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

Publication Date: Received 07 Apr 2021, Accepted 10 Jun 2022, Published online: 22 Jul 2022

# Examining the Impact of a First Grade Whole Number Intervention by Group Size

Ben Clarke
University of Oregon

Christian T. Doabler
University of Texas

Marah Sutherland
University of Oregon

Derek Kosty
Oregon Research Institute

Jessica Turtura
University of Oregon

Keith Smolkowski
Oregon Research Institute

The research reported here 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. The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Ben Clarke, and Chris Doabler are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain FUSION-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services. Additionally, the terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research. An independent external evaluator and coauthor of this publication completed the research analysis described in the article.

Correspondence concerning this article should be addressed to Ben Clarke, Center on Teaching and Learning, 1600 Millrace Drive, Suite 207, Eugene, OR 97403. Email: clarkeb@uoregon.edu.

## Abstract

This study utilized a partially nested randomized control design to investigate the impact of Fusion, a first grade math intervention. Blocking on classrooms, students were randomly assigned to one of three conditions: a Fusion two student group, a Fusion five student group, or a no treatment control group. Two primary research questions were examined: What was the overall impact of the Fusion intervention as compared to a business-as-usual comparison condition? and Was there a differential impact on student outcomes between the 2:1 Fusion and the 5:1 Fusion conditions? Analyses found a positive effects on four outcome measures favoring Fusion groups over control with two of the differences statistically significant. Results between Fusion groups found positive effects favoring the Fusion 2:1 group compared to the Fusion 5:1 group on all four outcome measures with two of the differences statistically significant. On a second grade follow up measure no difference was found between Fusion groups and control but a statistically significant difference was found between Fusion groups favoring the 2:1 Fusion group. Future research directions and implications for practice are discussed.


Keywords: Math, Numeracy, Intervention, Group Size

Examining the Impact of a First Grade Whole Number Intervention by Group Size

The importance of a successful start to learning mathematics has been a national priority for several decades (National Mathematics Advisory Panel, 2008; National Research Council, 2001). Mounting evidence indicates that trajectories of mathematics performance are relatively stable across time (Jordan et al., 2010; Morgan et al., 2016; Morgan et al., 2009). Unfortunately, opportunity gaps in mathematics are persistent and wide. Data compiled from the National Assessment of Educational Progress (NAEP) over the past 15 years indicate a consistent score gap of 22-24 points between students eligible and not eligible for the National School Lunch Program (National Center for Education Statistics [NCES], 2019). This may in part be due to substantial disparities in young students' access to early mathematics experiences and instruction (Anders et al., 2012), with preschool-aged students from upper- and middle-class backgrounds already outperforming their economically disadvantaged peers (Griffin et al., 1994; Morgan et al., 2016; Saxe et al., 1987; Starkey et al., 2004). The end result is that large numbers of students, especially those from marginalized and underserved communities, lack the necessary skills to engage in more advanced mathematics as they progress in their schooling.

**Early Mathematics Instruction**

Contributing to the struggle that students experience as they enter formal mathematics instruction in kindergarten is the transition from informal to formal mathematics (Gersten & Chard, 1999). For example, students must map their informal understanding of number or number concepts (e.g., recognizing a group of three objects) onto abstract representations such as numerals (e.g., the numeral "3"). Within the fields of mathematics and special education, significant strides have been made over the past several decades for developing strong intervention programs at the kindergarten level. Numerous mathematics intervention programs

have focused specifically on the development of whole number skills to ease the transition from informal to formal mathematics and to support students at risk for mathematics difficulties in developing mathematics proficiency (Clarke et al., 2016; Dyson et al., 2013; Sood & Jitendra, 2013; Wilson et al., 2009).

While a focus on developing strong kindergarten intervention curricula is undoubtedly important, these targeted efforts may be insufficient to permanently alter long term learning trajectories. One concern is the alignment of intervention programs to content standards as specified in the Common Core State Standards-Mathematics (CCSS-M; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Emerging evidence suggests that the content covered in kindergarten is largely already known by students and that covering this basic mathematics content is negatively associated with student learning (Engel et al., 2016). If intervention programs in kindergarten are not sufficiently challenging, students may be inadequately prepared for the significant increase in expectations and the more challenging content encountered in first grade and beyond (CCSS-M, 2010) . Mathematics content in first grade is increasingly complex and requires students to engage in higher-order thinking and flexible problem-solving with numbers. A brief review of the CCSS-M (2010) reveals a sizeable increase in what students are expected to know and demonstrate in first grade compared to kindergarten. For example, in kindergarten, the Number and Operations in Base 10 standards require students to build foundations for place value understanding for numbers 11 through 19. In first grade, this is expanded to understanding place value for any two-digit number, using place value understanding to add and subtract within 100, and using this knowledge to perform operations such as subtracting multiples of 10 from numbers in the 10-90 range. The accelerated learning progression from kindergarten to first extends into second grade.

For example, by the completion of second grade, students are expected to apply their understanding of place value, properties of operations, and the relationship between addition and subtraction to fluently add and subtract within 100, and to be able to add and subtract within 1000. Consequently, first-grade instruction and intervention programs must be designed to ensure that any deficits from kindergarten are addressed, first-grade content is learned, and the foundation for second grade is firmly established.

Despite the increased complexity and importance of first-grade mathematics, only a handful of researchers have investigated intervention programs at this grade level. Fuchs et al. (2005) evaluated Number Rockets, a small-group tutoring intervention delivered in groups of two to three at-risk first graders. The program centered on building conceptual understanding of mathematical ideas through the use concrete models and manipulatives. Number Rockets consisted of 17 topics taught in 45 sessions conducted over a 16 week period (approximately 3 sessions per week). The 17 topics included a focus on understanding magnitude through the use of a number line, building an understanding of the base 10 system and place value, understanding the concepts of addition and subtraction, and building fluency with basic facts. The program utilized an explicit and systematic instructional design framework (Gersten et al., 2009) with an emphasis on teacher modeling, opportunities to respond, and academic feedback. Each lesson was 40 minutes in length with the final 10 minutes focused on practice with math facts. The researchers primarily used graduate student tutors to implement the program and held weekly coaching sessions to ensure high levels of fidelity to the program. Positive impacts were observed on seven measures representing a range of proximal and distal measures, ranging from 0.11 to 0.70 standard deviations, with six of the seven impacts reaching statistical significance. Across sessions and tutors a high level of implementation fidelity was maintained. A replication

study of the Number Rockets program, conducted by Rolfhus et al. (2012), was implemented across four states. The researchers used conditions more closely approximating authentic educational conditions (e.g., tutors hired from the community, a more typical professional development model). Results indicated a slightly lower levels of fidelity but the program still had a positive effect ($g$ = .34) on a distal general measure of mathematics achievement.

