EXAMINING RELATIONSHIPS BETWEEN COMPONENTS OF IMPLEMENTATION FIDELITY AND STUDENT RESPONSE WITHIN THE CONTEXT OF AN EARLY NUMERACY INTERVENTION

by

CAYLA LUSSIER

A DISSERTATION

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

June 2024

DISSERTATION APPROVAL PAGE

Student: Cayla Lussier

Title: Examining Relationships Between Components of Implementation Fidelity and Student Response Within the Context of an Early Numeracy Intervention

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the School Psychology Department by:

Ben Clarke	Chairperson/Advisor
Derek Kosty	Core Member
Geovanna Rodriguez	Core Member
Kathleen Scalise	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

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

Degree awarded June 2024

DISSERTATION ABSTRACT

Cayla Lussier

Doctor of Philosophy

Department of Special Education & Clinical Sciences

June 2024

Title: Examining Relationships Between Components of Implementation Fidelity and Student Response Within the Context of an Early Numeracy Intervention

Evidence-based mathematics interventions are critical for supporting students with mathematics difficulties. In research and practice, collecting implementation fidelity is important for ensuring that all the core components of the intervention are implemented as designed. Historically, implementation fidelity has been defined as multi-faceted including examinations of adherence, instructional quality, and student engagement, though mathematics intervention studies rarely report on fidelity components outside of adherence. The current study examined relationships between these different components of fidelity and whether they are associated with student mathematics outcomes, intervention group size, and interventionist characteristics within the context of a first-grade mathematics intervention. Findings revealed relationships between components of fidelity with student initial mathematics skill, however no relationship was observed between fidelity components and student mathematics growth. Findings for group size and interventionist characteristics were mixed. Limitations, implications for research and practice, and future directions are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Cayla Lussier

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene

DEGREES AWARDED:

Doctor of Philosophy, School Psychology, 2024, University of Oregon
Master of Science, Special Education, 2022, University of Oregon
Bachelor of Science, Psychology, 2017, University of Oregon
Associates of Arts, Central Oregon Community College

AREAS OF SPECIAL INTEREST:

Academic Interventions
- Early Mathematics Intervention
- Early Reading Intervention

Implementation Science
MTSS and RTI Systems

PROFESSIONAL EXPERIENCE:

Graduate Employee, University of Oregon, 2020-2023
Involves assisting with teaching content to school psychology graduate students regarding academic assessment, collaboration, and professional service provision. Involves providing weekly group supervision with a small group of graduate students currently providing services as practicum students. Also involves providing feedback on psychoeducational reports.
Supervisors: Billie Jo Rodriguez, PhD, NCSP, BCBA and Angela Whalen, PhD

Research Assistant, Boston University, 2023-2023
Involves tracking and organizing study data as well as rating participant videos on a researcher-developed rubric. Also, collaborating with project staff and other research assistants to ensure coding/rating reliability.

Practicum Student, Bethel School District, 2020-2021
Involved supporting an on-site school psychologist with student special education evaluations, intervention planning, consultation, and professional development. Additionally, involves attending a variety of multi-disciplinary team meetings and working to develop skills in both academic and behavioral assessment and intervention implementation.
Supervisor: Danea deGlee, NCSP.

GRANTS, AWARDS, AND HONORS

Silvy Kraus Presidential Fellowship in Education, University of Oregon, 2023
Janette Gunther Drew Scholarship, University of Oregon, 2022
Summa Cum Laude, University of Oregon, 2017

PUBLICATIONS:

Sutherland, M., **Lussier, C.**, Nelson, G., Pilger Suhr, M., Turtura, J., & Clarke, B. (under review). A quantitative systematic literature review of self-monitoring components within mathematics instruction and intervention

Turtura, J., Sutherland, M., Doabler, C., Kosty, D., **Lussier, C.**, & Clarke, B. (under review). Measuring the quality of classroom management in an empirically validated Tier 2 kindergarten mathematics intervention

Lesner, T., Sutherland, M., **Lussier, C.**, & Clarke, B. (2023). Using the Number Line to Build Understanding of Fraction Arithmetic. Intervention in School and Clinic, Advance Online Publication. https://doi.org/10.1177/10534512231156878

Sutherland, M., Lesner, T., Kosty, D., **Lussier, C.**, Smolkowski, K., Turtura, J., Doabler, C.T. & Clarke, B. (2022). Examining interactions across instructional tiers: Do features of Tier 1 predict student responsiveness to Tier 2 mathematics intervention? *Journal of Learning Disabilities*, 1-14. https://doi.org/10.1177/00222194221102644

Fainstein, D., **Lussier, C.**, Cook, M., & Men, V. (2021). Professional learning checklist for a remote delivery format. *Literacy Information and Computer Education Journal, 12*(1), 3497 – 3502. https://doi.org/10.20533/licej.2040.2589.2021.0461

ACKNOWLEDGMENTS

I wish to express sincere appreciation and gratitude to all my professors throughout my undergraduate and graduate career for their care and support. Special thanks to my advisor, Dr. Ben Clarke, who has consistently inspired me to think critically about research, supported my interests and career goals, and has always provided quality feedback. I also want to thank the members of my committee for their support with conducting this research and developing this document. The investigation was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R324A090341 and R324A160046 to the Center on Teaching and Learning at the University of Oregon.

Dedicated to my parents, partner, and friends for supporting me through it all.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Improving academic achievement in the areas of language arts and mathematics has been a priority for the United States Department of Education for over a decade (U.S. Department of Education, 2007). Since, additional goals have also included prioritizing high quality elementary education to all students and closing achievement and opportunity gaps between low-performing schools and students (U.S. Department of Education, 2012). In line with these goals, policymakers and leaders in education have also prioritized the use of evidence-based practices and supports for struggling students in schools. Throughout the 2000s various policies and initiatives were put into place to support research on and implementation of evidence-based practices that aim to enhance elementary achievement across multiple academic domains. In 2002, the Education Sciences Reform Act created national centers for researching and promoting evidence-based practices in education through the newly established Institute of Education Sciences (IES). In the same year, the National Assessment of Educational Progress Authorization Act mandated the continued administration (first administered in 1969) of the National Assessment of Educational Progress (NAEP) to track U.S. student's academic progress over time. Furthermore, the Individuals with Disabilities Education Act (IDEA) was reauthorized in 2004 and provided additional guidance and assistance to schools across the US for supporting students with disabilities. In the late 2000s, with these general initiatives in place, efforts were being made to further support the teaching and learning of mathematics.

The National Mathematics Advisory Panel (NMAP) released their report in 2008 which focused on identifying the current evidence base across several areas related to math instruction and student outcomes including curricular content, learning processes, teachers and teacher

education, instructional practices, instructional materials, assessment, and research policies and mechanisms (U.S. Department of Education, 2008). Among the recommendations put forward from NMAP was that "the mathematics curriculum in Grades PreK-8 should be streamlined and should emphasize a well-defined set of the most critical topics in the early grades" (U.S. Department of Education, 2008, pg. 11). Towards this end a state-led initiative to develop the Common Core State Standards (CCSS) was launched in 2009 with the support of 48 states, two territories, and the District of Columbia (Common Core State Standards Initiative, 2010). The standards addressed 3 key shifts including a greater focus on fewer topics, coherence in topics across grades, and the pursuit of conceptual understanding, procedural skills and fluency, and application with equal intensity (Common Core State Standards Initiative, 2010). More recent policies such as the Every Student Succeeds Act (ESSA) continue to prioritize high academic standards and ensure accountability for schools in which groups of students are not making progress (Every Student Succeeds Act, 2015).

How are we, as a nation, progressing toward the goal of supporting mathematics achievement for all elementary-aged students? NAEP, cited above, provides a consistent source of information on the mathematics achievement of 4th, 8th, and 12th grade students across the US over the past 5 decades. NAEP mathematics assesses within the five core content areas of (1) number properties and operations, (2) measurement, (3) geometry, (4) data analysis, statistics, and probability, and (5) algebra. Based on historical NAEP data, 4th grade elementary scores had increased since 1990 with many more students scoring at or above the proficient range. Recently, scores have plateaued with only a one-point increase since 2003. Following the COVID pandemic scores have illustrated a slight decrease in proficiency. Despite the increase in scores over time, the most recent 2022 assessment indicated that 64% of 4th grade students are scoring

below the proficient range, and 43 states had score decreases for this population since 2019. Scores for 4th grade students with disabilities are even lower with 84% scoring below the proficient range. While NAEP provides a snapshot of mathematics proficiency from a large nationally representative sample, it only provides information for mathematics proficiency starting at the 4th grade.

However, we know that mathematics difficulties begin much earlier (Gersten et al., 2005). Findings from longitudinal studies have indicated that foundational math skills in kindergarten predict later math skills in 1st grade and later in 3rd grade (Jordan et al., 2007; Jordan et al., 2009). Furthermore, students that experience early mathematics difficulties often struggle to catch up to typically developing peers (Morgan et al., 2011). As more evidence for the importance of elementary math achievement emerges, work in evidence-based practices that support mathematics learning for all students continues. It is thus imperative that schools are able to support early mathematics learning.

Multi-Tiered Systems of Support

One way to support early mathematics achievement is through the adoption of a multi-tiered framework of instructional support. Many US elementary schools have adopted multi-tiered systems of support (MTSS) to guide the identification and distribution of academic support. Tiered models such as MTSS have been recommended within ESSA as one method of strengthening academic programs (ESSA, 2015). MTSS models generally consist of 3 tiers of support, with tier 1 representing universal support that every student receives, tier 2 representing additional supports provided to students that are not adequately responding to tier 1, and tier 3 representing more intensive and individualized support provided to students that are not adequately responding to tier 2. Core components of MTSS also include universal screening

(typically 3 times a year) to identify students who may be at risk of academic difficulties, continuous progress monitoring of students who are receiving additional supports to track their improvement, data-based decision-making processes, and the use of evidence-based and culturally responsive practices (American Institutes for Research, 2022). A recent study which examined the reported adoption of MTSS and other tiered models across all 50 states found that in 2017, 21 states had publicized the use of MTSS models within their schools (Berkeley et al., 2020). In this paper, Berkeley and colleagues reflected on their previous "snapshot" article published in 2009 and concluded that states had a much greater focus on the three-tiered model in 2017 than when they had observed in 2007.

The following sections include a review of work in early numeracy interventions, adaptations to interventions, and studies that allow researchers to better understand core components of their interventions. Understanding these core components is imperative for supporting intervention implementation. The purpose of the current study is to examine the implementation fidelity of an early numeracy intervention including the relationships between fidelity and intervention outcomes, intervention group size, and interventionist characteristics.

Mathematics Interventions & Early Numeracy

Although not as prevalent as reading, many schools have MTSS frameworks established for mathematics in the early grades (Balu et al., 2015). Within MTSS, a core component is the use of quality, evidence-based interventions that are necessary for supporting children with and/or at-risk for mathematics difficulty. Research-based recommendations from a panel of experts suggest that mathematics interventions should (a) focus on whole numbers in kindergarten through grade 5, (b) be explicit and systematic including utilizing models, verbalizing though processes, guided practice, corrective feedback, and cumulative review, (c)

include instruction on solving word problems, (d) include opportunities for students to work with visual representations, (e) include activities that build fluency in basic arithmetic facts, (f) include progress monitoring, and (g) include motivational strategies (Gersten et al., 2009). In addition to these recommendations, a more recent guide titled "Assisting Students Struggling with Mathematics Intervention in the Elementary Grades" added the recommendations that quality mathematics interventions should include instruction on mathematical language, concrete and semi-concrete representations, and the use of the number line to build students' understanding of material and prepare them for more advanced mathematics (Fuchs et al., 2021). Together, these recommendations provide curriculum developers with research-based strategies to incorporate into their mathematics interventions.

Although there is evidence supporting the effectiveness of specific strategies and instructional elements (Baker et al., 2002; Gersten et al., 2009), interventions themselves also need to be evaluated for effectiveness. A recent meta-analysis examined 34 studies of early numeracy interventions and found a moderate mean effect size (g = .64) suggesting that early numeracy interventions are generally effective although more research is needed to determine the sustainability of student gains over time (Nelson & McMaster, 2019). While large, randomized efficacy trials of mathematics interventions have historically represented a small proportion of published research (Seethaler & Fuchs, 2005), Nelson and McMaster's recent review illustrates an increasing trend in this type of research. Furthermore, when examining the articles included in this review an interesting trend reveals researchers conducting multiple studies of the same intervention with each study examining variations of the intervention. Examining the efficacy of an intervention within various contexts and through systematic alterations provides insight into what works, for whom, and under what conditions (Miller et al., 2014).

Exemplifying this process, a group of researchers from the University of Delaware have explored the efficacy of a number sense intervention for kindergarteners at risk for math difficulties. The team conducted a well-powered randomized control trial with students from schools serving low-income families (Dyson et al., 2011). The intervention included 24 scripted lessons that focused on mathematics vocabulary, identifying numbers 11-19, number sequencing, verbal subitizing, multiple representations, associating numeral to quantity, plus and minus one principle, number comparisons, part-whole relationships and using counting to solve problems. The authors noted that the curriculum was previously field tested to ensure that concepts being covered, and scripting were relevant to students and teachers. Results from this study showed that children in the intervention condition scored significantly higher than children in the control condition on a measure that assessed multiple domains of number sense when controlling for pretest scores. The authors also administered a more generalized measure of mathematics ability, the Applied Problems and Calculation subtests from the *Woodcock-Johnson III Tests of Achievement* (WJ; Woodcock et al., 2007). Results from the WJ subtests revealed non-significant differences at posttest and delayed posttest for the Applied Problems subtest, and delayed posttest for the Calculations subtest. Based on the findings from this study, the next study conducted by a similar author team on the same intervention included modifications to build on previous findings (Jordan et al., 2012). In this study, the authors implemented the same intervention but with revisions to better align the content to the CCSS. Additionally, in the 2012 study, the authors included a third treatment group that received a language intervention to further explore the effects of small group instruction on student gains. Results from this study illustrated significant differences in posttest and delayed posttest on a number sense assessment for students that received the number sense intervention in comparison to students in the control

condition. The authors also noted that this trial resulted in significant differences on the more generalized WJ Applied Problems subtest suggesting an improvement in the intervention from its previous version. This line of research demonstrates the process of refining mathematics interventions. The authors systematically revised their intervention content across studies to better meet the needs of their target population (kindergarten students with mathematics difficulties).

