeting, measurement). Students who struggle in early mathematics (i.e., preschool and kindergarten) often experience continued difficulties in later grades (Jordan et al., 2009). To improve these outcomes, students with early mathematics difficulties (MDs) require intervention and opportunity to develop a deep and robust understanding of foundational mathematics skills and concepts.

Promising Mathematics Interventions

Within a multi-tiered system of support (MTSS) model, Tier 2 interventions are provided to students with MD, in addition to Tier 1 or core instruction, to increase students' understanding of mathematics skills and concepts. A variety of mathematics interventions demonstrate promise in providing targeted and intensified instruction specifically for students with MDs, including interventions focused on improving students' understanding of whole number concepts, early number sense skills, and basic computation skills within a small-group setting (e.g., Clarke et al., 2020; Dyson et al., 2013; Fuchs et al., 2005).

Recently, Jitendra et al. (2021) investigated the effects of Tier 2 mathematics interventions designed for students with MD across prekindergarten through Grade 12; however, in 90% of the studies synthesized the participants were elementary age. Overall, Jitendra et al. (2021) found that Tier 2 interventions were implemented with high fidelity regardless of interventionist position (i.e., research staff or school personnel) and that these interventions demonstrated similar efficacy regardless of the severity of students' MD. Small groupings of two or three students with MD demonstrated the most promise, with an 0.29 increase in effect size above the average adjusted effect size of 0.46 ($p < .05$; Jitendra et al., 2021).

Requests for reprints should be sent to T. Brafford, The Meadows Center for Preventing Educational Risk at The University of Texas at Austin. Electronic inquiries should be sent to Brafford@utexas.edu.

Effective intervention programs are often based on the principles of explicit and systematic instruction (Fuchs et al., 2021), with interventionists providing models, guided practice, corrective feedback, and frequent review opportunities (Gersten et al., 2009). Researchers must use effective principles for learners with MD when designing mathematics interventions, and educators implementing these programs with students must be cognizant of these key principles to ensure instruction is provided as intended. Many intervention programs have demonstrated efficacy, including ROOTS (Clarke et al., 2012), a promising mathematics program backed by convincing evidence of its effect on improved student outcomes (Clarke et al., 2016, 2017; National Center on Intensive Intervention, 2021).

ROOTS Whole Number Foundation Program

ROOTS Whole Number Foundation Program (Clarke et al., 2012) is a 50-lesson intervention curriculum focused on whole number concepts and skills delivered in a Tier 2 setting of a MTSS model. Designed to be delivered outside of the Tier 1 core mathematics instruction, each ROOTS intervention session consists of 20 minutes of instruction and is delivered five days a week for 10 weeks beginning in late fall and ending in the spring of students' kindergarten year. ROOTS was developed using the principles of explicit and systematic mathematics instruction to include deliberate opportunities for teacher models, intentional practice opportunities, visual representation of mathematics, academic feedback, and frequent opportunities for students to respond and discuss the mathematics content. Lesson activities are described in detail in Clarke et al. (2017).

Research has demonstrated the efficacy of the ROOTS intervention for students with MD, regardless of group size (Clarke et al., 2020). Different measures of implementation have captured aspects of ROOTS instruction and how this instruction can vary based on interventionists' implementation of ROOTS.

Implementation as an Influence on Intervention Outcomes

Even when educators use an empirically established program, a wide variation of effectiveness may be seen in school settings. One aspect that may account for this variation is implementation, a multitude of factors that need to be unpacked to determine the influence on an intervention's observed effects on student achievement. Miller et al. (2014) made a call to the field of reading to evaluate the phenomenon of implementation, emphasizing the need for a focus on environmental conditions or school contexts that may strengthen an intervention's effect. The same investigative endeavor is needed in mathematics intervention contexts, where even less is known about the factors that impact intervention effectiveness.