Another evaluation of a first-grade mathematics intervention program was conducted by Bryant and colleagues (2008) using a regression-discontinuity design. The researchers evaluated the effects of a Tier 2 mathematics tutoring program aligned with the Texas Essential Knowledge and Skills standards, taught in groups of four to five students. Program content included counting up/back, number recognition and writing (0-99), number relationships of one and two more/less, understanding base 10 and initial place value, and basic addition and subtraction combinations through the learning of counting and decomposition strategies, number properties, and fact families. An explicit and systematic instructional framework was used and was coupled with an extensive use of a concrete-semi-concrete-abstract (CSA) to model and teach critical concepts (Butler et al., 2003). The program was taught four days per week across 23 weeks by two experienced tutors (one doctoral student and one former kindergarten teacher) who had taught the program previously for two years. Biweekly training on upcoming lessons was provided to the tutors, lasting two hours each session. The researchers found a significant and positive main effect across two of the four researcher-developed measures of the Texas Early Mathematics Inventories: Progress Monitoring, including number sequencing ($\beta$ = .19) and a timed addition/subtraction measure ($\beta$ = .20). An interaction effect was observed on the magnitude comparison measure benefitting students with lower initial pretest scores. No effect was observed on a measure of place value.

More recently, Clarke et al. (2014) conducted a pilot study of Fusion, a Tier 2, 60-lesson first-grade mathematics intervention program. Fusion is focused exclusively on whole number understanding and uses an explicit and systematic design with fully scripted lessons to promote high levels of implementation fidelity. Interventionists included nine experienced district employees who taught the program to groups of approximately five students. Lessons lasted 30 minutes each and were taught three to four days per week, across 20 weeks. The researchers found a significant effect size of Hedges' $g = 0.82$ on a proximal assessment focused on conceptual understanding. Effect sizes, while not significant, on two distal outcome measures were small and positive ($g = 0.11$ to $0.14$). Collectively to date, the efforts of researchers to study first grade mathematics interventions indicate both general promise but also the need to study more nuanced aspects of intervention programs and their use within school systems.

**Supporting Early Mathematics through Multi-Tier Systems of Support**

The provision of intervention programs in school systems is often linked to multi-tier systems of support (MTSS) designed to provide strong core instruction, screening systems to identify students at-risk, and with corresponding interventions and monitoring of student response to those at-risk students (Witzel & Clarke, 2015). Fundamental to these systems is a focus on intervention intensity with intervention intensity increased based on the severity of a student's need and their response to instruction (Fuchs et al., 2018). One mechanism to increase treatment intensity is providing students instruction in groups of decreasing size as the severity of the academic deficit increases. In the traditional MTSS model, this means that group size in Tier 3 is smaller than group size at Tier 2 (Vaughn & Swanson, 2015). The decrease in group size enables greater individualization of instruction and an increase in key teacher and student behaviors hypothesized to support student acquisition of critical instructional content (Codding

& Lane, 2015). The research base has identified key teacher-student interactions (Baker et al., 2002; Doabler et al., 2015; Gersten et al., 2009) that positively impact student mathematics achievement (e.g., teacher models, individual student response opportunities). These interactions can be provided at greater rates within the context of smaller instructional formats or groups. Yet, the provision of smaller groups comes with real and significant opportunity costs to schools (Clarke et al., in press). Given these costs, researchers have begun to explore the role of group size in understanding for whom and under what conditions interventions are effective in MTSS models (Fuchs & Vaughn, 2012).

The vast majority of the work examining group size has been conducted in the area of reading. Utilizing a randomized control trial, Vaughn et al. (2003) randomly assigned second-grade students at-risk in reading to one of three conditions: (a) 1 to 1 instruction, 1:1; (b) a small group of three students, 1:3; or (c) a small group of 10 students, 1:10. In each condition, students received the same intervention focused on critical early literacy skills including phonological awareness, word study, reading fluency, and comprehension. The duration of the intervention was kept constant (i.e., 58 30-min sessions). The use of a standard intervention of set duration across conditions allowed a systematic examination of the impact of group size on student outcomes. Across a range of reading outcome measures significant differences were found between the two smallest groups (1:1 and 1:3) and the 1:10 group. However, Vaughn et al. 2003 reported no differences between the two smallest groups. Results were similar across monolingual English speakers and English learners. In a study of middle school students, at-risk seventh and eighth graders were randomly assigned to large (10-12:1) or small (2-5:1) groups (Vaughn et al., 2010). The study also included a school-based intervention condition. Treatment content and duration were kept constant within the large and small groups. Although more

difficult to evaluate due to a lack of overall significant positive results, no significant group size differences were found on a range of reading outcome measures. In an analysis of interventions targeting oral reading fluency, Begeny and colleagues (2018) found that when interventions were comparable, 79% of students performed equally well in a small group as in a 1:1 instructional setting.

Although there is interest in documenting aspects of treatment intensity in mathematics (DeFouw et al., 2018) to date there have been limited efforts to investigate the impact of group size on student outcomes. Clarke and colleagues (Clarke et al., 2012-2016; Clarke et al., 2017; Doabler et al., 2019) conducted a systematic examination of the impact of a kindergarten mathematics intervention, ROOTS, and group size. ROOTS is a comprehensive intervention consisting of 50 lessons (20 min each) focused on building an in depth understanding of whole number concepts and skills. Using a partially nested randomized control trial, students within classrooms were randomly assigned to one of three conditions (a ROOTS 2:1 group; a ROOTS 5:1 group; or a no treatment control). Overall findings indicated significant positive impacts for ROOTS on a range of proximal and distal measures of mathematics achievement between treatment and control but no significant differences between the two treatment groups of varying sizes (Clarke et al., 2020).