Another line of inquiry within mathematics research for early elementary students exemplifies how researchers can also explore how implementation variation impacts effectiveness and how results vary across instructional contexts. Clarke and colleagues began research on a Tier 2 kindergarten mathematics intervention (ROOTS) in 2016 (Clarke et al., 2016a). The ROOTS intervention is a 50-lesson scripted curriculum that focuses on concepts of whole number and aligns with the CCSS. The program utilizes explicit and systematic instruction and includes various research-based elements described above such as modeling, using visual representations, prompting mathematics verbalizations, providing guided practice, and providing academic feedback. The initial efficacy trial found that at-risk students in the treatment condition that received the ROOTS intervention had statistically significantly higher gains on the Test of Early Mathematics Ability (TEMA) than the control group. From this study the authors went on to conduct a conceptual replication in a different geographical region with students receiving different core instruction programs. Critically, the researchers altered the intervention onset by varying when during the year intervention began. (Doabler et al., 2016). The authors continued to study ROOTS with different samples of students and with varying research questions and designs. Specifically, they examined the long-term effects of the intervention (Clarke et al., 2016b), the relationship between explicit instructional interactions and

student outcomes (Doabler et al., 2017), the effect of intervention group size (Clarke et al., 2017; Clarke et al., 2020; Doabler et al., 2019a), the efficacy for English language learners (Doabler et al., 2019b), the relationship between intervention effects and initial mathematics skill (Clarke et al., 2019; Clarke et al., 2020), and the efficacy of the intervention within different core instruction contexts (Clarke et al., 2022b). This line of research similarly exemplifies the iterative process of intervention efficacy trials with a lesser focus on the mathematical content of the intervention and a larger focus on participant characteristics, intervention procedures, and instructional context.

Within both lines of research described above, researchers documented intervention delivery using a variety of methods to ensure that the interventions were being delivered as designed within the context of the study. Methods utilized included examining adherence to the intervention lesson script via audio recordings (Dyson et al., 2011; Jordan et al., 2012), observing interventionists' adherence to the program (Clarke et al., 2016a), and observations of instructional interactions (Doabler et al., 2017). For adherence, researchers used different methods to obtain indicators of the delivery of core intervention components. For instructional interactions, Doabler and colleagues utilized a coding system that counted the number of instructional behaviors related to core components of the program including teacher models, group responses, individual responses, and teacher-provided academic feedback. Including these measures within their studies helps assure researchers that core components of the intervention are being implemented. It is imperative that researchers are able to document this when examining intervention effectiveness and defining core intervention components.

Core Components

Research studies like those highlighted above can help identify if the intervention is effective and further distinguish which components or aspects of the intervention are contributing to its effectiveness. This process leads to researchers and practitioners better understanding what works, for whom, and under what conditions (Miller et al., 2014; Ochsendorf, 2016). The parts of an intervention that lead to positive outcomes are often referred to as "core components" or "active ingredients". Part of the curriculum development and research process is defining and examining these core components (Clements, 2007). Within the medical research field, *theories of change* (ToC) or *change models* gained popularity as a tool for outlining core components and other important aspects of intervention including outcomes, impacts, population, and resources (Breuer et al., 2016; De Silva et al., 2014). Within education research authors rarely explicitly identified their ToC in published articles (Bos et al., 2022). Including ToC in mathematics intervention research may help readers further understand the core components of the intervention and help researchers draw conclusions based on results that may lead to future replication and program improvement (Kim, 2019). To measure the inclusion and presence of core components within their studies, many researchers utilize measures of implementation fidelity (Abry et al., 2015; Bos et al., 2022).

Implementation Fidelity

Implementation fidelity (IF) or fidelity of implementation has generally been defined as "the determination of how well an intervention is implemented in comparison with the original program design during an efficacy and/or effectiveness study" (O'Donnell, 2008, pg. 33). While this definition provides a summary of how IF is typically described, there have been a wide variety of definitions and conceptualizations across fields and authors, hence making it difficult to pinpoint a universally accepted definition (Sanetti & Kratochwill, 2009). Measures of IF are

imperative within research studies to ensure that the intervention was implemented as designed. When researchers can demonstrate that the core components of the intervention were present within the study, the validity of the study is strengthened (Stains & Vickery, 2017).

In practice, evaluating IF helps to ensure that the interventions being implemented in schools are being implemented as they were in the research studies that found them effective. Translating research to practice requires educators to understand the core components of a program and implement them with fidelity in real-world situations. While it is recommended that practitioners monitor students' progress within an intervention it is equally imperative to also monitor IF. Without evaluating fidelity alongside student progress, it may be unclear if student progress, or lack of progress, is due to intervention or other factors (Hagermoser Sanetti & Collier-Meek, 2019).

Previous research has examined links between IF and student outcomes (Dane & Schneider, 1998; O'Donnell, 2008; Crawford et al., 2012; Hill & Erickson, 2019). O'Donnell (2008) conducted a review of the literature that linked outcomes from K – 12 interventions and IF. The findings from this review suggest that few researchers examined statistical relationships between IF and intervention outcomes. Of the five identified studies that did examine relationships between IF and outcomes included in O'Donnell's review, all showed statistically significantly higher outcomes when the intervention was implemented with higher levels of IF. A recent review from Hill & Erickson examined the relationship between IF and student outcomes among a sample of 37 IES funded studies and 39 studies of preK-12 science, technology, engineering, and mathematics curricula. This review found that compared to studies with low levels of IF, studies with moderate to high levels of IF "had more than double the chance of yielding positive results than null results" (Hill & Erickson, 2019, pg. 593).

While these findings support the idea that high fidelity can contribute to positive student outcomes, the authors also concluded that more finite evidence is needed for determining the degree of fidelity necessary to yield positive outcomes. Additionally, the way that a researcher chooses to measure fidelity matters when determining connections between IF and student outcomes. Researchers have chosen to measure different components of IF including adherence to the program, quality of implementation, and student engagement with the program.

Approaches to Measuring Fidelity

When it comes to measuring IF, many different frameworks and measures have been used across studies. For example, Carroll and colleagues identified a conceptual framework for IF that includes the assessment of program adherence as well as a focus on intervention complexity, facilitation strategies, quality of delivery, and participant responsiveness (Carroll et al., 2007). Similar elements have been identified across other proposed models of IF including adherence, exposure (dosage), quality, participant responsiveness, and program differentiation (Sanetti & Kratochwill, 2009).

Adherence. The most common component of IF that researchers measure is adherence (Bos et al., 2022). Because IF measures core components of an intervention, it has been noted that "fidelity assessments are inherently unique to each intervention and thus rely primarily on guidance from developers" (Abry et al., 2015). Typically, IF measures are created by researchers based on the core components of their intervention as Abry suggested. These measures most often assess "adherence fidelity". Adherence measures assess the extent to which an interventionist implements the intervention as intended. For example, in their examination of a reading comprehension intervention Fogarty and colleagues (2014) discuss an IF assessment procedure in which independent observers recorded adherence data based on the presence of key

components of their intervention. Using this method, researchers can determine if the program was implemented as written and/or designed.

Quality. Researchers have examined different elements of IF beyond adherence. For example, quality of intervention delivery is often included in intervention ToC and has been examined in various ways. Measures of quality typically examine how well steps of the intervention were delivered and in the context of academic interventions may also refer to the quality of instruction. In their examination of a first-grade mathematics intervention, Clarke and colleagues observed intervention sessions and rated the quality of instructional interactions occurring between interventionists and students (Clarke et al., 2014). Other researchers have reported assessing the presence of features of explicit and systematic instruction, and the interventionist's ability to manage students' behavior as quality of implementation indicators (Bryant et al., 2021).

Student Engagement. Another component identified in various IF models is participant (student) responsiveness (Sanetti & Kratochwill, 2009). Student responsiveness has historically been defined as "a measure of participant response to program sessions, which may include indicators such as levels of participation and enthusiasm" (Dane & Schneider, 1998, pg. 45). Between measures of adherence, quality, and student responsiveness, previous reviews of the literature have found that student responsiveness is rarely reported (Bos et al., 2022; Dane & Schneider, 1998). A recent study from Doabler and colleagues operationalized students' responsiveness within the context of a first-grade mathematics intervention via direct observation (Doabler et al., 2021b) as individual practice and group practice opportunities. Study findings revealed that student mathematics gains from pre-test to post-test were associated with the rate of

group practice opportunities, suggesting that student responsiveness relates to intervention outcomes.

As an indicator of student participation and enthusiasm, student responsiveness may also be described as behavioral engagement. Other forms of engagement such as emotional engagement and cognitive engagement have been defined in the literature (Fredericks et al., 2004). Behavioral engagement has been described as students' attention and participation in instruction. This definition aligns most closely to definitions of student responsiveness as a component of IF. Throughout the current study, student responsiveness is defined as student engagement, referring to behavioral engagement during instruction. Ratings of student engagement provide a measure of the level of student participation and enthusiasm in accordance with this definition. Student engagement has been shown to be positively associated with early mathematics achievement gains in previous studies (Bodovski & Farkas, 2007; Robinson, 2013). Including measures of engagement or students' responsiveness provides researchers with one way of understanding how students are experiencing and interacting with the intervention. This data can further inform their understanding of intervention implementation and effectiveness at the student level.

Although rare, some researchers have included multiple measures of IF in accordance with the models discussed above. When assessing for IF in their examination of a reading comprehension intervention, Fogarty and colleagues measured other components of IF beyond adherence including quality and student responsiveness (Fogarty et al., 2014). To assess quality, they utilized a 5-point Likert scale in which independent observers rated the quality of each implemented intervention feature. To assess student responsiveness, independent observers rated the level of total classroom engagement during intervention. Fogarty and colleagues provide a

strong example of how to measure various aspects of IF within the context of a reading intervention, however similar approaches are rarely seen in the contexts of mathematics interventions. Measuring different components of IF (adherence, quality, and student engagement) in this way may help researchers identify patterns of IF and variation in implementation across interventionists and/or groups. For example, some interventionists may score high on adherence to the program and low on their quality of implementation while others may score high or low on both.

Variables Impacting Fidelity

Differences between interventionists when implementing an intervention may occur for a range of reasons. Interventionists often have varying levels of prior experience and confidence with implementing mathematics interventions. Some interventionists report that lack of time can be a barrier to intervention implementation (Strand Cary et al., 2017). Additionally, differences in how the intervention is delivered, such as group size, may also impact an interventionist's implementation of a program. For example, it could be hypothesized that when considering adherence, quality, and student engagement, interventionists may be able to ensure higher rates of IF when working with fewer students at one time. These differences and challenges may lead to variation in implementation based on the competing demands of adherence to the program, providing quality interactions, completing prescribed activities, and keeping students engaged.

An example of how interventionist characteristics may interact with implementation is explored in the work of Dusenbury and colleagues. In the area of drug prevention programing, Dusenbury and colleagues have explored relationships between teacher (implementer) prior experience and their implementation of a Life Skills Training (LST) program (Dusenbury et al., 2005). The authors found that from interviews and direct observations of the participating

teachers, those with more years of prior experience with prevention programs had higher adherence to the program, met more objectives, and had more engaged students than those with fewer prior years of experience. The link between prior interventionist experience and intervention implementation has not typically been examined in the context of mathematics intervention research. However, researchers typically report interventionist experience and/or describe training procedures (Bos et al., 2022; Clarke et al., 2016a). While reporting out prior experience may help contextualize study results, further examination of the impact of interventionist prior experience on IF within mathematics interventions is needed.

In addition to interventionist experience, interventionist perception of the program may also be related to implementation. Program perceptions may include general impressions of the program content or instructional components as well as an interventionist's beliefs on the feasibility of the program. Interventionists with more favorable views of an intervention may be more likely to implement the program as written. For example, Johnson and colleagues found that when teachers were given the choice to implement a preferred classroom management intervention, they did so with greater adherence to intervention procedures and with higher quality than teachers that were not given a choice (Johnson et al., 2014). A recent review of the literature suggests that very few researchers report teacher/interventionist acceptability or perceptions within mathematics intervention studies (Nelson et al., 2022). Further exploring the link between interventionist perceptions within mathematics intervention research may lead to additional insight into implementation and adoption of effective intervention programs.

Outside of interventionist-level characteristics, differences in intervention group size have been examined in relation to intervention effectiveness and IF. Within their examination of a kindergarten mathematics intervention, Clarke and colleagues included two treatment

conditions, one with a 5:1 student to interventionist ratio (larger group) and one with a 2:1 ratio (smaller group) (Clarke et al., 2017). In relation to student outcomes, the researchers found no significant differences in mathematics gains between students in larger groups and students in smaller groups. This finding held true in a more recent replication study from Doabler and colleagues (Doabler et al., 2019a). However, in the context of a first-grade mathematics intervention with students grouped in the same way, statistically significant differences were detected in some measures of student mathematics gains favoring students in the smaller groups (Clarke et al., 2022a). Additionally, researchers examined differences in implementation quality between large intervention groups and small intervention groups. Results revealed statistically significant differences in the number of individual practice opportunities such that students in small groups received more individual practice opportunities than those in large groups (Clarke et al., 2022a; Doabler et al., 2019a). In this most recent examination, Clarke and colleagues also examined differences in adherence fidelity between larger and smaller groups. Results indicated that interventionists teaching the same content and program to smaller groups taught more activities, met more instructional objectives, followed teacher scripting more closely, used more prescribed models, and had overall higher total fidelity scores than those leading larger groups. These results suggest that intervention group size may impact an interventionist's implementation of the intervention.