Intervention is a complex system (Bos et al., 2022), and investigating the domains of implementation that support or hinder student achievement within this system is imperative to improve student achievement. Implementation includes a myriad domains and underlying factors (Bos et al., 2022; O'Donnell, 2008) that are often discussed and measured differently across research groups (Harn et al., 2013). Though researchers may report on these aspects of implementation (Bos et al., 2022; Jitendra et al., 2021), few have evaluated the relation between intervention implementation and student outcomes (Capin et al., 2018; O'Donnell, 2008). Two aspects of implementation relevant to researchers and practitioners include treatment adherence and implementation quality. These two measures encompass both structural and process measures of implementation, which may help provide greater clarity on the key components of an intervention's effectiveness (O'Donnell, 2008). Therefore, two measures were targeted for investigation: treatment adherence—a structural measure—and implementation quality—a process measure.

Treatment Adherence

"Adherence" has been defined in the literature as the extent to which specific intervention components are or are not delivered as intended (Dane & Schneider, 1998; O'Donnell, 2008). Treatment adherence is often measured through checklists or frequency counts of specific instructional behaviors within curricula that are seen as most influential to student outcomes. The frequency rating of opportunities to respond is one example of a treatment adherence measure used in practice and research (e.g., Stichter et al., 2008; Sutherland et al., 2003). Researchers commonly report treatment adherence as a proportion of critical aspects that are completed versus prescribed in a curriculum or program (Bos et al., 2022; O'Donnell, 2008).

Implementation Quality

"Implementation quality" encompasses aspects of prescribed program content (e.g., interventionist-student communication, use of responsive pacing, appropriateness of feedback) and is often measured using teacher rating systems. Overall classroom instructional quality, commonly measured by the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), has been found to relate strongly to aspects of positive classroom environments (La Paro et al., 2004) and increases in children's social and academic outcomes (Perlman et al., 2016). However, while the CLASS is an effective evaluation tool for whole-class Tier 1 instruction, its efficiency (i.e., multiple observations required), time demands (i.e., observers spending at least two hours in the classroom), and appropriateness for small-group settings preclude wide use in real-life intervention settings (Hamre et al., 2009). Little is known about how implementation quality measures relate to student outcomes in intervention settings where learners with MD are present and demonstrate the greatest need. The field must identify measures that researchers and educators can use to evaluate intervention implementation quality effectively and efficiently.

Implementation of Elementary Mathematics Intervention Programs

Currently missing in the mathematics intervention literature is an investigation of whether or how treatment adherence and implementation quality affect an intervention's impact. Many author teams do not even report implementation quality in elementary mathematics intervention research (Bos et al., 2022). Furthermore, treatment adherence and implementation quality have not yet been investigated within the same study to determine their role in the effectiveness of an intervention for students with MD (Bos et al., 2022; Jitendra et al., 2021; Nelson et al., 2012). To fill these gaps in the literature, the current study sought to unpack the question: How does intervention implementation influence mathematics outcomes for young students with MD?

Purpose

Through a secondary analysis of implementation factors captured during a ROOTS efficacy trial (Clarke et al., 2012), the current study expanded the implementation literature by disentangling interventionists' adherence to an intervention protocol from implementation quality of an empirically validated kindergarten mathematics intervention and determined how each implementation domain accounts for student learning. Specifically, the following two research questions were addressed: (a) Is treatment adherence associated with implementation quality? And (b) which measure of implementation (i.e., treatment adherence, implementation quality) accounts for the most variance in student achievement?

METHODS

Data were derived from a randomized control trial (Clarke et al., 2012) examining the effects of ROOTS on kindergarten students' mathematics achievement. Interventionists provided instruction to students assigned to ROOTS based on a randomized block design; that is, students identified as at risk for MD using screening procedures within 60 classrooms were randomly assigned either ROOTS or business as usual (i.e., no treatment control). Only data from the students assigned to ROOTS were used for the current study.