The purpose of the current study was two-fold. First, we sought to examine the overall impact of a first-grade mathematics intervention, FUSION, focused on whole number concepts and skills. This first purpose contributes to the limited but growing research base on first-grade mathematics interventions. The second purpose was to examine whether differences in group size resulted in differing levels of critical teacher-student instructional interactions and student

mathematics outcomes. The second purpose sought to provide insight into how to allocate finite resources within MTSS service delivery systems to best support student achievement.

### Research Questions

1. What was the overall impact of the Fusion intervention as compared to a business-as-usual control condition?
2. Was there a differential impact on student outcomes between the 2:1 Fusion and the 5:1 Fusion conditions?
3. Was there a differential impact on the observed quantity or quality of explicit instructional interactions between the 2:1 Fusion and 5:1 Fusion conditions?

Based on previous results from studies of the ROOTS intervention (e.g., Clarke et al., 2020), we hypothesized that (a) the Fusion intervention program would positively impact student outcomes (b) there would be differences on rates of practice type across the small group conditions and (c) there would be no difference on student mathematics outcomes between the small group conditions.

## Method

### Research Design and Context

This study analyzed data collected from the first two cohorts of a multi-year, four-cohort federally-funded efficacy project involving the Fusion intervention, a Tier 2 first-grade mathematics intervention. The Fusion Efficacy Project (Clarke et al., 2016-2020) employed a partially nested randomized controlled trial (Baldwin et al., 2011). Blocking on classrooms, 460 first grade students were randomly assigned within first-grade classrooms to one of three conditions: (a) 2:1 Fusion group, (b) 5:1 Fusion group, and (c) a no-treatment control condition (i.e., business-as-usual). In all, 92, 230, and 138 students were assigned to the 2:1 Fusion, 5:1

Fusion, and no-treatment control groups, respectively. Students randomly assigned to the two treatment groups received the Fusion intervention in addition to district-approved core mathematics instruction. Collectively, Cohorts 1 and 2 provided a total of 92 Fusion intervention groups (46 = 2:1 Fusion, 46 = 5:1 Fusion). Institutional Review Board approval was obtained for all study methods and procedures and participants were treated in accordance with the ethical principles of the American Psychological Association.

## Participants

**Schools.** Nine elementary schools from three Oregon school districts participated in the current study. One school district was located in the metropolitan area of Portland and two districts were located in suburban areas of western Oregon. Across the three districts, student enrollment ranged from 5,492 to 40,495 students. Within the 9 participating schools, between 12% to 19% of students had disabilities, 11% to 27% were English learners, and 35% to 65% were eligible for free or reduced lunch. Between 1% to less than 1% identified as American Indian or Native Alaskan, 1% to 16% as Asian, 1% to 3% as Black, 20% to 25% Hispanic, 1% to less than 1% as Native Hawaiian or Pacific Islander, 48% to 67% as White, and 7% to 8% as more than one race.

**Classrooms.** The study took place in 53 first-grade classrooms from three school districts in Oregon. All classrooms provided mathematics instruction in English and operated 5 days per week. Classrooms had an average of 21 students ($SD$ = 5.6). The 53 classrooms (Cohort 1 n = 28 and Cohort 2 n = 25) were taught by 36 certified teachers. Of the 36 teachers, 17 participated in Cohorts 1 and 2, 11 participated in Cohort 1 only and 8 participated in Cohort 2 only. 89% of teachers identified as female, 86% as White, 3% Asian American or Pacific Islander, 3% as American Indian or Alaskan Native, 6% as two or more races, and 3% declined to respond.

Teachers had an average of 13.2 years of teaching experience (*SD* = 9.1) and 7.4 years of first grade teaching experience (*SD* = 6.4); 72% had a master's degree in education; and 75% of teachers had completed an algebra course at the college level. Roughly half of teachers reported using teacher created materials. Teachers also reported regular use of commercially-available core instructional programs, including Houghton Mifflin Harcourt, Engage New York, and Curriculum Associates Common Core. Grouping arrangements for instruction included use of 1:1 instruction (67% of classes) and whole group (100% of classes). Roughly half the teachers utilized mathematics centers and engaged in some form of student led instruction.

**Students and inclusion criteria.** In each participating classroom, all students with parental consent were screened in late fall of their first-grade year. A total of 1,076 first-grade students from Cohorts 1 and 2 were screened for Fusion eligibility using the four measures of the first grade Assessing Student Proficiency in Early Number Sense battery (ASPENS; Clarke et al., 2011). These measures included Magnitude Comparison, Missing Number, Basic Arithmetic Facts and Base-10. Students were considered eligible for the Fusion intervention and thus considered at risk for mathematics learning disabilities (MLD) if they had an ASPENS' composite score in the *Strategic* or *Intensive* categories based on winter benchmarks. Composite scores at or below the Strategic category suggest that students have less than a 50% chance of meeting end-of-year grade level expectations in mathematics (Clarke et al., 2011).

Students with ASPENS composite scores in the *Strategic* or *Intensive* categories were rank ordered in each participating classroom by an independent evaluator. Within each classroom, the independent evaluator then randomly assigned the 10 students with the lowest ASPENS composite scores to one of three conditions: (a) Fusion intervention group with a 2:1 student-teacher ratio, (b) Fusion intervention group with a 5:1 student-teacher ratio, or (c) a

control (i.e., business-as-usual) condition. Of 1,076 students screened, 460 met the eligibility criteria. Classrooms with fewer than 10 eligible students were combined to form virtual randomization blocks. This procedure resulted in 46 randomization blocks from the 53 classrooms. Students were then randomly assigned within classrooms or virtual classrooms to the three conditions. Demographic data for the 460 Fusion-eligible students indicated that 17% received special education services, 14% identified as English learners, and 54% as females. While the majority racial group of Fusion students identified as White (65%), 21% as Hispanic, 2% as Black, 3% Asian, 1% as American Indian, and 8% as Multiple Races.