Fidelity in Mathematics Intervention Research

A recent review of the mathematics intervention literature from 1990 to 2018 examined if mathematics intervention studies included measures of fidelity, and if so, what types of measures (Bos et al., 2022). Based on common IF frameworks, the authors coded studies for components of IF including adherence, quality, and student engagement. However, the authors extended their

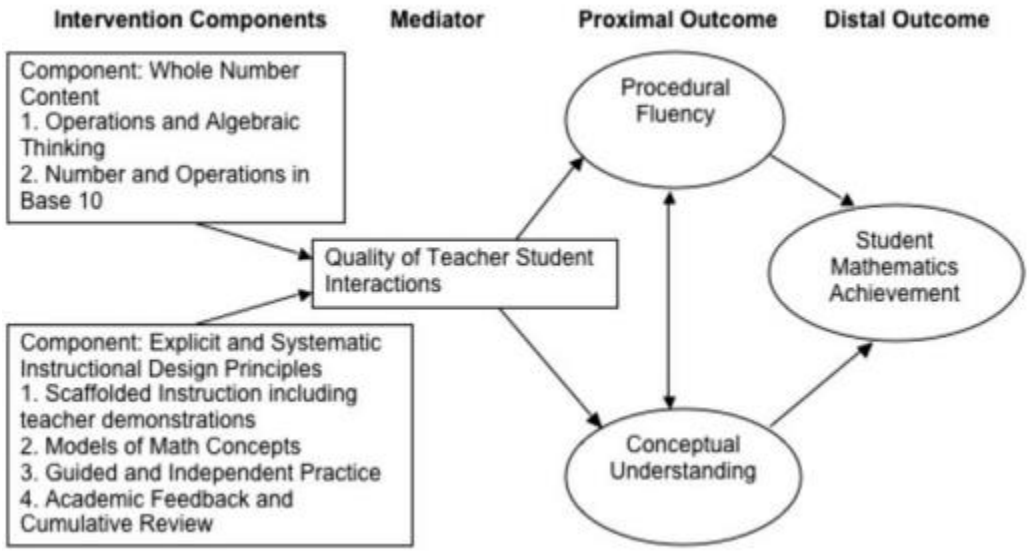
review of IF elements by examining additional IF elements such as the presence of a ToC, described logistics, implementor knowledge, and student engagement. Among the 99 included studies from 1990 to 2018, Bos and colleagues found that only 3 studies included ToCs. While most studies reported teacher experience (82%) only 29% of studies reported on implementer knowledge of the intervention. A large portion (75%) reported collecting quantitative adherence fidelity data. Of 6 total studies that reported quality data, only 4 reported quantitative quality data. Some studies reported a combined fidelity score composed of both adherence and quality but few reported collecting quality data alone. Additionally, only 36% of studies included a report of student engagement, though most were strictly narrative. Bos and colleagues noted that both quality and student engagement data will be important for researchers to consider "when creating the theories of change that elucidate the causal chains that bring about greater student outcomes" (Bos et al., 2022, pg. 14).

Based on Bos and colleagues' findings, very few studies include all relevant aspects of IF. One example of a mathematics intervention study that includes many elements comes from Clarke et al. (2014). Clarke and colleagues conducted a randomized control trial examining the efficacy of a first grade, tier 2 mathematics intervention: Fusion. Results from this study indicated that students in the treatment condition made significantly greater gains than students in the control condition on a proximal measure of conceptual understanding but not on a proximal measure of procedural fluency or a distal measure of conceptual understanding. Within this study, the authors illustrated their ToC model (see Figure 1) which was comprised of core intervention components including both content and process components. The ToC also included quality of teacher student interactions as a mediator of the proximal outcomes (procedural fluency and conceptual understanding). To examine the effects of each component on student

outcomes the authors included a measure of adherence fidelity that provided an overall level of implementation rating and a rating of if the interventionist taught the first three activities in each observed lesson. The authors also reported collecting observation data on the Ratings of Classroom Management and Instructional Support (RCMIS) as a measure of instructional interaction quality. This measure included ratings of instructional quality including the organization of materials, group management techniques, teacher modeling, pacing, practice opportunities, and feedback among others.

Figure 1.

Fusion Intervention Theory of Change Model



Note. Figure retrieved from Clarke et al., 2014.

Additionally, the authors examined the relationships between adherence fidelity scores, quality scores, and student outcomes. Results from this analysis revealed no significant relationship between adherence fidelity scores and student outcomes or between quality of implementation scores and student outcomes. The researchers noted that the lack of significant results may be due to a lack of variation in the fidelity scores as most sessions were scored

highly. This study also provides insight into how mathematics intervention researchers can utilize ToCs that include implementation variables to examine intervention effects.

A more recent study from Nelson and colleagues further explored the relationship between adherence and quality IF on student outcomes within a mathematics intervention while also examining the role of student engagement (Nelson et al., 2020). This study examined adherence, quality, and engagement within the context of a scripted whole and rational number intervention (Math Corps) for students in grades 5-8. Using multi-level regression models, the authors found that in a model containing free and reduced-price lunch status, adherence fidelity, quality, and student engagement, only free and reduced-price lunch status and student engagement were found to have a statistically significant association with student outcomes. Results from this study add to the literature by identifying another component of IF (student engagement) that may be significantly related to student positive response to mathematics intervention. A key takeaway from this work is the importance of including measures of student engagement when examining the relationship between student outcomes and IF.

Purpose of The Current Study

The purpose of the current study is to explore the relationships between IF measures of adherence, quality, student engagement, and student mathematics outcomes within the context of a first-grade mathematics intervention. This study will utilize the framework outlined by Nelson and colleagues (2020) using a multilevel modeling approach to determine the amount of variance in mathematics scores explained by the different components of IF. The current study will extend the work of Nelson et al. by examining the 3 components of IF within a different mathematics intervention (Fusion) and with a first-grade sample. The Fusion intervention is described in more detail in the methods section. Engagement as a component of IF has not been previously

explored in many mathematics interventions (Bos et al., 2022), however it has been identified as a significant predictor of mathematics scores within Nelson and colleague's analysis. It is expected that exploring the role of student engagement within an already established first grade tier 2 mathematics intervention such as the Fusion intervention may help further establish its importance for consideration as a potential piece of future ToCs, providing researchers and practitioners with more strategies to positively affect students' mathematics outcomes.

The current study will address two primary research questions (RQs 1 and 2) and two exploratory research questions (RQs 3 and 4):

1. To what extent do different components of implementation fidelity (adherence, quality, and engagement) correlate with each other?
2. To what extent does each component of implementation fidelity (adherence, quality, and engagement) predict gains in student outcomes?
3. How does intervention group size relate to each component of implementation fidelity at the group-level?
4. How do factors such as interventionist experience and perception of the intervention relate to each component of implementation fidelity at the interventionist-level?

The first two research questions will add to the current literature by examining the relationship between IF components and mathematics outcomes. The third and fourth exploratory research questions will contribute to the literature by providing additional insight into how intervention group size and interventionist-level experience and perception are related to different components of intervention implementation.

Based on previous literature showing moderate correlations between components of fidelity (Abry et al., 2015) it is hypothesized that there will be significant correlations between

the three components of IF. Additionally, it is hypothesized that each component will positively predict mathematics gains, however based on previous findings from Nelson and colleagues not all relationships may be statistically significant (Nelson et al., 2020). Regarding research question three, previous findings suggest statistically significant differences in adherence scores between small 2:1 and large 5:1 intervention groups (Clarke et al., 2022a). It is hypothesized that IF components will be higher for small groups compared to large. Lastly, it is hypothesized that at the interventionist-level, higher IF component scores will be observed for interventionists with higher levels of prior teaching experience and for interventionists with more positive perceptions of the intervention. While these relationships are expected to be positive, there is little previous research on the associations between interventionist characteristics and IF to inform the strength of these relationships.

CHAPTER II

METHODS

The current study analyzed data collected from The Fusion Efficacy Project (Clarke et al., 2016-2020), a multi-year research project funded through IES. The Fusion Efficacy Project included four independent cohorts and utilized a partially nested randomized control trial design blocking on classrooms across cohorts. With this design 970 students were randomly assigned within classrooms to either (1) receive the Fusion intervention in a small group (two students), (2) receive the Fusion intervention in a large group (five students), or (3) a business-as-usual control condition in which students did not receive the Fusion intervention. In total, 194 students making up 97 groups were assigned to the small group intervention condition, 485 students making up 97 groups were assigned to the large group intervention condition, and 291 students were assigned to the business-as-usual control condition. All students were identified as experiencing mathematics difficulty based on their scores on a screening measure. Students included in the treatment conditions received the Fusion intervention in addition to their business-as-usual mathematics instruction. For the purposes of the current study, only data from students in the treatment conditions were analyzed.

Participants

Schools. Data included in the current study was collected from 26 elementary schools representing six school districts in Oregon and Massachusetts. Of the six school districts included two were in large suburban areas in Massachusetts and four were in small and medium sized cities in Oregon. Student enrollment across the participating districts ranged from 5,492 to 40,495 students. Within the participating schools, between 12% to 19% of students had disabilities, 4% to 38% were English learners, and 19% to 65% were eligible for free or reduced-

price lunch. Additionally, between 1% to less than 1% were American Indian or Alaskan Native, 1% to 16% Asian, 1% to 5% Black, 9% to 87% Hispanic, less than 1% to 2% were Native Hawaiian or Pacific Islander, 7% to 73% were White, and 1% to 8% were more than one race.

Classrooms and Teachers. Across the 26 participating schools, the current study includes data from 109 first-grade classrooms. All classrooms operated on a 5-day per week schedule and provided mathematics instruction in English. Teachers utilized a variety of core mathematics curricula such as Houghton Mifflin Harcourt, Engage New York, Envision, iReady, and others. On average, classrooms contained 22.3 first-grade students (*SD* = 5.01). Classrooms were taught by 89 certified teachers, 17 of which participated in multiple years of the Fusion Efficacy Project. A majority of the 89 participating teachers identified as female (95.5%) and White (88.8%). Teachers averaged 13.5 years of teaching experience (*SD* = 8.4) and 8.1 years specifically teaching first grade (*SD* = 6.2). Of the participating teachers 69.7% had a master's degree in education and 58% had completed an advanced mathematics course such as calculus, algebra, statistics, or trigonometry at the college level.

Interventionists. Fusion intervention groups were taught by interventionists that were either district-employed instructional assistants or hired specifically for the study. A total of 87 interventionists participated across cohorts. Among the interventionists a majority identified as female (94.4%) and White (77.5%), with 3.4% identifying as Hispanic, 6.7% two or more races, 2.2% African American, 1.1% Asian American/Pacific Islander and the remaining 8.9% identified as another race or ethnicity or declined to respond. Many interventionists had previous experience with teaching small groups (90.1%) and with math instruction (62.0%). On average, interventionists had 7.3 years of teaching experience (*SD* = 9.4). Of the 89 interventionists,

15.7% had a current teaching license; and 76.3% had taken an advanced math course such as calculus, algebra, and statistics at the college level.

Students. Parental consent was obtained for all participating students. All participating students (2,304 in total) were screened in the fall of their first-grade year. Four measures from the Assessing Student Proficiency in Early Number Sense battery (ASPENS; Clarke et al., 2012) including the Magnitude Comparison, Missing Number, Basic Arithmetic Facts, and Base-10 were administered during the screening process. Students were considered eligible for the Fusion intervention if they had an ASPENS composite score in the *Strategic* (raw score between 13-26) or *Intensive* (below 13) categories based on winter benchmarks. Students who score at or below the *Strategic* category have less than a 50% chance of meeting end-of-year grade level expectations in mathematics (Clarke et al., 2011).

Students who were found eligible for Fusion were ranked in each participating classroom by an independent evaluator. The 10 students with the lowest ASPENS composite scores were then randomly assigned into one of the study conditions: (1) small group Fusion intervention, (2) large group Fusion intervention, or (3) a business-as-usual control condition. Of the 2,304 students screened for eligibility, 1,455 met eligibility criteria. Randomization blocks consisted of the 10 students in each participating classroom with the 10 lowest ASPENS scores. If a classroom had fewer than 10 students eligible for Fusion, classrooms were combined to form virtual randomization blocks. In total, 97 classrooms (including virtual classrooms) were formed containing 10 students each. Students were then randomly assigned within classrooms to the Fusion small group (n = 194), Fusion large group (n = 485), or the control condition (n = 291). Students assigned to the control condition were not further divided into groups. For the current study, only data from the two treatment conditions (large and small Fusion groups) were

included in analyses due to the focus on intervention IF. Demographic data for the 970 randomly assigned students are reported in Table 1.

Procedures

Fusion. The Fusion intervention is a Tier 2 first-grade mathematics intervention focused on teaching whole number concepts and skills. The Fusion intervention is highly scripted and scaffolded for interventionists in that it includes built-in teacher scripting, models, opportunities for practice, and corrective feedback. Fusion is comprised of 60 lessons and can be conceptualized within a three-component framework (1) understanding of whole number concepts and skills, (2) principles of instructional design and delivery, and (3) high-quality instructional interactions. Within the first component, Fusion content is aligned to the CCSS (2010) and includes base 10, place value, number to 100, basic number combinations, operations with 2-digit numbers, story problems, and number properties. Lessons identify and review prerequisite knowledge so that students can access new material. Instructional examples are strategically sequenced across lessons to promote success with new math content and increase in complexity as lessons progress. Lessons 1 – 30 involve naming, counting, writing, and sequencing numbers to 100 and working with number combinations to 5. Lessons 30 – 60 involve comparing numbers within 100, solving 2-digit addition and subtraction problems, solving story problems, mentally adding and subtracting 10, and working with number combinations to 10. Fluency checks are built into the Fusion program to ensure that students are progressing as the lesson complexity increases.

Table 1

Fusion-eligible student-level demographics by condition.