Participants

Twenty-three schools participated in the project; all schools were Title 1 eligible. Tier 1 mathematics instruction in the participating classrooms was provided in English five days a week.

Students

Approximately 10 students per classroom were identified as being at risk for MD based on their performance on two standardized measures of early mathematics: Assessing Students Proficiency in Early Numeracy (ASPENS; Clarke et al., 2011) and Number Sense Brief Screener (NSB; Jordan et al., 2008). A total of 880 students were assigned to ROOTS groups, including students receiving special education (n = 70, 8%) or English language learner (n = 201, 24%) services. Approximately half of the students (n = 425, 51%) were identified as female. Furthermore, over half of the students were identified as white (n = 500, 64%), and nearly one quarter were identified as Hispanic (n = 185, 24%). There was between 0.5%−11% (n = 4−97) missingness in each of the demographic categories. Information regarding the screening and randomization procedures may be found in Clarke et al. (2017).

Interventionists

All interventionists were either employed by the participating school district or hired for the study. Interventionists had an average of 10.4 years of experience in education; a majority identified as female (93.5%). Most 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 teaching license.

Intervention Implementation Measures

Treatment Adherence

Adherence to the critical components of ROOTS was measured by observer ratings on the following components using a four-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 total number of the lesson's activities. The number of activities completed was also recorded. Interclass correlation (ICC) stability estimates indicate a need for more than three observation occasions (ICC = .30; Shoukri et al., 2004); however, this was not feasible in the current study as the instructional observation investigations were not the primary research questions of the main project. Finally, interobserver agreement ICCs were calculated across observers for individual fidelity ratings, indicating moderate to nearly perfect agreement (.59–.92; Clarke et al., 2019).

Implementation Quality

Ratings of Classroom Management and Instructional Support (RCMIS) was used to evaluate implementation quality. Each of 14 items was rated on a four-point scale: 1 = *not present*; 2 = *somewhat present*; 3 = *present*; and 4 = *highly present*. Table 1 outlines each item on the RCMIS. Stability estimates were moderate (ICC = .62 for summed RCMIS score) for three observation occasions (Shoukri et al., 2004).

TABLE 1
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 frequent 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
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 in which they appear on the RCMIS. Descriptors provide additional information regarding the behaviors observed for each item.

Data Collection

All observations were scheduled with the interventionists in advance. Across the study, each group was observed on three separate occasions. The 12 trained observers included former educators, doctoral students, faculty members, and experienced data collectors who received approximately 10 hours of training on direct observations, kindergarten mathematics, and each observational measure. All observers completed two reliability checks and met interobserver agreement of at least .85 prior to conducting observations on their own. During observations ($M = 20.8$ minutes, $SD = 3.8$), observers completed both the adherence and the quality measures. Of the 740 observations conducted, 139 included two observers for evaluation of interobserver agreement.

Student Outcome Measures

Test of Early Mathematics Achievement, Third Edition (TEMA)

The Test of Early Mathematics Achievement, Third Edition (TEMA; Ginsburg & Baroody, 2003), a norm-referenced assessment, was used to measure students' mathematics achievement pre- and postintervention. The TEMA is an individually administered assessment for children 3 to 8 years old and takes 30−40 minutes to administer. For this study, the TEMA served as the distal measure of student mathematics performance as it was (a) not developed by the intervention curriculum team and (b) not aligned specifically with the ROOTS intervention curriculum. According to the creators of the TEMA, it has high internal reliability, with coefficient alphas ranging from .94 to .96, and test-retest reliability, with coefficient alphas ranging from .82 to .93 (Ginsburg & Broody, 2003). For the purposes of the current study, TEMA posttest scores were used as one of the student outcome measures, with pretest scores nested within students in multilevel models.