**Interventionists.** Fusion intervention groups were taught by district-employed instructional assistants and by interventionists hired specifically for this study. A total of 39 interventionists taught the 92 Fusion groups. Of the 39 interventionists, five taught 2:1 groups only, four taught 5:1 groups only, and 30 taught both 2:1 and 5:1 groups. Among the interventionists, 94% identified as female and the majority were White (83%), with 3% identifying as Hispanic. The remaining 14% identified as another race or ethnicity or declined to respond. Most interventionists had previous experience providing small group mathematics instruction (72%) and had a bachelor's degree or higher (75%). Interventionists had an average of 5.3 years of teaching experience ($SD$ = 7.2); 19% had a current teaching license; and 64% had taken an algebra course at the college level.

## Procedures

**Fusion Intervention.** Fusion is a 60-lesson, Tier 2 first grade mathematics intervention aimed at building students' proficiency with critical concepts and skills of whole number. Each 30-min lesson addresses mathematical content from two strands focused on whole number understanding of the first-grade Common Core State Standards for Mathematics (CCSS-M,

2010): (a) Operations and Algebraic Thinking, and (b) Number and Operations in Base Ten. Fusion's scope and sequence introduces new concepts and skills in "tracks," with students practicing a variety of different skills each lesson. Activities within lessons build over time as increasingly advanced content is introduced. This sequencing allows for frequent review of previously taught content and supports students' maintenance of mathematical skills.

In Lessons 1-30, students build proficiency with numbers up to 100 through identifying, modeling, writing, and sequencing numbers. Students are also explicitly taught strategies to fluently recall addition and subtraction number combinations within 10. As lessons progress, students encounter increasingly complex content to expand on skills taught earlier in the program. For example, Lessons 1-30 build place value understanding of two-digit numbers, whereas in Lesson 31-60, students learn to solve two-digit addition and subtraction problems and compare two-digit numbers using "greater than" and "less than" terminology. Additionally, in the second half of the program, students expand their repertoire of number combinations, including doubles facts and common number families (e.g., 3, 4, and 7). Lessons 31-60 are also designed to build a deep understanding of mathematical problem-solving. Students are taught the underlying structures of the word problem types identified in the first-grade CCSS-M (2010), and learn to represent and solve add to, take from, put together, and take apart problems.

Fusion incorporates mathematical models through a concrete-representational-abstract (CRA) framework to build students' conceptual understanding (Gersten et al., 2009). For example, when learning place value of two-digit numbers, students use base-ten blocks and unit cubes to model the tens and ones in a given number on a place value chart. In Fusion, the CRA framework is typically applied across multiple lessons, providing students with concurrent exposure to visual representations and mathematical symbols. Other mathematical models used

in the program include number lines, number families, layered place value cards, and a hundreds chart.

To promote high-quality mathematics instruction, the Fusion intervention offers scripted lessons to support teachers in (a) delivering clear demonstrations and explanations of targeted mathematics content, (b) facilitating frequent student practice opportunities, and (c) offering timely academic feedback. The lesson scripting also enables teachers' use of precise and consistent mathematical language and capacity to promote high-quality instructional interactions centered on whole number concepts and skills. These interactions are intended to facilitate deep mathematical thinking and reasoning, through individual or group student mathematics verbalizations (Doabler et al., 2021; Fuchs et al., 2021). For example, when teaching the commutative law of addition, the interventionist writes two problems on the board (e.g., 3 + 1 =, 1 + 3 =) and asks students to discuss with their partner how the two problems are alike. In the latter part of the lesson, students practice explaining the commutative law to their partner using their own words.

In this study, the Fusion intervention was delivered in 30-min, small group formats (i.e., two or five students per interventionist), five days per week for approximately 12 weeks. Because Fusion is designed as a supplemental intervention, instruction occurred at times that did not conflict with core Tier 1 mathematics instruction. For all students, instruction began in early winter and ended in the spring. The early winter start date was selected to provide students with opportunities to respond to core mathematics instruction and therefore minimize the false identification of typically-achieving students during the screening process.

**Professional development.** All interventionists participated in two 4-hr professional development workshops delivered by project staff. The first workshop was held prior to the start

of implementation of the Fusion intervention and focused on content from Lessons 1-30, whereas the second focused on Lessons 31-60. Both workshops centered on validated practices in early mathematics instruction, small-group instruction, and classroom management. Staff leading the workshops explicitly modeled instructional practices, such as group response signals, immediate correction of student errors, and pacing of activities within lessons. Interventionists were provided opportunities to practice and receive feedback on lesson delivery from project staff. To promote implementation fidelity and enhance the quality of instruction, all interventionists received, on average, two coaching visits from coaches during Fusion implementation. Coaching visits consisted of direct observations of lesson delivery, followed by feedback on instructional quality and fidelity of Fusion implementation.

**Fidelity of implementation**

Fidelity of Fusion implementation was measured via direct observations by trained research staff, with each Fusion group observed three times during the course of the intervention. On a 4-point scale (4 = all, 3 = most, 2 = some, 1 = none), observers rated the extent to which the interventionist (a) met the lesson's instructional objectives, (b) followed the lesson's teacher scripting, and (c) used the lesson's mathematics models. Observers also recorded whether the interventionist taught the number of activities prescribed in the lesson. Overall, the majority of prescribed activities were taught with high levels of fidelity ($M$ = 3.4, $SD$ = 0.5). Interventionists were found to meet instructional objectives ($M$ = 3.4, $SD$ = 0.5), follow scripting ($M$ = 3.2, $SD$ = 0.6), and use prescribed models ($M$ = 3.5, $SD$ = 0.5). However, the 2:1 small groups experienced more activities taught ($g$ = 0.46, $p$ = .0306) and had greater total fidelity ($g$ = 0.25) than the 5:1 groups. The 2:1 small groups were also rated higher on meeting instructional objectives ($g$ = 0.36), following teacher scripting ($g$ = 0.08), and using prescribed models ($g$ = 0.23).