	Control (*n* = 291)		Large Group (*n* = 485)		Small Group (*n* = 194)		Total (*n* = 970)	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Female	159	54.6	246	50.7	111	57.2	516	53.2
Race/Ethnicity								
American Indian or Alaska Native	-	-	2	0.4	2	1	4	0.4
Asian	11	3.8	13	2.7	3	1.5	27	2.8
Black or African American	8	2.7	18	3.7	6	3.1	32	3.3
Hispanic or Latino	73	25.1	130	26.8	49	25.3	252	26.0
Native Hawaiian or Pacific Islander	-	-	5	1	1	0.5	6	0.6
Two or More Races	17	5.8	42	8.7	15	7.7	74	7.6
White	166	57	253	52.2	111	57.2	530	54.6
English as a Second Language	36	12.4	70	14.4	33	17.0	139	14.3
Special Education Status								
Enrolled	47	16.2	76	15.7	23	11.9	146	15.1

Note. Demographic data were not provided for 16 students in the control condition, 22 students in the large group condition, and 7 students in the small group condition.

The second core component of Fusion's conceptual framework, instructional design and delivery, is the use of systematic and explicit instruction including teacher modeling, scaffolding of content, and opportunities for student feedback. Specifically, Fusion utilizes an iterative "I do, we do, you do" format to scaffold students' use of taught strategies. With this format the teacher first models the task, then guides students in practice before students independently practice. Fusion includes mathematical models and representations used to teach mathematical concepts. Commonly used models include the number line, hundreds chart, and place value models. These models are incorporated within a concrete-representational-abstract (CRA) framework. Both the CRA framework and the "I do, we do, you do" format have been recommended as effective strategies for teaching conceptual understanding to diverse learners (Miller & Hudson, 2006). Lastly, the Fusion intervention emphasizes frequent and specific academic feedback. Providing students with specific feedback has been identified as a component in various effective mathematics interventions (Gersten et al., 2009). Within the Fusion intervention, interventionists are instructed to confirm correct student responses and immediately provide corrective feedback for incorrect responses. It is also suggested that interventionists review content that students initially answered incorrectly.

The last piece of Fusion's conceptual framework includes the facilitation of high-quality instructional interactions between students and interventionists. These interactions include opportunities to practice their mathematics skills, engaging in math talk with the interventionist, and interaction with peers. Providing students with frequent practice opportunities is one way that Fusion works to build procedural fluency. As students engage in frequent practice opportunities, they begin to build automaticity with mathematical procedures. The lesson

scripting also facilitates the use of precise mathematical language from the interventionist while also eliciting verbalized mathematics reasoning from students.

Across all four cohorts included in the current study, the Fusion intervention was delivered to groups of students (either small groups of two or large groups of five), five days per week for approximately 12 weeks. Each intervention session was approximately 30-minutes and occurred outside of Tier 1 mathematics instruction. Intervention began for all students in the early winter and ended in the spring to allow students time to respond to core instruction.

Professional Development & Coaching. Two four-hour professional development training workshops were delivered to all interventionists by project staff. The first workshop was held prior to beginning intervention and detailed Fusion lessons 1 – 30. The initial workshop also included an overview of the conceptual framework of Fusion, practice with using response signals, and general lesson structure. The second workshop detailed Fusion lessons 31 – 60 and was held before interventionists began lesson 31. During both workshops, project staff explicitly modeled instructional practices including using group response signals, correcting student errors, and pacing of activities within lessons. Interventionists were provided with opportunities to practice implementing Fusion lessons with feedback from project staff in both workshops. Feedback surveys were distributed to interventionists after both workshops. Generally, interventionists agreed that the training was worthwhile and rated the presentation highly across five components: (1) the training was logical and well organized, (2) the information presented was clear, (3) the presenters had good rapport with participants, (4) the information was relevant and useful, and (5) the participants feel that they can teach Fusion with fidelity.

Project staff also provided coaching support throughout Fusion implementation. Each interventionist received two coaching visits which each consisted of an observation of a lesson

followed by feedback. Some interventionists received additional coaching sessions at the coaches' discretion to provide additional feedback and practice. Coaches provided feedback on the interventionist's implementation of core Fusion principles including organization, student engagement, modeling, practice opportunities, use of response signals, feedback, pace, and completion of prescribed activities. Because interventionists were not required to meet any mastery criteria, coaches did not provide quantitative ratings of interventionist implementation.

Measures

Test of Early Mathematics Ability – Third Edition (TEMA-3). The TEMA-3 was the primary distal mathematics outcome measure for the study. All students were individually-administered the TEMA-3 at pretest in the winter of first grade and at posttest in the spring of first grade. TEMA-3 (Ginsburg & Baroody, 2003) is a standardized, norm-referenced assessment that measures mathematics ability in children ages 3 – 8-years 11 months. Content on the TEMA-3 includes numbering skills, number-comparison, numeral literacy, mastery of number facts, calculation skills, and understanding of concepts. Based on a sample of 1,219 children, alternate-form and test-retest reliabilities are reported at .97 and .82 to .93, respectively. Concurrent validity with other early mathematics assessments ranged from .54 to .91.

Observations of Fusion Intervention

Implementation measures were collected via observation. Each Fusion intervention group was observed approximately three times with approximately three weeks separating each observation. In total 672 observations were completed and 35.1% were coded by an additional independent observer. The average observation lasted approximately 25 minutes. All observers were trained prior to collecting data and were blind to the study goals and research questions. Inter-observer agreement is reported out per measure below.

Fidelity – Adherence. Adherence fidelity was measured via direct observation per the process above. The adherence fidelity measure was researcher-developed and measured the interventionists' implementation of the intervention as intended. The adherence fidelity measure is provided in Appendix A. Observers rated adherence fidelity on a 4-point scale (4 = all, 3 = most, 2 = some, 1 = none). Using this scale, observers rated the extent to which the interventionist (a) met the lesson's instructional objectives, (b) followed the teacher scripting, and (c) used the lesson's prescribed mathematics models. A total adherence fidelity score was computed by calculating the mean score of the three items above. Cronbach's Alpha was calculated at .81 for this measure. Inter-observer agreement was calculated via intraclass correlations coefficients (ICCs) at .95. A stability ICC was also calculated to describe the proportion of variance in adherence between groups versus within intervention groups. Stability was calculated at .29.

Fidelity – Quality. The quality dimension of fidelity was measured using 6 items from The Quality of Explicit Mathematics Instruction (QEMI; Doabler & Clarke, 2012). QEMI items are provided in Appendix B. Items relate to the quality of delivery of instruction including pacing, interventionist modeling, group practice opportunities, individual practice opportunities, academic feedback, and instructional scaffolding. Post observation observers rated each item on a 1 – 4 scale with a score of 1 representing that the item was not present, a score of 2 representing that the item was somewhat present, a score of 3 representing that the item was present, and a score of 4 representing that the item was highly present. A total quality score was derived by calculating the average score across the 6 items. ICCs were calculated to estimate inter-observer agreement. Inter-observer agreement for the six-item version of the QEMI was calculated at .97. Additionally, a stability ICC was also calculated to describe the proportion of

variance in instruction quality between versus within intervention groups. The stability ICC was calculated at .53. Internal consistency of the 6-item measure utilized in this study was calculated at .95 (coefficient alpha).

Fidelity – Engagement. Student engagement was measured via an item from the QEMI (item number 2 in Appendix B). Unlike the other 6 items within the QEMI, engagement was measured from the lens of the student and was therefore pulled out to be analyzed separately from the other interventionist-focused items. The single item was scored by independent observers via the procedures outlined above. Observers rated the level of student participation and engagement based on students' active involvement in the intervention, their compliance with the interventionist, and their completion of work during the lesson. Engagement was rated on a 4-point scale with a score of 1 representing that student engagement was not present, a score of 2 representing that student engagement was somewhat present, a score of 3 representing that student engagement was present, and a score of 4 representing that student engagement was highly present. Inter-observer agreement and stability was calculated using ICCs. Interobserver agreement was calculated at .88 and stability was calculated at .37.

Interventionist Experience. Interventionist surveys were distributed towards the end of the intervention period. Interventionists were asked to provide information regarding their previous teaching experience. Interventionists indicated if they had prior teaching experience, prior experience with teaching small groups, prior experience teaching math, and/or prior experience teaching first grade students. Using this data, interventionists were divided into two categories, high experience and low experience. Interventionists with high experience had prior experience teaching small groups, teaching mathematics, and first grade students (n = 29) and

interventionists with low experience had mixed prior experience teaching small groups, mathematics, or first grade students (n = 32).

Interventionist Perception. Interventionist perceptions of the Fusion program were also collected via surveys distributed towards the end of the intervention period. Perception questions asked interventionists to rate their acceptability of the Fusion intervention for first grade students struggling in math, as well as the feasibility of implementing the intervention. A full list of questions included in the perception section of the survey can be found in Appendix C. Interventionists rated their perceptions on a 1-7 scale with scores of 1-2 representing a currently untrue statement, scores of 3-5 representing a currently somewhat true statement, and scores of 6-7 representing a currently very true statement. Scores were averaged across items to create a total perception score. Total scores ranged from 3 – 7. Cronbach's α for this survey was calculated at .77.

Statistical Analysis

Research Question 1: To what extent do different components of implementation fidelity (adherence, quality, and engagement) correlate with each other? To address this research question, descriptive statistics for all students in the treatment groups who received intervention across fidelity components were computed. Descriptive statistics included means, standard deviations, and ranges for each component of IF (adherence, quality, and engagement) and the TEMA. Descriptive statistics for IF components were computed at the group-level using R Studio. These statistics will be later utilized in the interpretation of IF components across analyses. Parametric (Pearson's r) correlations were conducted between components of IF. Assumptions of normality were assessed by examining the skewness and kurtosis of each variable as well as examining each variable's distribution.

Research Question 2: To what extent does each component of implementation fidelity (adherence, quality, and engagement) predict gains in student outcomes? To address the second research question, a set of multi-level models were analyzed. Within each model, the dependent variable was student TEMA scores. Each model included one of the components of IF (adherence, quality, or engagement) as a level 3 predictor. Utilizing a multi-level model to address this research question accounts for the nested structure of the data. Each model included time (*t*) as a level-1 predictor, students (*i*) at level-2 nested within groups at level-3 (*j*). This structure allows for the control of group-level variables. Each model followed the general structure:

Level 1 Model: $TEMAScore_{tij} = \pi_{0ij} + \pi_{1ij}(Time_{tij}) + e_{tij}$

Level 2 Model: $\pi_{0ij} = \beta_{00j} + r_{0ij}$

$\pi_{1ij} = \beta_{10j} + r_{1ij}$

Level 3 Model: $\beta_{00j} = y_{000} + y_{001}(IFComponent) + u_{00j}$

$\beta_{10j} = y_{100} + y_{101}(IFComponent) + u_{10j}$

Mixed Model: $Y_{tij} = y_{000} + y_{101}(IFComponent) + y_{100}(Time) +$

$y_{101}(IFComponent*Time) + e_{tij} + r_{0ij} + u_{00j}$

To facilitate the interpretation of the intercept at mean levels of implementation within each model, IF component variables were mean centered. Each model contains 4 main effects including the intercept (y_{000}), the effect of the IF component (y_{101}), the effect of time (y_{100}, coded as 0 = pre-test, and 1 = post-test), and the interaction between the IF component and time (y_{101}). Full maximum likelihood estimation was run for each model and r^2-equivalents (Rosnow & Rosenthal, 2003) were calculated after the completion of running each multilevel model. All

analyses were conducted using R Studio software. Assumptions of multilevel models including the calculation of ICCs and homoscedasticity were estimated when running the models.

Research Question 3: How does intervention group size relate to each component of implementation fidelity at the group-level? To address the exploratory third research question multiple independent-means *t*-tests were performed at the group-level to assess differences in IF component means between small Fusion groups (2:1) and large Fusion groups (5:1). Within these analyses each IF component was a dependent variable while the independent categorical variable was group size (either small or large). An independent-means *t*-test was utilized due to the independent nature of the data in which groups were defined as small or large, but never both. Assumptions of normality and homogeneity of variance were assessed prior to running the *t*-tests.

Research Question 4: How do factors such as interventionist experience and perception of the intervention relate to each component of implementation fidelity at the interventionist-level? To address the exploratory fourth research question independent-means *t*-tests were performed at the interventionist level. The dependent variable for each analysis was the IF component (adherence, quality, or engagement) and the independent variable was the categorical variable of interventionist experience (high or low). Assumptions of normality and homogeneity of variance were assessed prior to running the *t*-tests. To assess associations between IF and interventionist perceptions, Pearson's *r* correlations were run between scores on the interventionist perceptions survey and each implementation component.

CHAPTER III

RESULTS

Descriptive Statistics

Prior to running analyses descriptive statistics were computed for each IF component as well as pre- and post-intervention TEMA scores across treatment conditions. Descriptive statistics for student- and group-level variables are displayed in Table 2. IF component descriptives are based on mean ratings aggregated across multiple observations per group. Generally, based on the mean and median values of each component which were all rated on a 4-point scale, ratings were typically high across observations. This is especially true for adherence fidelity with a mean score of 3.4 and a median of 3.4.

Table 2

Individual and group-level descriptive statistics across treatment conditions.

Component	N	Mean	SD	Median	Minimum	Maximum	Skew	Kurtosis
Individual-Level								
Pre-TEMA	663	34.5	7.3	35.0	4.0	54.0	-0.66	0.98
Post-TEMA	612	41.9	8.2	41.0	9.0	65.0	-0.04	0.46
Group-Level								
Adherence	188	3.4	0.4	3.4	1.9	4.0	-0.75	0.42
Quality	188	3.1	0.5	3.0	1.7	4.0	0.09	-0.57
Engagement	188	3.1	0.6	3.0	2.0	4.0	0.00	-0.92

Note. SD = Standard Deviation, TEMA = Test of Early Mathematics Achievement.