ROOTS Assessment of Early Numeracy Skills (RAENS)

The ROOTS Assessment of Early Numeracy Skills (RAENS; Doabler et al., 2012) was developed by the ROOTS intervention curriculum team and is related to the content of the intervention; it served as the proximal measure of student mathematics achievement in the present study. Untimed, the RAENS was individually administered at pre- and posttest. During the 32-item assessment, students were asked questions related to counting and cardinality (e.g., verbally count; compare groups of objects and numbers), numbers and operations (e.g., write and order numbers; solve single-digit addition problems), and the base-10 system (e.g., label 10-frame). 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).

Analytic Method

Association of Implementation Domains

To evaluate the extent to which treatment adherence was associated with implementation quality, a linear regression model was used with the measure of treatment adherence as the independent variable. The following equation was tested for Research Question 1: $Y = b_0 + b_1(\textit{Treatment Adherence}) + e$, where Y was implementation 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.

Implementation Accounting for Student Outcomes

For Research Question 2, three-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 independent variable. The repeated measures of each observation measure were averaged. 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: $TEST_{ijk} = \pi_{0jk} + \pi_{1jk} \times (Time) + e_{ijk}$

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

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

Level 3 Model: $\beta_{00k} = \gamma_{000} + \gamma_{001} \times (\textit{Implementation-Measure}) + u_{00k}$

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

Mixed Model: $Y_{ijk} = \gamma_{000} + \gamma_{001}(\textit{ImplementationMeasure}) + \gamma_{100}(\textit{Time}) + \gamma_{101}(\textit{ImplementationMeasure} \times \textit{Time}) + e_{ijk} + r_{0jk} + r_{1jk}(\textit{Time}) + u_{10k}(\textit{Time})$

Models were run for each implementation measure separately, first with treatment adherence and then implementation quality. Full maximum-likelihood estimation was used for all analyses. The Benjamini and Hochberg (1995) correction procedure was used to account for multiple tests of significance. $r^2_{\text{equivalent}}$ (Rosnow & Rosenthal, 2003) was calculated to determine the amount of variance in outcomes that was accounted for by each model.

RESULTS

Item-level descriptive statistics for the group-level implementation measures demonstrated high levels of adherence to the program protocol and implementation quality (M = 3.64, SD = 0.40, range = 2−4 for treatment adherence; M = 3.14, SD = 0.48, range = 2−4 for RCMIS). Specifically, on average, interventionists completed most to all the aspects of an individual treatment adherence item. Additionally, interventionists delivered the intervention with quality, as measured by the individual RCMIS items.

Association of Implementation Domains

A regression equation with mean-centered treatment adherence as the independent variable and mean-centered implementation quality was used. The regression model was statistically significant, $R^2 = .60$, $F(1, 253) = 380.62$, MSR = 0.09, $p < .001$. The intercept was not statistically significant, $t(1, 254) = 0.00$, $SE = 0.02$, $p = 1.00$. Treatment adherence was a statistically significant predictor of implementation quality, $t(1, 253) = 19.51$, $SE = 0.05$, $p < .001$.

Implementation Accounting for Student Outcomes

TEMA

The first HLMs included the TEMA as the outcome measures, initially with the treatment adherence as a Level 3 predictor variable and then with RCMIS as a Level 3 predictor variable.

Treatment adherence. Table 2 presents the results of the HLMs regressing student gains on the TEMA across the intervention on the treatment adherence and RCMIS measures. For the first HLM, the *Predictor* × *Time* variable represents the difference in change in TEMA score from pretest to posttest due to a unit increase in treatment adherence score. The Predictor × Time variable indicated that the predicted gains in TEMA score from pretest to posttest were not significantly associated with treatment adherence score (p = .19, $r^2_{\text{equivalent}} = .009$). The association between treatment adherence and TEMA pretest mathematics performance was not statistically significant, $p = .77$, although 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.