**Student Mathematics Outcome Measures**

Students were administered four mathematics outcome measures at pretest ($T_1$) in January of first grade and posttest ($T_2$) in May of first grade and one measure at a delayed posttest ($T_3$) in February of second grade. All measures focused on critical whole number concepts and skills and were administered by trained research staff, who met an interscorer reliability criteria ≥.85 for all assessments.

**ProFusion** is a researcher-developed assessment designed to assess students' conceptual and procedural knowledge of number and numeration, place-value concepts, basic number combinations, and problems involving multi-digit addition and subtraction. In an untimed, small group setting, students are asked write numbers from dictation and numbers missing from a sequence, write numbers matching base-10 block models, and decompose double-digit numbers. Moreover, students complete addition and subtraction problems, and two word problems. Students also complete 1-min, timed addition and subtraction fluency measures and work with proctors individually to complete a set of number-identification items. The correlation between pretest and posttest ProFusion scores was .63 and the standardized alpha for the subscales was .75. Criterion validity of ProFusion with other mathematics outcome measures, including the SAT-10, ranges (r) from .56 to .68 (Clarke et al., 2014).

**Test of Early Mathematics Ability-Third Edition** (TEMA-3; Ginsburg & Baroody, 2003) is a standardized, norm-referenced, individually administered measure of beginning mathematical ability. The TEMA-3 assesses mathematical understanding at the formal and informal levels for children ranging in age from 3 to 8 years 11 months. The TEMA-3 addresses children's conceptual and procedural understanding of math, including counting and basic calculations. The TEMA-3 reports alternate-form and test-retest reliabilities of .97 and .82 to .93,

respectively. For concurrent validity with other mathematics outcome measures, the TEMA-3 manual reports coefficients ranging from .54 to .91.

**ASPENS** (Clarke et al., 2011) is a set of CBMs validated for screening and progress monitoring in first grade mathematics. Each 1-min fluency-based measure assesses an important aspect of early numeracy proficiency, including number identification, magnitude comparison, missing number identification, and arithmetic facts and base-10. Test authors report test-retest reliability ranges from the .70s to .90. Criterion concurrent validity with the TerraNova 3 is reported as ranging from .51 to .63.

**easyCBM Math** (Alonzo et al., 2006) is an online benchmark screening and progress monitoring system for kindergarten to eighth grade. The test items are multiple choice and testing occurs on a secure web site. Reliability and validity of the assessments are well established. Internal reliabilities of first grade easyCBM Math measures are high (.81-.84). Concurrent validity of easyCBM Math scores on the winter benchmark, with the Stanford Achievement Test, Tenth Edition (SAT-10), ranges from .75 to .82. In the current study, the first-grade easyCBM measure was administered at posttest, whereas the second-grade version served as the follow-up assessment in second grade.

**Observations of Fusion Instruction**

Each Fusion group was observed approximately three times over the course of the intervention, with approximately three weeks separating each observation occasion. A total of 274 observations were conducted, of which 70 (26%) included two observers who simultaneously evaluated inter-observer agreement. Observations were scheduled in advance and observers remained for the duration of Fusion instruction, with an average observation lasting

25.4 minutes (*SD* = 2.6 min.). Trained observers, who were blind to our research hypotheses, conducted all observations using two observation measures.

**The Classroom Observations of Student-Teacher Interactions-Mathematics** (COSTI-M; Doabler et al., 2015-2017; Smolkowski & Gunn, 2012) is a low-inference observation instrument that has been empirically validated to document the frequency of teacher demonstrations, individual and group student practice opportunities, teacher-provided academic feedback, and student mistakes. As documented by the COSTI-M, teacher models represent a teacher's verbalizations of thought processes and physical demonstrations of mathematical content. For example, observers coded a teacher model if the teacher explicitly described the structural features of an "add to" word problem. Academic feedback was operationalized as a teacher's verbal reply or physical demonstration to affirm or correct a student response. For example, observers recorded an academic feedback code if the teacher restated an correct answer or corrected a student error. Group practice opportunities were defined as a mathematics-related verbalization produced by two or more students in unison. Individual practice opportunities were coded whenever a single student had the opportunity to verbalize or physically demonstrate her mathematical thinking, such as when a teacher asked a specific student to answer a mathematical question (e.g., "Lamar, use the place value bocks to show 82?"). Rates per minute for each targeted behavior were computed as the frequency of the behavior divided by the duration of the observation in minutes. Doabler et al. (2015) reported predictive validity of the COSTI-M with the TEMA-3 ($p$ = .004, Pseudo-$R^2$ = .08) and the EN-CBM ($p$ = .017, Pseudo-$R^2$ = .05).

**The Quality of Explicit Mathematics Instruction** (QEMI; Doabler & Clarke, 2012) comprises seven items that target the quality of explicit instructional interactions, including group and individual practice opportunities, student participation, teacher modeling, academic

feedback, efficiency of instructional delivery, and instructional scaffolding. Internal consistency of the measure was high, .94 (coefficient alpha). To rate the quality of each item, observers used a 4-point rating scale, with scores of 1–2 representing the lower quality range and 3–4 representing the upper quality range. A Total QEMI score was computed as the mean across all items. The mean across the three observations was used in subsequent analyses.

**Observation training.** Trained observers conducted all direct observations. The observers included former educators, doctoral students, faculty members, and experienced data collectors. Observers received approximately 10 hours of training, with an initial training lasting six hours and a four-hour follow-up training prior to the third round of observations to recalibrate observers, help minimize observer drift, and increase interobserver reliability. Training focused on direct observation procedures, kindergarten mathematics, and use of the COSTI-M and QEMI observation instruments. Prior to observing classrooms on their own, observers were required to complete two reliability checkouts and meet an interobserver agreement criterion of .85 or higher on each checkout. The first was a video checkout, which had observers code a 5-minute video of kindergarten mathematics instruction. Second, observers completed a real-time classroom checkout with a primary observer from the research team. All observers met the minimum interobserver agreement level for both checkouts.