Research Question 1: To what extent do different components of implementation fidelity (adherence, quality, and engagement) correlate with each other?

To assess the extent of correlation between the different components of IF, parametric Pearson's *r* correlations were analyzed. Each IF component's skewness and kurtosis values were within ± 2 suggesting adequate distributions for correlation analyses. Overall, correlation coefficients ranged from .60 - .79 suggesting strong correlations between all IF components (Cohen, 1992). There was a strong significant correlation between group adherence fidelity and

group instructional quality ($r = .73$, $p < .001$). The correlation between group adherence fidelity and group student engagement ratings was calculated at $r = .60$ ($p < .001$). Lastly, the strongest correlation was calculated at $r = .79$ ($p < .001$) between group instructional quality and group student engagement.

Research Question 2: To what extent does each component of implementation fidelity (adherence, quality, and engagement) predict gains in student outcomes?

To address research question 2 a set of multi-level growth models were analyzed. Missing data were examined for patterns to determine if data were missing completely at random. Secondly, IF component predictor variables were grand mean centered to better account for scale. Next, models with the predictor variable, time, and a time x predictor interaction term were run. Lastly, diagnostics were examined, and plots were generated. All analyses were conducted in R using the following packages: *lme4, sjPlot, performance, robustlmm,* and *lmerTest*.

When examining missing data, of the 679 participants across treatment conditions, 88% (595) had completed data for all outcome and predictor variables. The 12% incomplete cases represented 6 missing data patterns. Missing data patterns included missing student TEMA scores at pretest, posttest, or both, missing IF component observations, missing IF component observation and TEMA posttest scores, or no missing data across measures. Most missing data included missing post-test TEMA scores. To determine if data were missing completely at random (MCAR), Little's MCAR test was completed to ensure that TEMA means and covariances for students with complete data did not significantly differ when compared to students with incomplete data. Little's MCAR test was non-significant suggesting that data were missing completely at random ($p = .456$).

Each multi-level model included fixed effects for a component of fidelity (adherence, quality, or engagement), time (coded 0 at pretest and 1 at posttest), and the time × predictor interaction predicting TEMA scores. Table 3 displays the unstandardized beta coefficients, their standard errors, significance values, degrees of freedom, intraclass correlation coefficients (ICC), and *r*-squared equivalence values (Rosnow & Rosenthal, 2003) for each model. To check model assumptions diagnostics were run utilizing the *performance* package in R. Across all three models results revealed that the standardized residuals were normally distributed. Additionally, checks for multicollinearity revealed low correlations between fixed effects across models. When utilizing a *z*-score method with the threshold of 2.5 standard deviations above or below the mean results revealed 24 potential outliers or influential observations. When taken out, model estimates did not substantively change thus these observations were kept in the model as they did not influence outcomes. Lastly, all three models violated the assumption of heteroskedasticity (Breusch & Pagan, 1979) and standardized residuals did not appear constant. To further explore the impact of this violation, models were run with robust standard errors utilizing the *robustlmm* package (Koller, 2016). Results for the robust models were comparable to those reported in the original models and the significance levels did not change. Because multilevel models are robust to violations of heteroskedasticity (Schielzeth et al., 2020), results of the original multi-level models are presented.

Table 3

Multilevel analysis results.

Model Parameters	Model 1 – Adherence	Model 2 – Quality	Model 3 – Engagement
Fixed Effects			
Intercept	34.6 (0.3) ***	34.6 (0.3) ***	34.6 (0.3) ***
Time	7.2 (0.2) ***	7.2 (0.2) ***	7.2 (0.2) ***
Predictor	2.6 (0.8) ***	2.4 (0.7) ***	2.1 (0.6) ***
Time x Predictor	0.5 (0.6)	0.4 (0.5)	0.4 (0.4)
Variances			
Group	5.4 (2.3)	5.0 (2.2)	5.4 (2.3)
Student	36.5 (6.0)	36.5 (6.0)	36.3 (6.0)
Residual	16.9 (4.1)	16.9 (4.1)	16.9 (4.1)
p values			
Predictor	.001	< .001	.001
Time x Predictor	.360	.453	.335
df			
Predictor	209.2	215.0	211.5
Time x Predictor	604.7	602.1	603.3
$r^2_{equivalent}$			
Predictor	.047	.059	.052
Time x Predictor	.002	.001	.002
ICC	0.09	0.08	0.09

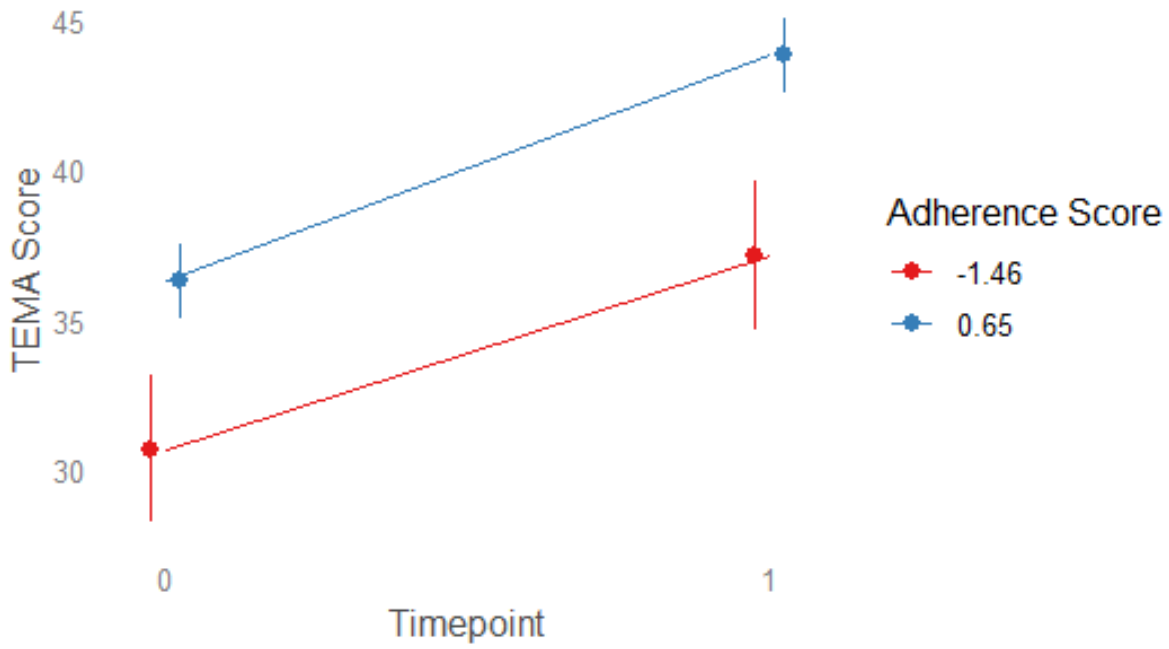
Note. Table cells include non-standardized beta coefficients with standard errors in parentheses. TEMA = Test of Early Mathematics Achievement. *df* = degrees of freedom. ICC = intraclass correlation coefficient.

***$p < 0.001$.

The first model included adherence fidelity as the predictor variable. Results demonstrate that time from pre-test to post-test was a significant predictor of TEMA scores in groups with mean adherence fidelity, $p < .001$. Specifically, students grew an average of 7.2 points on the TEMA from pre-test to post-test. Additionally, adherence fidelity was a significant predictor of TEMA scores at pre-test ($p = .001$, $r^2_{equivalent} = .047$). The interaction between time and adherence fidelity was not significant ($p = .360$, $r^2_{equivalent} = .002$). For every one unit increase in adherence fidelity students grew 0.5 points on the TEMA from pretest to posttest. Furthermore, Figure 2 displays the pretest and posttest TEMA scores for students in intervention

groups with adherence fidelity scores 1.5 points below the mean and 0.7 points above the mean adherence score.

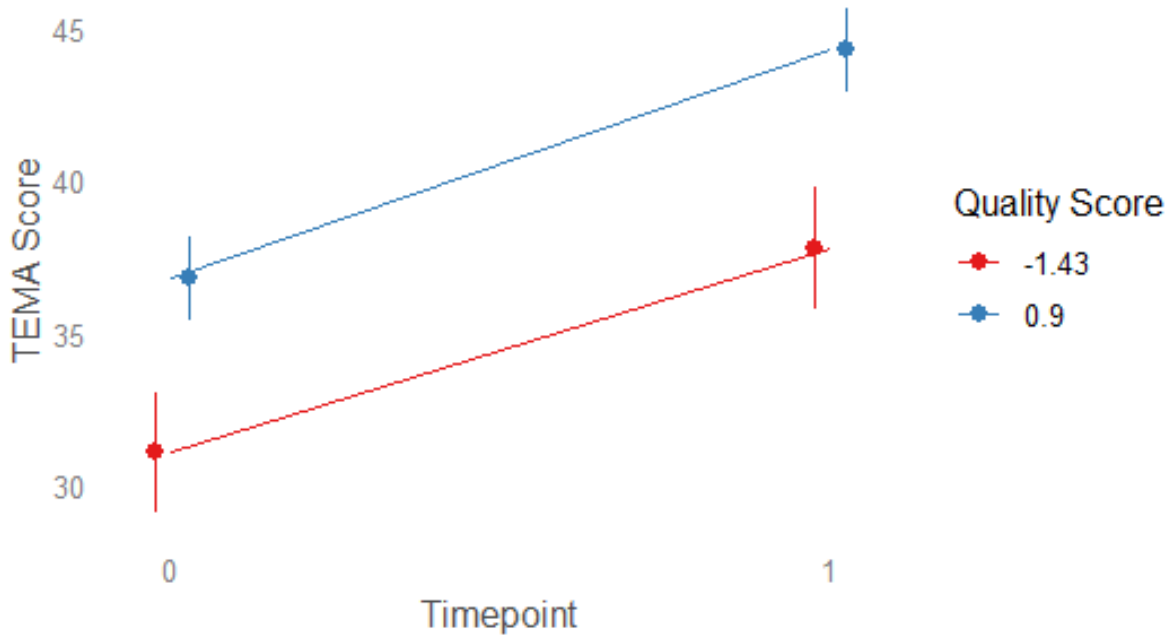
Figure 2
TEMA scores over time across adherence scores.



The second model included instructional quality as the predictor variable. Similarly, to model 1, in model 2 time was a significant predictor of TEMA score ($p < .001$) such that students in groups with average quality scores gained 7.2 points on the TEMA from pretest to posttest. Additionally, quality of instruction was a significant predictor such that students with higher pretest TEMA scores were in groups with higher ratings of instructional quality ($p < .001$, $r^2_{equivalent} = .059$). Also, similarly to the adherence model, the interaction between instructional quality and pre-posttest growth was non-significant ($p = .454$, $r^2_{equivalent} = .001$). A one unit increase in instructional quality was associated with a 0.4 increase in TEMA score from pretest to posttest. Figure 3 displays this relationship by plotting the pretest and posttest TEMA scores

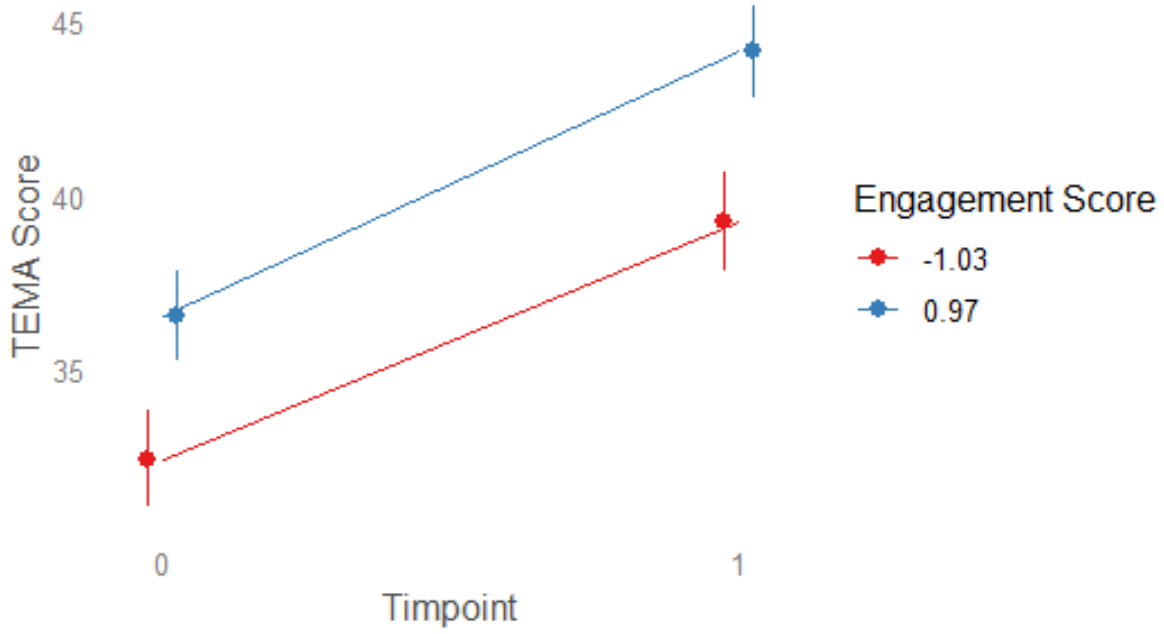
for groups with a quality score 1.4 points below the mean and groups with a quality score of 0.9 above the mean.

Figure 3
TEMA scores over time across quality scores.



Lastly, model 3 included student engagement as the predictor variable. Like in previous models, time was a significant predictor of student TEMA scores ($p < .001$). Students' TEMA scores grew 7.2 points from pre-test to post-test across average engagement scores. Engagement was also a significant predictor such that students in groups with higher engagement scores also had significantly higher pre-test TEMA scores ($p = .001$, $r^2_{equivalent} = .052$). The interaction between time and student engagement was non-significant suggesting that student engagement scores did not predict student mathematics gains from pretest to posttest ($p = .336$, $r^2_{equivalent} = .002$). For every one unit increase in engagement scores, students grew 0.4 points on the TEMA from pre-test to post-test. Figure 4 displays the pretest and posttest TEMA scores for groups with engagement scores 1.03 points below the mean and groups that are 0.97 points above the mean.