RCMIS. Using RCMIS score as a predictor, similar patterns emerged. That is, gains in mathematics achievement were not significantly associated with RCMIS score (p = .28, $r^2_{\text{equivalent}} = .006$), meaning there was not a statistically significant difference in change in TEMA score from pretest to posttest due to a unit increase in RCMIS. Furthermore, 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

TABLE 2
Coefficients Analysis of the RCMIS and Treatment Adherence Measures Predicting Outcomes

Model Parameters	*TEMA*		*RAENS*	
	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 × 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	0.001	0.001	0.001	0.001
Predictor	0.767	0.767	0.645	0.704
Time	0.001	0.001	0.001	0.001
p Values				
Predictor × Time	0.284	0.189	0.029	0.010
$r^2_{\text{equivalent}}$				
Predictor × Time	0.006	0.009	0.022	0.030

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.
$*p < .05$.
$**p < .01$.
$***p < .001$.

receiving average implementation quality, as measured by the RCMIS.

RAENS

Treatment adherence. The results using the RAENS as the outcome measure, first run with the treatment adherence as a Level 3 predictor variable and then with RCMIS as a Level 3 predictor variable, are presented in the final two columns of Table 2. Time of testing, dichotomously coded (0 = pretest, 1 = posttest) was again a Level 1 predictor. The Predictor × Time variable indicated that the predicted gains in RAENS score from pretest to posttest were significantly associated with treatment adherence score ($p = .01$, $r^2_{\text{equivalent}} = .030$). The association between treatment adherence and TEMA pretest mathematics performance was not statistically significant, $p = .70$. The average change in RAENS score from pretest to posttest given treatment adherence average score 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.

RCMIS. As with the TEMA, similar patterns emerged when using RCMIS as a Level 3 predictor. In these models, the Predictor × 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 demonstrated that gains in mathematics achievement were significantly associated with RCMIS score ($p = .03$, $r^2_{\text{equivalent}} = .022$), 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 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 given the average score on the RCMIS was 12.41, $p = .001$, meaning there was about a 12-point increase for students in groups receiving average implementation quality according to the RCMIS.

DISCUSSION

This study reexamined the results from an efficacy study of a mathematics intervention for at-risk kindergarteners through an examination of (a) the relation between implementation quality and treatment adherence and (b) the extent to which each measure (i.e., implementation domain) accounted for variance in student achievement. Specifically, we evaluated how two implementation measures related to one another to identify which would be the best to use across observational opportunities.

Our results indicated that treatment adherence was statistically and highly related to implementation quality ($R^2 = .60$), leading to similar results when investigating how the implementation domains accounted for the variance in students' proximal (RAENS) and distal (TEMA) performance. The distal and proximal status of these measures were denoted by each measure's alignment to the curricular content, not the timing of measure completion. These findings indicate that, with high treatment adherence, one would expect high-quality implementation and vice versa. In the following, we discuss these findings as they relate to both practical and measurement factors and how these findings can guide future intervention curricula and measurement work.

Relation of Implementation Measures to Each Other

Treatment adherence is often seen as the gold standard for researchers, whereas practitioners value measures that can be used to facilitate specific instructional feedback (Cook et al., 2012). Treatment adherence measures are often shorter and more structural in nature, which can make them easier to administer in school settings. Structural measures, including checklists, are more objective than process measures (Mowbray et al., 2003). Implementation quality measures can facilitate feedback to practitioners but tend to take longer to administer and are more process in nature (Fritz et al., 2019). Process measures, which are often evaluated through rating scales, can be more subjective in nature, and include interactions between the program staff and clients, treatment delivery, or program (Mowbray et al., 2003). For example, the RCMIS can provide interventionists a rating on the level of student participation and engagement as well as whether the interventionist is using effective teacher modeling and demonstrations throughout a specific lesson. Other aspects of the RCMIS, such as establishing a community of positive learning, may require more in-depth discussion between the interventionist and the person providing feedback, and may include videos or demonstrations to ensure the interventionist understands what each item truly means. Another way to establish a common understanding involves creating behavioral descriptors of each RCMIS item.