**Inter-observer agreement and stability intraclass correlations coefficients (ICCs).** To estimate inter-observer agreement in observation measures, we calculated ICCs to describe the proportion of variance in each observation measure occurring between versus within paired observation occasions. Inter-observer agreement ICCs for COSTI-M and QEMI scores ranged from .77 to .99, which based on guidelines proposed by Landis and Koch (1977) represented substantial to nearly perfect agreement. To estimate stability across time, we calculated ICCs to

describe the proportion of variance in each observation measure occurring between versus within Fusion groups. Stability ICCs were .24 for teacher demonstrations, .11 for individual practice, .34 for group practice, .20 for student mistakes, .59 for academic feedback, and .58 for the QEMI scale. Reliability of mean scores across the three observation occasions were fair and ranged from .26 (for individual practice) to .80 (for QEMI scores).

**Statistical Analysis**

Analyses were conducted to address three research questions. First, we assessed overall Fusion intervention effects, with 2:1 and 5:1 Fusion groups as the intervention condition, on student outcomes using a mixed model (multilevel) Time × Condition analysis (Murray, 1998) designed to account for students partially nested within small groups (Baldwin et al., 2011; Bauer et al., 2008). The study design called for the randomization of individual students to receive Fusion, nested within 2:1 or 5:1 Fusion groups, or a nonnested comparison condition, and the analytic model must account for the potential heterogeneity among variances across conditions (Roberts & Roberts, 2005). In particular, the Fusion groups required a group-level variance, while the unclustered controls did not. Furthermore, because the residual variances may have differed among conditions, we tested the assumption of homoscedasticity of residuals. The analysis tested for differences among conditions on gains in outcomes from the fall ($T_1$) to spring ($T_2$) of first grade and is described in detail by Clarke et al. (2016) and Doabler et al. (2016). The statistical model included time, coded 0 at $T_1$ and 1 at $T_2$; condition, coded 0 for control and 1 for Fusion; and the interaction between time and condition. These models test for net differences between conditions (Murray, 1998), which provide an unbiased and straightforward interpretation of the results (Allison, 1990; Jamieson, 1999). For two outcomes not collected at pretest— first-grade easyCBM and second-grade easyCBM—we used the

analysis of covariance approach described by Bauer et al. (2008) and Baldwin et al. (2011).

Second, we examined the effects of the 2:1 versus the 5:1 Fusion group size on student outcomes using a fully nested mixed-model (multilevel) Time × Group Size analysis (Murray, 1998) to account for the intraclass correlation associated with students nested within Fusion groups. Similar to the first set of analyses, the model included time, coded 0 at $T_1$ and 1 at $T_2$; group size, coded 0 for 5:1 Fusion and 1 for 2:1 Fusion conditions; and the interaction between time and group size. Mixed analysis of covariance models were used for the first-grade and second-grade easyCBM scores.

Third, we tested whether 2:1 and 5:1 Fusion groups experienced differential rates of observed instructional interactions using independent-samples *t* tests.

**Model estimation.** We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute, 2016) using restricted maximum likelihood (REML), generally recommended for multilevel models (Hox, 2002). Maximum likelihood estimation for the Time × Condition analysis uses of all available data to provide potentially unbiased results even in the face of substantial attrition, provided the missing data were missing at random (Graham, 2009). We did not believe that attrition or other missing data represented a meaningful departure from the missing at random assumption, meaning that missing data did not likely depend on unobserved determinants of the outcomes of interest (Little & Rubin, 2002). The majority of missing data involved students who were absent on the day of assessment (e.g., due to illness) or transferred to a new school (e.g., due to their family moving).

The models assume independent and normally distributed observations. We addressed the first, more important assumption (Van Belle, 2008) by explicitly modeling the multilevel nature of the data. In a sensitivity analysis, accounting for the classroom-level resulted in negligible

differences in effect size and their statistical significance, so the classroom-level was excluded. The data in the present study also do not markedly deviate from normality; skewness and kurtosis fell with ± 1.0 for all measures. Nonetheless, multilevel regression methods have also been found quite robust to violations of normality (e.g., Hannan & Murray, 1996).

**Effect sizes.** To ease interpretation of intervention effects, we computed Hedges' *g* effect sizes (Hedges, 1981) and Improvement Index values as recommended by the WWC (2017) . The Improvement Index represents the expected change in percentile rank for an average comparison group student if the student had received Fusion. Although we set alpha to .05 for all statistical tests, we provided unadjusted and adjusted *p*-values based on the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

## Results

Table 1 presents means, standard deviations, and sample sizes for the five dependent variables by assessment time and condition. In what follows, we present results from tests of bias due to attrition, efficacy effects for Fusion (Research Question 1), effects of the 2:1 versus 5:1 Fusion group size on student outcomes (Research Question 2), and differential rates of instructional interactions between the 2:1 and 5:1 Fusion conditions (Research Question 3).

### Attrition

The overall rate of missingness at posttest was 6.3% for the measures available at pretest, and the difference in rates of missingness among study conditions was below 1.0% for posttest measures. "The proportions of the treatment and control groups that provide information are not particularly important, at least for internal validity" (Foster & Bickman, 1996, p. 698), so we tested the potential for *differential attrition effects*, which may threaten internal validity. This attrition analysis tests whether pretest measures were associated with (a) study condition (Fusion

versus control in this case), (b) attrition status, and (c) the interaction between the two (Biglan et al., 1987; Graham & Donaldson, 1993). This specific analysis used a mixed-model analysis of variance designed to test whether outcome variables were differentially affected across study conditions by attrition while accounting for the nested structure of the data. We found no statistically significant interactions between attrition and study condition predicting baseline outcomes ($p > .500$), suggesting that the effect of attrition on outcomes would not likely threaten internal validity.

**Effects of Fusion versus Control on Student Outcomes**

Table 2 presents the results of the partially nested statistical models comparing gains between nested Fusion students and unclustered control students. The table presents the results of the homoscedastic model for each outcome because it was deemed equivalent to the more complicated heteroscedastic model. The bottom two rows of the table show the likelihood ratio test results that compared homoscedastic residuals to heteroscedastic residuals.