Figure 4
TEMA scores over time across engagement scores.



Research Question 3: How does intervention group size relate to each component of implementation fidelity at the group-level?

To assess research question 3 independent-means t-tests were ran for each IF component comparing scores between large intervention groups (5 students) and small intervention groups (3 students). Assumptions of t-tests were assessed by examining Q-Q plots of each fidelity component for normality and testing the homogeneity of variance between large and small intervention groups. Across IF components, Q-Q plots revealed that the data were normally distributed within both small and large groups. The homogeneity of variances between large and small groups was assessed with Levene's Test for Homogeneity of Variance. Results revealed that there was no significant difference between the variances in adherence scores between group

sizes (p = .470). Similar results were found for quality (p = .790) and engagement (p = .525). Having met the assumptions of normality and homogeneity, independent samples t-tests were run between large groups and small groups for each IF component at the group level. Group-level means and standard deviations by group size are presented in Table 4. Across IF components, scores were higher in the small group condition on average.

For large intervention groups with 5 students, the mean adherence fidelity score was 3.31 while for small intervention groups with 2 students, the mean was 3.42. The result from the adherence t-test suggests that there was not a statistically significant difference in adherence fidelity scores between large and small intervention groups, t (186) = -1.82, p = .07, 95% CI [-0.22, 0.01]. A Hedge's g effect size was calculated at 0.27 for this comparison.

The average quality score was 3.09 for large intervention groups and 3.11 for small intervention groups. The result from the t-test comparing instructional quality did not show a statistically significant difference in quality scores between large and small intervention groups, t (186) = -0.24, p = .81, 95% CI [-0.16, 0.13]. A Hedge's g effect size for this comparison was calculated at 0.04.

Student engagement scores were higher in small groups with a mean of 3.15 compared to a mean score of 2.98 for large intervention groups. The t-test result comparing engagement scores between large and small groups was statistically significant, t (186) = -1.99, p = .05, 95% CI [-0.32, -0.00] with small groups having higher student engagement than large groups. A Hedge's g effect size was calculated at 0.31. Table 4 provides a summary of t-tests and effect sizes across fidelity components.

Table 4

T-test results across IF components between large and small groups.

Component	Small Group M (SD)	Large Group M (SD)	*t*	*p*	Hedge's *g*
Adherence	3.42 (0.40)	3.31 (0.42)	-1.82	.07	0.27
Quality	3.11 (0.51)	3.09 (0.50)	-0.24	.81	0.04
Engagement	3.15 (0.54)	2.98 (0.58)	-1.99	.05	0.31

Note. SD = Standard Deviation.

Research Question 4: How do factors such as interventionist experience and perception of the intervention relate to each component of implementation fidelity at the interventionist-level?

Interventionist Experience

To complete analyses for the fourth research question, data were aggregated at the interventionist-level such that IF component scores for interventionists that taught more than one intervention group were averaged. Of the 87 participating interventionists 61 provided prior experience information. Of those 61 interventionists, 29 were classified as having "high" experience due to their previous experience teaching first-grade students, small groups, and mathematics. The other 32 interventionists were classified as having "low" experience as they reported only some prior experience teaching first-grade students, small groups, or mathematics. Table 5 provides descriptive statistics and *t*-test results for both high and low experienced interventionists across IF components.

Table 5

T-test results across IF components for high and low experienced interventionists.

Component	High Experience M (SD)	Low Experience M (SD)	*t*	*p*	Hedge's *g*
Adherence	3.31 (0.42)	3.38 (0.29)	-0.77	.45	0.20
Quality	3.14 (0.47)	3.07 (0.42)	0.62	.53	0.16
Engagement	3.13 (0.51)	2.98 (0.51)	1.11	.27	0.29

Note. SD = Standard deviation.

To ensure that the assumption of normality was met Q-Q plots were examined. Across IF components Q-Q plots revealed that the data were normally distributed within high experience and low experience groups. To ensure that the assumption of homogeneity of variance was met, Levene's test for homogeneity of variance was performed for each fidelity component. For adherence fidelity, Levene's test was non-significant (p = .125). Levene's test was also non-significant for quality (p = .290) and engagement (p = .900), suggesting that the assumption of homogeneity of variances is supported across fidelity components.

The average adherence fidelity score for interventionists with "high" levels of prior experience was 3.31 compared to the average score of 3.38 for interventionists with "low" levels of prior experience. An independent samples t-test between "high" and "low" experienced interventionists was nonsignificant, t (59) = -0.77, p = .45, 95% CI [-0.25, 0.11]. This result suggests that there was no statistically significant difference in adherence fidelity between interventionists with high and low prior experience. Hedge's g was calculated at 0.20.

The average quality score for interventionists with "high" levels of prior experience was 3.14 and the average quality score for interventionists with "low" levels of prior experience was 3.07. The independent samples t-test revealed no statistically significant difference between quality of implementation between interventionists with "high" prior experience and those with "low" prior experience, t (59) = -0.54, p = .54, 95% CI [-0.16, 0.30]. The effect size for this test was calculated at g = 0.16.

Interventionists with "high" levels of prior experience had an average student engagement score of 3.13 and interventionists with "low" levels of prior experience had an average student engagement score of 2.98. Results from the independent samples t-test suggest no statistically significant difference in student engagement scores between "high" and "low"

experienced interventionists, t (59) = 1.11, p = .27, 95% CI [-0.11, 0.41]. The effect size was calculated at g = 0.29.

Interventionist Perceptions

To assess how interventionist perception relates to fidelity of implementation, correlations were performed between interventionist perception scores and each IF component. Complete perception survey data was available for 79 of the 87 participating interventionists. The average perception score was 5.17 with a standard deviation of 0.94, minimum of 3.00, median of 5.31, and maximum of 7.00. The skewness of perception scores was -0.36 and the kurtosis was -0.69. Based on the examination of the distribution of perception scores the data are normally distributed.

Parametric Pearson's r correlations were performed between perception scores and each IF component. Across IF components correlations were non-significant (p-values > .05) and small, ranging from 0.10 to 0.15 with the strongest relative relationship between perception and adherence fidelity (r = .15) and the weakest relative relationship between perception and student engagement scores (r = .10). The correlation between perception and quality was calculated at r = .11. These coefficients suggest little association between the interventionist's perceptions of the intervention and their implementation of the intervention across the various IF components.

CHAPTER IV

DISCUSSION

The current study examined relationships between different components of IF and student outcomes within the context of a highly scaffolded and supported first-grade mathematics intervention. Additionally, the current study explored the relationships between group size, interventionists characteristics, and intervention implementation across the different components of IF. Importantly, this study included measures of IF beyond adherence including quality of instruction and student engagement. These components have been historically identified in conceptualizations of IF, but recent reviews have shown that they are not often evaluated within mathematics intervention research (Bos et al., 2022; Dane & Schneider, 1998). Results indicated that IF components were strongly correlated with each other. IF components were not significantly related to student mathematics growth. However, findings revealed that fidelity scores across IF components were significantly higher for groups with students that had higher initial mathematics skill based on pre-test TEMA score. Results for the relationship between group size and IF components were mixed with small groups receiving higher ratings of student engagement and adherence fidelity compared to large groups but little difference in ratings of quality. Furthermore, results illustrated weak relationships between interventionist variables such as previous experience and perception scores on ratings of IF across the different IF components. These results are discussed further in the sections below alongside study limitations, future research directions, and implications for practice.

Implementation Fidelity Components

Descriptive statistics for each IF component revealed that ratings of fidelity were generally strong. Each IF component was rated on a scale from 1 – 4 and mean scores were all

above 3.0 illustrating high fidelity across components. It is important to consider the context in which these components were rated when interpreting their values. The Fusion intervention program has built in supports including organized lesson structures and scripting. Teacher scripting that is built into the program was designed to ensure a minimum level of fidelity and quality as these components are a large part of Fusion's ToC and conceptual framework. Alongside the supports built into Fusion, the current study provided interventionists with a total of eight hours of professional development which outlined the core components of the intervention program and provided interventionists with the opportunity to receive feedback on their delivery of the program. Interventionists that provided feedback on these trainings rated their ability to implement the Fusion intervention with fidelity strongly. Additionally, interventionists were also provided with multiple coaching sessions during implementation to further support their delivery of the program. It is likely that these built-in and additional supports contributed to the high-fidelity scores observed across adherence, quality, and student engagement. Ratings of IF may be more variable when implemented outside of a research study with the resources to provide this level of support. Ratings may also be more variable when programs do not include built-in supports such as teacher scripting.

Results from the current study also illustrated strong correlations between adherence fidelity, quality of instruction, and student engagement. These results were consistent with previous findings showing significant correlations between measures of fidelity (Abry et al., 2015). The strong correlations between IF components provide additional information regarding the conceptual relationships between different components. In part, these results are expected based on the design of the Fusion intervention program. For example, if an interventionist has high adherence fidelity, that indicates that they're following the lesson scripting and design as

written and because lessons were written with instructional quality and student engagement in mind, it follows that high adherence to the program would also result in high quality and student engagement.

Fidelity Components and Mathematics Outcomes

Results from the multi-level models examining the relationship between IF components and student mathematics outcomes revealed that ratings of IF components did not predict student growth in mathematics skills from pre-test to post-test. The *r*-squared equivalents calculated for each time x predictor interaction term revealed little effect of each IF component on student mathematics growth. Findings from these analyses align well with results from Nelson and colleagues (2020) which also showed nonsignificant relationships between student outcomes and adherence fidelity and quality of instruction. Notably, while Nelson and colleagues found that student engagement predicted post-test mathematics scores, the current investigation did not find that student engagement was a significant predictor of mathematics gains. These results suggest that students can make gains in mathematics skills and knowledge when provided with an evidence-based mathematics intervention and that these gains may not be impacted by variations in implementation. As detailed previously, there was little variation across implementation components and scores were generally high which may have limited the current study's capacity to investigate this research question. Studies of intervention programs with a greater range of IF scores would potentially enable a more thorough investigation of the relationship between IF and student math outcomes.

Results from the multi-level models also revealed a significant relationship between IF components and pre-test mathematics scores. This finding demonstrates that intervention groups with students that had higher initial mathematics scores (as demonstrated by higher pre-test

scores) were in intervention groups with higher adherence, quality, and engagement ratings. This finding aligns well with previous examinations which found that intervention groups with higher initial skill received more practice opportunities (Doabler et al., 2021b). A reasonable hypothesis is that students with higher initial mathematics scores were more likely to be successfully acquiring the skills taught in the intervention and thus more highly engaged during instruction. This level of engagement could have made it easier for interventionists to implement the intervention with adherence and quality. Additional inquiry into the impact of student initial skill on IF is needed to further quantify and understand this relationship.

Group Size and Implementation

Alongside examinations of student outcomes, the current study also included a focus on factors that may relate to intervention implementation including intervention group size. It was initially hypothesized that the small intervention groups would have higher ratings of IF across components, however results were mixed. Specifically, small groups with 2 students had significantly higher student engagement ratings than large groups with 5 students. Additionally, small groups also had higher ratings of adherence fidelity compared to large groups, however this difference was not statistically significant. The Hedge's *g* effect size of .27 for this comparison suggests potential for clinical or practical significance and is similar to previous findings (Clarke et al., 2022a). Differences in instructional quality between small and large groups were minimal and non-significant, suggesting that group size was not related to the quality of instruction. These results were consistent with previous findings that quality of instruction did not differ between small and large Fusion intervention groups (Clarke et al., 2022a).

Overall, these results suggest some variation in IF based on group size. It may be that only having two students in a group allows the interventionist more time to complete more intervention activities thus contributing to slightly higher adherence scores in the small groups compared to the large groups. It may also be easier for interventionists to engage smaller groups of students with more individual opportunities to respond and practice. For example, within their examination of differences in Fusion intervention outcomes by group size, Clarke et al. (2022a) found that students in small groups had greater gains as measured by the TEMA than those in large groups. They also found a statistically significant difference in the number of independent practice rates favoring the small intervention groups over the large groups, but no differences between the overall quality of instruction. These findings along with those illustrated by the current study support further inquiry into the role that group size plays in intervention delivery and student outcomes.

Interventionist Characteristics and Implementation

The current study also examined the relationship between interventionist-level characteristics including prior experience and perception of the intervention and implementation. Results demonstrated a weak relationship between prior experience and IF components. Notably, the Fusion intervention includes built-in supports such as teacher scripting and all interventionists received eight hours of professional development as a part of the study design. Providing professional development and coaching support throughout the intervention may have mitigated the role of prior experience. These findings suggest that with interventionist support such as scripting, professional development, and coaching in place, interventionists can achieve similar levels of implementation regardless of prior experience.

Although differences between “high” and “low” experienced interventionists were not significant across IF components, effect size calculations suggest potential relationships between experience and implementation in practice. Namely, the largest differences between high and low experienced interventionists were in student engagement scores and adherence scores. Interventionists that had greater prior experience scored higher on average than those with lower prior experience on engagement. Differences in adherence scores demonstrated the opposite relationship with interventionists that had low amounts of prior experience scoring higher on average than those with high prior experience. This pattern suggests that prior experience may matter more for engaging students in instruction and less so for adhering to the program and delivering quality instruction. A potential hypothesis for these differences would be that more experienced teachers who have taught first grade students and small groups in the past may have behavior management skills which could allow them to better engage students in the intervention material. Previous studies have noted relationships between behavior management, instructional quality, student engagement, student motivation, and mathematics outcomes (Lekwa et al., 2019; van Dijk et al., 2019). Although not examined in the current study, behavior management skills may be another factor related to interventionist implementation. Other hypotheses could also be made regarding the relatively smaller differences in instructional quality between high and low experienced interventionists. For one, it may be that because many elements of quality instruction are built into the teacher scripting within the Fusion intervention, little prior experience is needed to facilitate quality instructional interactions. Similarly, the level of supports provided within the intervention and professional development may have influenced adherence scores and diminished the role of prior experience to the extent that interventionists with less prior experience had higher levels of adherence. Interventionists with low prior

experience may have also been less likely to make their own adjustments to the intervention and relied more heavily on the intervention scripting, thus resulting in higher adherence. Due to the findings of the current investigation being non-significant, interpretations should be done cautiously and serve as a catalyst for future research directions.