Relation of Implementation 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), yet they have been shown to be predictive of student learning. Different measures or constructs of implementation have been found to relate to diverse types of student outcomes or content (Boardman et al., 2016; Odom et al., 2010).

In the current study, we found that both measures of implementation were related to the proximal measure of student outcomes (p-values $< .03$), but not the distal outcome measure (p-values $> .19$). Similar patterns emerged across both implementation measures, which is not surprising given how highly correlated the measures were with one another. The amount of variance in student outcomes accounted for by these measures ($r^2_{\text{equivalent}} = .01$–$.03$) was akin to that explained by other observational measures evaluating implementation quality and treatment adherence in research settings (i.e., Doabler et al., 2021; Varghese et al., 2021).

Though not statistically significant, the implementation quality and treatment adherence measures trended in the expected direction, with higher implementation scores corresponding to higher student outcomes. This trend was statistically significant with the proximal measure; a one-point increase in treatment adherence or implementation quality score related to a one-unit increase—one more correct answer—on the RAENS. One extra point at the cut point on a screening measure may be meaningful, but one additional point (i.e., problem correct) on a proximal measure may not be practically significant. Since the mean item score was used, both implementation 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 or less difficult than moving from a 3 to 4 on the treatment adherence tool. The difference in these changes has not yet been investigated.

Relevance for Researchers

These implementation measures were found to be related to one another and could be used in conjunction with one another or individually, depending on the context and needs within an educational setting. Treatment adherence is more often seen in research, but implementation quality measures can also provide valuable information, such as contextual factors (i.e., behavioral expectations, explicit instruction components) that may influence implementation and overall instruction, therefore affecting student outcomes.

Implementation should be measured to truly evaluate the effectiveness of an intervention. Given that each of the implementation domains studied here related to student proximal outcomes, researchers should consider capturing measures of implementation in intervention research and evaluating how implementation factors affected student outcomes. Such investigations can aid in determining under what conditions an 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 student groups or under different conditions (i.e., low or high quality of implementation or treatment adherence; Odom et al., 2010). Future research should evaluate how other measures of implementation, such as those described by Dane and Schneider (1998; i.e., dosage, participant responsiveness, program differentiation) relate to student outcomes, and if these relations differ by content area or instructional setting. Through the collection of implementation data in control and treatment conditions, future research can provide valuable information about other implementation dimensions, such as program differentiation, and how these implementation aspects affect student outcomes (Halle et al., 2013).

Another consideration is the low ICCs for each measure, indicating that additional observation points are needed to establish a stable estimate of treatment adherence and implementation quality (Shoukri et al., 2004) and may attenuate the associations between the observation measure and student outcomes. With higher ICCs, we would expect to be able to better capture the true nature of "implementation quality" and "treatment adherence." With better estimates of implementation in the different ROOTS groups, we would be better able to see differentiation between these implementation domains, resulting in possible differences in their ratings and their link to student outcomes. Consequently, future research should include multiple observation time points to improve the accuracy of implementation domain estimates.

Relevance for Practitioners

Since the implementation measures studied here are highly related to one another, practitioners could develop a more efficient observation schedule by interchanging these measures based on the observational purpose. For example, if feedback is necessary to improve practice or if the goal is to gain a qualitative understanding of intervention instruction, implementation quality measures may be necessary (Fritz et al., 2019; Harn, 2017). Conversely, if observation is occurring as a checkpoint before more in-depth observations, then a treatment adherence measure may be more appropriate. Though other implementation domains (i.e., dosage, program differentiation, participant responsiveness) were not investigated in this project, the two measures of implementation investigated demonstrate that at least some of the constructs of implementation are related both to each other and to student outcomes, providing critical information on instruction.