The models in Table 2 tested fixed effects for differences among conditions at pretest (condition effect), gains across time for the control condition (time effect), and differential gains for the Fusion condition (Time × Condition interaction). We found no statistically significant differences in mathematics outcomes at pretest (|Hedges' $g$|'s $\leq 0.04$ and $p$'s $> .6971$ for all pretest measures), suggesting similar mathematics outcomes at pretest by condition. We found statistically significant differences by condition in gains from fall to spring for two dependent variables. Students in the Fusion condition made greater gains than control students on the ASPENS ($t(250) = 2.43$, $p = .0160$) and ProFusion assessment ($t(213) = 9.02$, $p < .0001$). We did not detect statistically significant differences between conditions in gains on TEMA-3 ($p = .3729$) or differences between conditions on posttest first-grade easyCBM scores ($p = .8132$) or

follow-up second-grade easyCBM scores ($p$ = .9135), both tested with TEMA-3 and ASPENS scores as pretest covariates. The Time × Condition model estimated differences in gains between the Fusion and control conditions of 0.6 for the TEMA-3 (Hedges' $g$ = 0.07, 95% CI [-0.09, 0.23], Improvement Index = 2.9%), 3.6 for the ASPENS ($g$ = 0.20 [0.04, 0.36], Improvement Index = 7.9%), and 9.2 for ProFusion ($g$ = 0.77 [0.60, 0.93], Improvement Index = 27.8%). The analysis of covariance model estimated differences between Fusion and control conditions of 0.1 for the first-grade easyCBM ($g$ = 0.02 [-0.16, 0.20], Improvement Index = 0.9%) and -0.1 for the second-grade easyCBM ($g$ = -0.01 [-0.20, 0.18], Improvement Index = -0.4%).

**Effects of 2:1 versus 5:1 Fusion Groups on Student Outcomes**

Table 3 presents the results of the fully nested Time × Group Size models comparing gains between 2:1 and 5:1 Fusion groups. The models in Table 3 tested fixed effects for differences among group sizes at pretest (2:1 Fusion group effect), gains across time for the 5:1 Fusion condition (time effect), and differential gains for the 2:1 Fusion condition (Time × Group Size interaction). We found no statistically significant differences in outcomes at pretest ($|g|$'s ≤ 0.10 and $p$'s > .3800 for all pretest measures), suggesting similar mathematics outcomes at pretest across group sizes.

Students in 2:1 Fusion groups made greater gains than students in 5:1 Fusion groups on the TEMA-3 ($t(89) = 2.21$, $p$ = .0296), and scored higher on the first-grade easyCBM ($t(80)$ = 2.36, $p$ = .0205) and the second-grade easyCBM ($t(83) = 2.77$, $p$ = .0069), both tested with TEMA-3 and ASPENS scores as pretest covariates. We did not detect statistically significant differences among group sizes in gains on ASPENS ($p$ = .1040) or ProFusion ($p$ = .1096). The Time × Group Size model estimated differences in gains between 2:1 and 5:1 Fusion conditions of 1.8 for the TEMA-3 ($g$ = 0.21 [0.02, 0.39], Improvement Index = 8.2%), 2.9 for the ASPENS

($g$ = 0.17 [-0.04, 0.37], Improvement Index = 6.7%), and 1.95 for ProFusion ($g$ = 0.16 [-0.04, 0.37], Improvement Index = 6.5%). The analysis of covariance model estimated differences between 2:1 and 5:1 Fusion groups of 1.5 for the first-grade easyCBM ($g$ = 0.29 [0.05, 0.53], Improvement Index = 11.4%) and 1.7 for the second-grade easyCBM ($g$ = 0.34 [0.09, 0.58], Improvement Index = 13.1%).

**Effects of Group Size on the Quantity and Quality of Explicit Instructional Interactions**

Table 4 presents descriptive statistics for the quantity (rates per minute) and quality of explicit instructional interactions as well as results of independent-samples $t$ tests comparing these observation measures by Fusion group size. Compared to the 5:1 Fusion groups, 2:1 Fusion groups experienced higher rates of individual practice ($t(89) = 4.96$, $p < .0001$, $g$ = 1.04 [0.62, 1.46]). Non-significant differences were observed for teacher model rates (g = -0.11 [-0.53, 0.30] favoring the 5:1 Fusion groups), group practice rates ($g$ = -0.25 [-0.67, 0.16], favoring 5:1 Fusion groups), academic feedback rates ($g$ = 0.13 [-0.29, 0.55], favoring 2:1 Fusion groups), student error rates (g = -0.11 [-0.53, 0.30], with more errors occurring in the 5:1 Fusion groups), and the overall quality of explicit instruction rating ($g$ = 0.07 [-0.35, 0.49], favoring 2:1 Fusion groups).

## Discussion

Comparing Fusion conditions to control, we found positive impacts ($g$ = 0.02 to 0.77) on all four first grade measures with two of the impacts statistically significant. The greatest impact was found on the proximal ProFusion measure ($g$ = 0.77) with lower impacts on distal measures. At second-grade follow up, there was not a statistically significant difference between the Fusion and control conditions. When comparing between Fusion small group conditions, outcomes favored the 2:1 Fusion group on all four first grade outcomes ($g$ = 0.16 to 0.29) with two differences reaching statistical significance. Greater impacts were found on distal measures

(TEMA-3 and first-grade easyCBM) of mathematics achievement. Second-grade follow up results showed a statistically significant difference favoring the 2:1 small group ($g$ =0.34) on the second-grade easyCBM outcome measure. In terms of instructional delivery, across treatment conditions overall fidelity was strong. However, the 2:1 small groups completed significantly more activities and a global indicator of fidelity was rated significantly stronger along with multiple components of overall fidelity, including meeting instructional objectives, the use of mathematics models and following teacher scripting. An overall rating of quality of explicit instruction found no significant differences between treatment conditions. However, a significant difference was observed on the number of individual student practice opportunities, favoring the 2:1 group condition. In the remainder of the discussion, we interpret the study's findings, note limitations to the work, and suggest directions for future research.

**Implications and Future Research**

The overall positive impact of the Fusion intervention adds to the growing research base on early elementary intervention programs in mathematics with the majority of work focused on kindergarten (Nelson & McMaster, 2019). The results from this study augment the limited research to date that has been conducted in first grade (e.g., Bryant et al., 2011; Fuchs et al., 2005). Across intervention programs, similar design features include the use of systematic and explicit instruction (Fuchs et al., 2021) and a focus on whole number content (Frye et al., 2013; Gersten et al., 2009). The overall positive impact of Fusion supports the continued importance of adherence to these general principles when considering how best to support the mathematics learning needs of students at-risk.