Findings from the current study also suggest little to no relationship between interventionist perception and implementation. Correlations between interventionist perceptions and each IF component were all non-significant and below .20, illustrating a weak relationship. Within the current study, perception scores were relatively high making it difficult to draw conclusions regarding the relationship between lower or negative interventionist perceptions. Additionally, perception surveys were distributed towards the end of intervention implementation, thus ratings did not provide information regarding initial interventionist buy-in. Across examinations of interventionist experience and perception within the current study, fidelity data were collapsed across the two treatment conditions (small group and large group). Because findings from the current study found differences in implementation when interventionists were teaching small groups compared to large groups it may be that group size matters for the strength of the relation between interventionist characteristics and implementation. For example, interventionists that taught small groups may find the intervention easier to implement because they only needed to manage 2 students instead of 5. Additionally, prior experience may become more important when the interventionist is expected to deliver the intervention to 5 students instead of 2. Caution is warranted when making these claims as the current study did not explicitly examine interactions between interventionist characteristics, group size, and implementation.

Limitations

Several limitations should be considered when interpreting findings from the current study. Firstly, ratings of IF were high across components. This limitation is due in part to the high level of interventionist supports that were built into the Fusion intervention, provided through professional development, and provided via ongoing coaching that then aided interventionists in delivering the intervention with fidelity.

While these high ratings are promising when considering the feasibility of the Fusion intervention, they make it difficult to draw conclusions regarding associations between IF components and student outcomes. It was hypothesized that students in intervention groups with high IF ratings would experience greater gains from the Fusion intervention. However, with little variation in IF scores, the current study was limited in investigating the relationship between IF and student outcomes.

Secondly, stability ICCs were low across IF component ratings suggesting little stability in IF ratings across observation sessions. This lack of stability may be in part due to variation in implementation across intervention sessions. The low stability ICCs make it difficult to conclude that each IF component rating is a strong representation of implementation within a group across the course of the intervention. Additional observations of intervention groups could result in higher stability across IF component ratings, however additional observations would also require additional resources from the research team.

Lastly, measures were limited to those collected in the original Fusion efficacy trials. This limitation did not allow for more in-depth measures of IF across components. For example, the engagement measure consisted of a single item rated on a 4-point scale. Additional items assessing student engagement could result in more accurate ratings and could allow for assessment across forms of student engagement (including emotional and cognitive

engagement). A measure of academic engagement that includes ratings of behavioral, emotional, and cognitive engagement has been called for in previous literature (Fredricks et al., 2004). Within the context of early mathematics intervention research, measuring across forms of engagement would require ratings of students' feelings, interests, and attitudes towards math (emotional), their investment in learning and self-efficacy in mathematics (cognitive), and their attention and participation in mathematics instruction (behavioral; Fredericks et al., 2004; Kwan Lo & Foon Hew, 2021).

Measures of interventionist-level characteristics also limited the current investigation's ability to explore relationships between initial perception and different levels of perception and prior experience. As noted earlier, interventionist perception scores were elicited towards the end of intervention implementation. It may be that interventionists perceptions changed from the beginning of implementation to the end, especially as interventionists became more familiar with the intervention procedures and their students. Assessing perception across implementation would provide additional insight into how perceptions change over time and how these perceptions relate to IF. Additionally, ratings of both interventionist perception and experience were relatively high which limited the scope of the current study. Specifically, when dichotomizing interventionist prior experience, the low experience group included interventionists with some prior experience teacher either mathematics, first grade students, or small groups of students. In real world contexts, it is possible that interventionists may have no prior experience with teaching mathematics, small groups, or first grade students however very few interventionists included in the current study fit into this category. Therefore, the current study was unable to fully examine the relationship between a larger range of prior interventionist experiences and IF components.

Future Directions

Several future directions are recommended based on the findings and limitations of the current study. To start, additional examinations of IF components within different settings are needed to further examine the relationships between IF and student outcomes. Specifically, recording IF components within more naturalistic environments that do not provide the same level of interventionist supports found within an efficacy trial where the primary goal is to investigate impact may result in more variation in IF component scores. This variation would allow for a more nuanced analysis of the relationship between IF components and student mathematics gains that more closely aligns with real-world contexts.

Future studies may also include different types of mathematics programs as findings from the current study are also limited to the Fusion intervention specifically. Examining these relationships with interventions across grade-levels and complexity of mathematics content is important for identifying intervention characteristics related to IF. Because the Fusion intervention covers early mathematics content that many interventionists are likely more comfortable and familiar with than more advanced mathematics content, lesson delivery may be easier for interventionists. Further examination of IF components within more complex mathematics interventions is needed to better understand if complexity level has any additional association with implementation. Similarly, examining these relationships within the context of non-scripted programs would also be important. As described earlier, the Fusion intervention is a heavily scripted program, and this level of scripting may assist interventionists in implementation across components. It could be hypothesized that when examined within the context of an early mathematics program that has less built-in support, implementation ratings would be more variable and IF components may have larger associations with student outcomes.

Interventionist experience may also play a greater role in intervention implementation within these different contexts as the level of built-in support decreases and/or the level of complexity increases. By exploring these relationships within various contexts more insight is gained into the magnitude and contexts in which IF matters for student outcomes. Future inquiries into IF components should consider examining thresholds for which IF predicts student outcomes. For example, it is possible that once an interventionist has achieved a certain level of IF there is no additional association with student outcomes.

Additional work is also needed to better understand how individual student differences matter for intervention implementation and mathematics gains. Although the student sample from the current study lacked racial diversity, the current study included diverse learners in that 14.5% of students in the treatment conditions were receiving special education services and 15.1% were identified as English language learners. While results from the current study suggest that overall students made significant mathematics gains from pre-test to post-test across levels of implementation, individual differences were not examined. Future examinations of mathematics intervention implementation should include analyses of individual student characteristics such as disability status, language proficiency, socio economic status, and race/ethnicity. Student-level examinations of engagement for example, may provide additional insight into how educators can differentiate supports and modify intervention procedures to better meet the needs of individual students.

Results from the current study also suggest that more work is needed to develop measures of IF that include multiple IF components. Ratings of adherence fidelity within the current study were collected via a measure that was separate from the measure used for quality and engagement. Results from previous reviews have shown that few mathematics intervention

studies report quality and engagement data while relatively more report adherence fidelity data (Bos et al., 2022). Development of new tools that incorporate multiple components of IF would provide researchers with opportunities for collecting, analyzing, and reporting IF data across components. Within their review, Bos and colleagues (2022) suggest that researchers reflect on their theories of change to identify which key fidelity components need to be measured. Bos and colleagues also call for future mathematics intervention research to include multiple measures of IF, including quality and engagement. As noted above, there have been previous calls for measures of academic engagement to include multiple forms including behavioral, emotional, and cognitive (Fredricks et al., 2004). Including items that address each of these forms within a measure of IF would provide a more diverse and comprehensive rating of student engagement. Based on the low stability ICCs reported within the current study, future research on the development of IF measures should also consider the feasibility of these measures. While Bos et al.'s recent review found that a majority of studies reporting adherence and quality fidelity data utilized live observation as their method for data collection, live observation often can be resource intensive. Live observation often requires a large team of personnel to be trained and requires an acceptable level of inter-rater reliability to be met. When developing new measures of IF researchers need to consider balancing the inclusion of various components with the feasibility of data collection. Alternative methods of data collection such as video recording, audio recording, self-report, and interventionist-report should also be considered to maximize utility and feasibility.

Lastly, findings from the current study revealed several variables that did or did not relate to IF across components. Future research should continue to explore what factors are associated with IF including teacher mathematics knowledge (Sutherland et al., 2022), interventionist

behavior management skill (Lekwa et al., 2019; van Dijk et al., 2019), student initial skill and intervention group composition (Doabler et al., 2021a). Identifying factors that influence interventionist's implementation will provide further assistance to educators implementing mathematics interventions. Based on the limitations of the current study, additional examinations of interventionist perception over time, and a wider range of interventionist prior experience would also be beneficial in better understanding the relationships between perception, experience, and implementation. Future IF research should be conducted in different contexts and with different curricula to further investigate the relationship between IF and factors that may relate to IF outside of the current context.

Implications for Practice

Results from the current study have several implications for practice within school settings. Firstly, the finding that IF components did not predict student outcomes suggests that students can show growth in mathematics skills even with some variation in intervention implementation. This may aid schools in determining the types of interventions and the level of support needed to implement them effectively. Furthermore, the finding that initial mathematics skill matters for implementation can help educators decide where to place support resources. Interventionists teaching students with lower initial skill may benefit from additional implementation supports such as ongoing coaching or more intensive professional development. Educators may also consider differentiated supports in groups with students with lower initial skills to better engage students. Additionally, the correlations between ratings of IF across components may imply that practitioners could utilize adherence, quality, and/or student engagement measures when assessing intervention implementation.

Findings around group size and interventionist-level characteristics also have implications for practitioners within school settings. Firstly, group size was related to some IF components including adherence and student engagement. Educators should take these findings into consideration when determining the sizes of intervention groups. Smaller groups may be warranted when practitioners want to prioritize implementation. This may be especially beneficial when considering the level of intensity needed for specific students as smaller group sizes may allow for more individualization based on previous findings that smaller groups received more individual practice opportunities (Clarke et al., 2022a). More individual practice has been identified as a strategy to intensify supports for students with mathematics difficulties (Fuchs et al., 2017). Because smaller group sizes have been positively related to individual practice opportunities, adherence fidelity, and student engagement between previous studies and the current study, students with lower initial mathematics skill may benefit from participating in smaller intervention groups.

Unlike group size, interventionist-level characteristics did not have strong associations with IF components. These findings have promising implications for educators as they suggest that with supports, including professional development and coaching, interventionists with varying levels of prior experience can implement interventions with fidelity. However, because findings related to experience also revealed that experience tended to matter more for supporting student engagement, schools may also consider additional supports in these areas for interventionists with little prior experience. Schools can then make decisions to leverage the staff they have or hire new staff to support mathematics intervention delivery depending on the resources they have available to support implementation. While acceptability and perception are

important to consider as schools decide which interventions to provide, findings from the current study suggest that interventionists can implement with fidelity regardless of perception.

Conclusion

Within both research and practice, it is imperative that IF is measured and utilized in the interpretation of mathematics intervention effects. Although a majority of mathematics intervention studies only report adherence fidelity data (Bos et al., 2022), conceptual models of IF have historically included components such as instructional quality and student responsiveness (Dane & Schneider, 1998; Nelson et al., 2020). The current study examined relationships between IF across multiple components (adherence, quality, and student engagement) and student mathematics gains within the context of a highly supported and scripted evidence-based mathematics intervention (Fusion; Clarke et al., 2014). The current study also examined the relationships between intervention group size and interventionist characteristics on IF. Findings revealed that ratings of adherence, quality, and student engagement are highly correlated with each other. Results also illustrated that while adherence, quality, and engagement scores were all positively related to student mathematics skills at pre-test, they were not related to student gains in mathematics skills and knowledge from pre-test to post-test. Additionally, while there were differences in IF component ratings favoring small intervention groups over large groups, there was little to no association between interventionist prior experience and perception with implementation. Based on previous reviews (Bos et al., 2022), inquiries (Nelson et al., 2020), and the results from the current study, there's a need for the development and use of IF measures that include multiple components within mathematics intervention research. The field would also benefit from additional explorations of IF components and student outcomes across various contexts, including more naturalistic contexts

where implementation support is more limited and/or variable. Further exploration of relationships between IF components, student outcomes, and other environmental characteristics will aid educators in continuing to support students struggling with mathematics.

APPENDIX A

ADHERENCE FIDELITY ITEMS

Across the lesson activities, the interventionist…

A. Met the math objectives.

B. Followed the teacher scripting.

C. Used the prescribed math models.

1 = None, 2 = Some, 3 = Most, 4 = All

APPENDIX B

QUALITY OF EXPLICIT MATHEMATICS INSTRUCTION (QEMI) ITEMS

1. Efficient delivery of instruction
    a. Uses appropriate pacing; consistent language; minimizes student confusion.
2. Student participation and engagement
    a. Active involvement, compliance, completion of work.
3. Effective teacher modeling and demonstrations
    a. Models skills and concepts clearly; uses math representation effectively.
4. High-quality opportunities for group practice
    a. Offers frequent and rich opportunities for guided and independent practice.
5. Checks of student understanding
    a. Provides timely academic feedback; actively monitors practice opportunities.
6. High-quality practice opportunities for individuals
    a. Distributes individual practice opportunities, both guided and independent.
7. Instructional Scaffolding and Support
    a. Provides adequate think/response time and independent learning opportunities.

1 = Not Present, 2 = Somewhat Present, 3 = Present, 4 = Highly Present

APPENDIX C

INTERVENTIONIST PERCEPTION SURVEY

1. I am able to set aside 30 minutes per day for the Fusion intervention.
2. I see whole number understanding as a high priority for students' mathematical development.
3. I see problem solving skills as a high priority for students' mathematical development.
4. I think that the Fusion intervention would work well as a Tier II or supplemental intervention.
5. I think that the Fusion intervention is feasible for **teachers** to implement during the regular school day.
6. I think that the Fusion intervention is feasible for **instructional assistants** to implement during the regular school day.
7. I have the necessary background knowledge and instructional experience to effectively implement the Fusion intervention.
8. I have the necessary resources to effectively implement the Fusion intervention in my school.
9. I think that the Fusion intervention is appropriate for first grade students who are struggling in early mathematics.
10. I think that the Fusion intervention is appropriate for first grade students who are achieving at grade level in mathematics.
11. I think that the Fusion intervention is appropriate for first grade students who are English learners (Els).