Limitations

Carroll et al. (2007) and others (Doabler et al., 2021) have suggested the possibility of a moderating or mediating relation affecting intervention delivery. According to Carroll et al. (2007), these moderators are process in nature (e.g., implementation quality, participant responsiveness). This theory could not be evaluated within the current study due to the cross-sectional nature of the data. With treatment adherence and implementation quality measures being observed at the same time in the current study, the use of mediation might have led to bias and result in biased estimates (Mitchell & Maxwell, 2013; Smolkowski, n.d.). Due to the procedures outlined in the efficacy trial, the treatment adherence and implementation quality measures were collected on the same day by the same observer, resulting in simultaneous observations that are "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 implementation quality measure studied here, influence the relation between treatment adherence and student outcomes.

Future Directions

Analyses of how implementation impacts student outcomes in other content areas or with different types of student outcomes are necessary. This is especially important in Tier 2 settings where little oversight can occur in practice (Harn, 2017). 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., 2013). The implementation quality measure related 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 implementation domains are being measured and how they relate to student outcomes.

Reporting of treatment adherence in research is not sufficient as this reporting of interventionists' adherence to the prescribed protocol is inadequate to inform educators on *how* implementation affects student performance. High treatment adherence may insinuate that the instruction is effective, but interventionists may deviate from the instructional protocol to ensure a range of opportunities for students to engage with the material, particularly if students demonstrate difficulty answering questions correctly. This example demonstrates why multiple factors of implementation should be considered in tandem, as the interventionist's instructional quality in addition to treatment adherence provides a more robust picture of what is truly happening during the intervention instruction.

Educators should be evaluating implementation in practice to gain critical information regarding how to evaluate students' response to intervention. Such measures should be used to monitor instruction and provide feedback to practitioners to increase treatment adherence and implementation quality, as both have been found here and in previous work (Fritz et al., 2019) to relate to student outcomes. Specifically, higher treatment adherence and implementation quality were related to higher student outcomes. In practice, if students are not making adequate progress in an intervention setting, the standard decision is to increase intervention intensity. But an essential, but often ignored, factor is the investigation of the adherence to or implementation quality of the intervention. If treatment adherence or implementation quality is poor, we must improve intervention delivery rather than concluding that the lack of student progress requires increased intervention intensity. The true problem here is actually the lack of measuring and evaluating key implementation factors and making related improvements in intervention implementation.

REFERENCES

Benjamini, T., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

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

Bos, S. E., Powell, S. R., Maddox, S. A., & Doabler, C. T. (2022). A synthesis of the conceptualization and measurement of implementation fidelity in mathematics intervention research. *Journal of Learning Disabilities*. Advance online publication. https://doi.org/10.1177/00222194211065498

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 Review*, *30*, 885–919. https://doi.org/10.1007/s10648-017-9429-z

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

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., Doabler, C. T., Turtura, J., Smolkowski, K., Kosty D. B., Sutherland, M., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2020). Examining the efficacy of a kindergarten mathematics intervention by group size and initial skill. *The Elementary School Journal*, *121*, 125–153. https://doi.org/10.1086/710041

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency in early number sense (ASPENS)*. Cambium Learning Group.

Cook, B. G., Smith, G. J., & Tankersley, M. (2012) Evidence-based practices in education. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 495–525). American Psychological Association. https://doi.org/10.1037/13273-017

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. https://doi.org/10.1016/S0272-7358(97)00043-3

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

Doabler, C. T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. K. (2021). Measuring the quantity and quality of explicit instructional interactions in an empirically validated tier 2 kindergarten mathematics intervention. *Learning Disability Quarterly*, *44*, 50–62. https://doi.org/10.1177/0731948719884921

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergarteners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*, 166–181. https://doi.org/10.1177/0022219411410233

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, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*, 493–513. https://doi.org/10.1037/0022-0663.97.3.493

Fuchs, L. S., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K. S., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R., Jyanthi, M., Keating, B., & Morgan, S. (2021). Assisting students struggling with mathematics: Intervention in the elementary grades (WWC 2021006). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/WWC2021006-Math-PG.pdf

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools (NCEE 2009–4060). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/practiceguide/2

Ginsburg, H., & Broody, A. (2003). Test of early mathematics ability (3rd ed.). Pro-Ed.