12. I think that the Fusion intervention is appropriate for first grade students who receive special education in mathematics.
13. I currently have access to interventions that focus on developing students' whole number understanding.

0 = Irrelevant, 1-2 = Not true of me now, 3-5 = Somewhat true of me now, 6-7 = Very true of me now

References Cited

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. E. (2015). Using Indices of Fidelity to Intervention Core Components to Identify Program Active Ingredients. American Journal of Evaluation, 36(3), 320–338. https://doi.org/10.1177/1098214014557009

American Institutes for Research. (2022). *Center on multi-tiered system of supports.* https://mtss4success.org/

Baker, S., Gersten, R., & Lee, D. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *The Elementary School Journal, 103*(1), 51-73.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading.* U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED560820.pdf

Berkeley, S., Scanlon, D., Bailey, T.R., Sutton, J.C., & Sacco, D.M. (2020). A snapshot of RTI implementation a decade later: New picture, same story. *Journal of Learning Disabilities, 53*(5), 332-342. https://doi.org/10.1177/0022219420915867

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, 00*(0), 1-21. https://doi.org/10.1177/00222194211065498

Bodovski, K. & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal, 108*(2)

Breuer, E., Lee, L., De Silva, M., & Lund, C. (2016). Using theory of change to design and evaluate public health interventions: A systematic review. Implementation Science, 11(1), 63. https://doi.org/10.1186 /s13012-016-0422-6

Breusch, T. S., and Pagan, A. R. (1979) A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287-1294

Bryant, D.P., Pfannenstiel, K.H., Bryant, B.R., Roberts, G., Fall, A., Nozari, M., & Lee, J. (2021). Improving the mathematics performance of second-grade students with mathematics difficulties through an early numeracy intervention. *Behavior Modification, 45*(1), 99-121. https://doi.org/10.1177/0145445519873651

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40), 1-9. https://doi.org/10.1186/1748-5908-2-40

Clarke, B., Doabler, C. T., Kosty, D., Kurtz Nelson, E., Smolkowski, K., Fien, H., & Turtura, J. (2017). Testing the Efficacy of a Kindergarten Mathematics Intervention by Small Group Size. AERA Open, 3(2), 233285841770689

Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2016a). Examining the Efficacy of a Tier 2 Kindergarten Mathematics Intervention. Journal of Learning Disabilities, 49(2), 152–165. https://doi.org/10.1177/0022219414538514

Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016b). Testing the Immediate and Long-Term Efficacy of a Tier 2 Kindergarten Mathematics Intervention. Journal of Research on Educational Effectiveness, 9(4), 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(2), 129–146. https://doi.org/10.1177/0014402918799503

Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D., Baker, S., Fien, H., & Smolkowski, K. (2014). Preliminary Evaluation of a Tier 2 Mathematics Intervention for First-Grade Students: Using a Theory of Change to Guide Formative Evaluation Activities. School Psychology Review, 43(2), 160–178. https://doi.org/10.1080/02796015.2014.12087442

Clarke, B., Doabler, C.T., Sutherland, M., Kosty, D., Tutura, J., & Smolkowski, K. (2022a). Examining the impact of a first-grade whole number intervention by group size. *Journal of Research on Educational Effectiveness*, 1-24. https://doi.org/10.1080/19345747.2022.2093299

Clarke, B., Doabler, C., Fien, H., & Smolkowski, K. (2016–2020). A randomized control trial of a Tier 2 first grade mathematics intervention (Project No R324A160046, awarded $3,498,258). Institute of Education Sciences (IES): Special Education Research. NCSER-Mathematics, Efficacy and Replication, Goal 3, CFDA No. 84.324. http://ies.ed.gov/funding/grantsearch/ details.asp?ID=1815

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: Implications for Practice and Policy. The Elementary School Journal, 121(1), 125–153. https://doi.org/10.1086/71004

Clarke, B., Rolfhus, E., Dimino, J., & Gersten, R. M. (2012). Assessing student proficiency of number sense (ASPENS). Longmont, CO: Cambium Learning Group, Sopris Learning

Clarke, B., Turtura, J., Lesner, T., Cook, M., Smolkowski, K., Kosty, D., & Doabler, C. T. (2022b). A Conceptual Replication of a Kindergarten Math Intervention Within the Context of a Research-Based Core. Exceptional Children, 89(1), 42–59. https://doi.org/10.1177/00144029221088938

Clements, D.H. (2007). Curriculum research: Toward a framework for "research-based curricula". *Journal for Research in Mathematics Education, 38*(1), 35-70

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101

Common Core State Standards Initiative. (2010). Common core standards for mathematics. Retrieved 12/15/10, from http://www.corestandards.org/the-standards/mathematics

Crawford, L., Carpenter II, 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*(4), 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(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

De Silva, M. J., Breuer, E., Lee, L., Asher, L., Chowdhary, N., Lund, C., & Patel, V. (2014). Theory of Change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials, 15(1), 267. https://doi.org/10.1186/1745-6215-15-267

Doabler, C. T., & Clarke, B. (2012). Quality of explicit mathematics instruction [Unpublished observation instrument]. Center on Teaching and Learning, University of Oregon

Doabler, C.T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. (2021a). Measuring the quantity and quality of explicit instructional interactions in an empirically validated tier 2 kindergarten mathematics intervention. *Learning Disability Quarterly, 44*(1), 50 – 60. https://doi.org/10.1177/0731948719884921

Doabler, C.T., Clarke, B., Kosty, D., Turtura, J.E., Sutherland, M., Maddox, S.A., & Smolkowski, K. (2021b). Using direct observation to document "practice-based evidence" of evidence-based mathematics instruction. *Journal of Learning Disabilities, 51*(1), 20 -35. https://doi.org/10.1177/0022219420911375.

Doabler, C. T., Clarke, B., Kosty, D. B., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2016). Testing the Efficacy of a Tier 2 Mathematics Intervention: A Conceptual Replication Study. Exceptional Children, 83(1), 92–110. https://doi.org/10.1177/0014402916660084

Doabler, C. T., Clarke, B., Kosty, D., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2019a). Examining the Impact of Group Size on the Treatment Intensity of a Tier 2 Mathematics Intervention Within a Systematic Framework of Replication. Journal of Learning Disabilities, 52(2), 168–180. https://doi.org/10.1177/0022219418789376

Doabler, C. T., Clarke, B., Kosty, D., Smolkowski, K., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2019b). Building number sense among English learners: A multisite randomized controlled trial of a Tier 2 kindergarten mathematics intervention. Early Childhood Research Quarterly, 47, 432–444. https://doi.org/10.1016/j.ecresq.2018.08.004

Doabler, C. T., Clarke, B., Stoolmiller, M., Kosty, D. B., Fien, H., Smolkowski, K., & Baker, S. K. (2017). Explicit Instructional Interactions: Exploring the Black Box of a Tier 2 Mathematics Intervention. Remedial and Special Education, 38(2), 98–110. https://doi.org/10.1177/0741932516654219

Dusenbury, L., Brannigan, R., Hansen, W.B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research Theory and Practice, 20*(3), 308-313

Dyson, N.I., Jordan, N.C., Glutting, J. (2011). A number sense intervention for low-income kindergarteners at risk for mathematics difficulties. *Journal of Learning Disabilities, 46*(2), 166-181. https://doi.org/10.1177/0022219411410233

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Fogarty, M., Oslund, 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

Fredericks, J.A., Blumenfeld, P.C. & Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74* (1), 59-109

Fuchs, L.S., Fuchs, D., & Malone, A.S. (2017). The taxonomy of intervention intensity. *Teaching Exceptional Children, 50* (1), 35-43. https://doi.org/10.1177/0040059917703962.

Fuchs, L.S., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K.S., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R., Jayanthi, M., Keating, B., and Morgan, S. (2021). Assisting Students Struggling with Mathematics: Intervention in the Elementary Grades (WWC 2021006). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from http://whatworks.ed.gov/

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). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Gersten, R., Jordan, N.C., & Flojo, J.R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*(4), 293-304

Ginsburg, H. P., & Baroody, A. J. (2003). Test of early mathematics ability- Third edition (TEMA-3). ProEd

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

Hill, H. C., & Erickson, A. (2019). Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review. Educational Researcher, 48(9), 590–598. https://doi.org/10.3102/0013189X19891436

Johnson, L. D., Wehby, J.H., Symons, F.J., Moore, T.C., Maggin, D.M., & Sutherland, K.S. (2014). An analysis of preference relative to teacher implementation of intervention. *The Journal of Special Education, 48*(3), 214-224. https://doi.org/10.1177/0022466913475872

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. Journal of Educational Psychology, 104(3), 647–660. https://doi.org/10.1037/a0029018

Jordan, N.C., Kaplan, D., Locuniak, M.N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research and Practice, 22*(1), 36-46

Jordan, N.C., Kaplan, D., Ramineni, C., & Locuniak, M.N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*(3), 850-867. https://doi.org/10.1037/a0014939

Kim, J. S. (2019). Making Every Study Count: Learning From Replication Failure to Improve Intervention Research. Educational Researcher, 48(9), 599–607. https://doi.org/10.3102/0013189X19891428

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. Journal of Statistical Software, 75(6), 1-24. doi:10.18637/jss.v075.i06

Kwan Lo, C., & Foon Hew, K. (2021). Student engagement in mathematics flipped classrooms: Implications of journal publications from 2011 to 2020. *Frontiers in Psychology, 12:672610*. https://doi.org/10.3389/fpsyg.2021.672610

Lekwa, A.J., Reddy, L.A., & Schernoff, E.S. (2018). Measuring teacher practices and student academic engagement: A convergent validity study. *School Psychology, 34*(1), 109-118. https://doi.org/10.1037/spq0000268

Miller, S.P., & Hudson, P. (2006). Helping students with disabilities understand what math means. *Teaching Exceptional Children*, 28-35

Miller, B., Vaughn, S., & Freund, L. (2014). Learning disabilities research studies: Findings from NICHD funded projects. *Journal of Research on Educational Effectiveness, 7*(3), 225-231. https://doi.org/10.1080/19345747.2014.927251

Morgan, P.L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities, 44*(5), 472-488. https://doi.org/10.1177/0022219411414010

Nelson, G., Johnson, A., & Sawyer, M. (2022). A systematic review of treatment acceptability in mathematics interventions for students with learning disabilities. *Learning Disabilities: A Contemporary Journal, 20*(1), 1-26

Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. Journal of Educational Psychology, 111(6), 1001–1022. https://doi.org/10.1037/edu0000334

Nelson, P. M., Pulles, S. M., Parker, D. C., & Klaft, J. (2020). Implementation fidelity for math intervention: Basic quality ratings to supplement adherence. School Psychology, 35(1), 72–79. https://doi.org/10.1037 /spq0000338

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/10.3102/003465430731379

Ochsendorf, R. (2016). Advancing understanding of mathematics development and intervention: Findings from NCSER-funded efficacy studies. *Journal of Research on Educational Effectiveness, 9*(4), 570-576. https://doi.org/10.1080/19345747.2016.1222144

Robinson, K. (2013). Early disparities in mathematics gains among poor and non-poor children: Examining the role of behavioral engagement in learning. *The Elementary School Journal, 114*(1)

Rosnow, R.L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*(3), 221-237

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward Developing a Science of Treatment Integrity: Introduction to the Special Series. School Psychology Review, 38(4), 16

Scheilzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Allegue, H., Teplitsky, C., Reale, D., Dochtermann, N.A., Garamszegi, L.Z., Araya-Ajoy, Y.G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*, 1141-1152. https://doi.org/10.1111/2041-210X.13434

Seethaler, P. M., & Fuchs, L. S. (2005). A Drop in the Bucket: Randomized Controlled Trials Testing Reading and Math Interventions. Learning Disabilities Research and Practice, 20(2), 98–102. https://doi.org/10.1111/j.1540-5826.2005.00125

Stains, M., & Vickrey, T. (2017). Fidelity of Implementation: An Overlooked Yet Critical Construct to Establish Effectiveness of Evidence-Based Instructional Practices. CBE—Life Sciences Education, 16(1), rm1. https://doi.org/10.1187/cbe.16-03-0113

Strand Cary, M. G., Clarke, B., Doabler, C. T., Smolkowski, K., Fien, H., & Baker, S. K. (2017). A Practitioner Implementation of a Tier 2 First-Grade Mathematics Intervention. Learning Disability Quarterly, 40(4), 211–224. https://doi.org/10.1177/0731948717714715

Sutherland, M., Clarke, B., Kosty, D.B., Baker, S.K., Doabler, C.T., Smolkowski, K., Fien, H., & Goode, J. (2022). Investigating the interaction between teacher mathematics content knowledge and curriculum on instructional behaviors and student achievement. *The Elementary School Journal, 123*(2), 292 – 317. https://doi.org/10.1086/721877

Van Dijk, W., Gage, N.A., & Grasley-Boy, N. (2019). The relation between classroom management and mathematics achievement: A multilevel structural equation model. *Psychology in the Schools, 56*, 1173 – 1186. https://doi.org/10.1002/pits.22254

U.S. Department of Education. (2012). *Strategic Plan for Fiscal Years 2011-2014.* https://www2.ed.gov/about/reports/strat/plan2011-14/plan-2011.pdf

U.S. Department of Education. (2008). *The Final Report of the National Mathematics Advisory Panel*. https://files.eric.ed.gov/fulltext/ED500486.pdf

U.S. Department of Education. (May 2007). *Strategic Plan For Fiscal Years 2007-12*. https://www2.ed.gov/about/reports/strat/plan2007-12/2007-plan.pdf

Woodcock, R.W., McGrew, K.S., & Mather, N. (2007). *Woodcock-Johnson III*. Itasca, IL: Riverside