Halle, T. G., Metz, A. J., & Martinez-Beck, I. (Eds.). (2013). *Applying implementation science in childhood programs and systems*. Paul H. Brookes Publishing Co.

Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). Classroom Assessment Scoring System (CLASS) implementation guide. https://www.boldgoals.org/wp-content/uploads/CLASSImplementationGuide.pdf

Harn, B. (2017). Making RTI effective by coordinating the system of instructional supports. *Perspectives on Language and Literacy*, *43*(4), 15–18.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*, 181–193. https://doi.org/10.1177/001440291307900205

Jitendra, A. K., Alghamdi, A., Edmunds, R., McKevett, N. M., Mouanoutoua, J., & Roesslein, R. (2021). The effects of tier 2 mathematics interventions for students with mathematics difficulties: A meta-analysis. *Exceptional Children*, *87*, 307–325. https://doi.org/10.1177/0014402920969187

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.

Jordan, N. C., Kaplan, D., Ramineni, C., & Louniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*, 850–867. https://doi.org/10.1037/a0014939

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*, 409–426.

Miller, B., Vaughn, S., & Freund, L. S. (2014). Learning disabilities research studies: Findings from NICHD-funded projects. *Journal of Research on Educational Effectiveness*, *7*, 225–231. https://doi.org/10.1080/19345747.2014.927251

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 Center on Intensive Intervention. (2021). Academic intervention tools chart. https://charts.intensiveintervention.org/aintervention

Nelson, M. C., Cordray, D. S., Julleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research*, *39*, 374–396. https://doi.org/10.1007/s11414-012-9295-x

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. https://doi.org/10.3102/0034654307313793

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PLoS ONE*, *11*, e0167660. https://doi.org/10.1371/journal.pone.0167660

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom assessment scoring system*. Paul H. Brookes.

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.

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

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

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. https://doi.org/10.1177/10634266030110040501

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

About the Authors

Tasia Brafford, PhD, BCBA, is a postdoctoral fellow at The Meadows Center for Preventing Educational Risk at The University of Texas at Austin. Dr. Brafford's work focuses on mathematics assessment and intervention, implementation in school settings, and preservice teacher education.

Beth Harn, PhD, is an associate professor in special education who teaches graduate-level courses in special education and school psychology. She has expertise in early literacy assessment, instruction, and intervention development and implementation.

Ben Clarke, PhD, is an associate professor in the School Psychology Program and the Director of the Center on Teaching and Learning at the University of Oregon. His research is broadly focused the development of mathematical thinking and how school systems can support the mathematical learning needs of all students.

Christian T. Doabler is an assistant professor in the Department of Special Education and a Research Fellow of the Mathematics and Science Institute for Students with Special Needs at The Meadows Center for Preventing Educational Risk at The University of Texas at Austin. Dr. Doabler's research focuses on designing and delivering effective science, technology, engineering, and mathematics (STEM) instruction for students who are at risk for learning difficulties, including multilingual students and students from marginalized and underserved communities. His research also includes supporting teachers' use of evidence-based practices to promote equitable access to rich and engaging STEM instruction for the full range of learners.

Derek Kosty, PhD, studies problematic substance use across the lifespan, effects of academic and behavioral interventions, and applied quantitative research methods. His methodological expertise includes group- and single-case designs, advanced statistical modeling of longitudinal and multilevel data, and structural equation modeling with latent variables.

Kathleen Scalise is a professor at the University of Oregon in the Department of Methodology, Policy and Leadership. Dr. Scalise employs data science at the intersection with measurement and assessment for applied and theoretical research, including for learning in digital social networks and science/engineering education, as well as for network analyses of leadership and collaboration.