Group size results ran counter to the hypothesis that there would be no differences between the 2:1 and 5:1 Fusion conditions. Of note, was the finding that differences reached

statistical significance on two distal measures of achievement in first grade and, of particular interest, on a second-grade follow up measure of mathematics achievement. The goal of early intervention is to alter learning trajectories, yet to date little impact has been found for early mathematics interventions on long term outcomes (Bailey et al., 2020). Thus, the finding that a long term impact was found is important both from a research standpoint but also from a practical one as schools weigh how to best to structure resources and spend finite resources (Clarke et al., in press). Future research on early mathematics interventions should default towards the inclusion of follow-up measures of intervention impact (e.g., in the subsequent grade). While the cause of fadeout is potentially based on an array of factors and how best to sustain intervention effects is complex (Bailey et al., 2017), the finding here suggests that there may be malleable factors directly controlled by schools (i.e., group size) that can be manipulated in ways that meaningfully impact later student mathematics outcomes. Future research should continue to investigate and explore factors that predict, sustain, and impact intervention effectiveness as the field grapples with the best approach to designing interventions and environments for long-term student success (Bailey et al., 2020).

The findings also illustrate the importance of not extrapolating beyond the results of a specific study or a limited research base. To date, findings indicate a lack of group size differences in reading interventions (e.g., Vaughn et al., 2010). More relevant to the current research is work testing the efficacy and impact of group size in the context of kindergarten mathematics (Clarke et al., 2017; Clarke et al., 2020; Doabler et al., 2019). Given the similarities shared between these kindergarten studies and the current research (i.e., randomized control trials, small-group intervention formats, whole number focused interventions), the contrast in outcomes regarding group size is more striking. Reasons for the different outcomes are

speculative but may range from the increased complexity of first-grade mathematics content to implementation variables. For example, across the ROOTS kindergarten studies (i.e., Clarke et al., 2017; Clarke et al., 2020; Doabler et al., 2019) and the current research, 2:1 small groups had greater rates of independent practice opportunities compared to 5:1 groups. In discussing the role of individual practice opportunities in kindergarten, it was hypothesized that a threshold effect may exist in which additional practice opportunities do not increase student understanding (Clarke et al., 2017; Doabler et al., 2019). However, with more complex content it could be hypothesized that the additional individual practice opportunities are necessary to support student acquisition of key first-grade concepts and skills. Findings from the current study point to the need to continue investigating group size in different grade levels to further flesh out the interaction between group size, intervention content, and other critical variables including teacher-student instructional interactions. Such work is critical in light of a growing emphasis on exploring the conditions under which interventions are effective (Miller et al., 2014) and their relative cost compared to alternate treatments (Levin & Belfield, 2015).

**Limitations and Conclusion**

A limitation to the current work was conducting the study in one geographic region of the U.S., resulting in a sample demographically non-representative of the general population. Future research should investigate the Fusion intervention across different geographic regions and with diverse demographic samples intentionally selected to increase the generalizability of results. The importance of systematic replication is garnering increased focus (Chhin et al., 2018; National Science Foundation et al., 2018) to address the current lack of replication research (Cook et al., 2014) and it's perceived value (Makel & Plucker, 2014). The importance of replication is highlighted by the key findings of this study which add to the research base on

effective mathematics interventions. The intervention was generally effective but results varied by group size in direct contrast to previous investigations of group size (Clarke et al., 2020). Critically, the difference in group size results was present at both immediate and delayed posttest. Given the paucity of sustained intervention effects (Bailey et al., 2020), this finding is noteworthy for considering how schools can best to support long term mathematics development through the allocation of finite resources. Additional direct and conceptual replications (Coyne et al., 2016) will aid in furthering investigating Fusion and other early mathematics intervention as the field attempts to better understand the conditions under which interventions are effective and how they are delivered in schools to maximize student outcomes.

## References

Allison, P. D. (1990). Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology, 20*, 93–114. https://doi.org/10.2307/271083

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. University of Oregon. http://easycbm.com

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*(2), 231–244. https://doi.org/https://doi.org/10.1016/j.ecresq.2011.08.003

Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2020). Prevention: Necessary But Insufficient? A 2-Year Follow-Up of an Effective First-Grade Mathematics Intervention [https://doi.org/10.1111/cdev.13175]. *Child Development, 91*(2), 382–400. https://doi.org/https://doi.org/10.1111/cdev.13175

Baker, S. K., Gersten, R. M., & Lee, D.-S. (2002). A synthesis of empirical research on teaching mathematics to low- achieving students. *Elementary School Journal, 103*, 51–73. https://doi.org/10.1086/499715

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

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43*, 210–236. https://doi.org/10.1080/00273170802034810

Begeny, J. C., Levy, R. A., & Field, S. A. (2018). Using Small-Group Instruction to Improve Students' Reading Fluency: An Evaluation of the Existing Research. *Journal of Applied School Psychology, 34*(1), 36–64. https://doi.org/10.1080/15377903.2017.1328628

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289–300. http://www.jstor.org/stable/2346101

Biglan, A., Severson, H., Ary, D., Faller, C., Gallison, C., Thompson, R., Glasgow, R., & Lichtenstein, E. (1987). Do smoking prevention programs really work? Attrition and the internal and external validity of an evaluation of a refusal skills training program. *Journal of Behavioral Medicine, 10*(2), 159–171. https://doi.org/10.1007/BF00846424

Bryant, D. P., Bryant, B. R., Gersten, R. M., Scammacca, N., & Chavez, M. M. (2008). Mathematics intervention for first- and second-grade students with mathematics difficulties: The effects of tier 2 intervention delivered as booster lessons. *Remedial and Special Education, 29*, 20–32. https://doi.org/10.1177/0741932507309712

Bryant, D. P., Bryant, B. R., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. (2011). Early Numeracy Intervention Program for First-Grade Students with Mathematics Difficulties. *Exceptional Children, 78*, 7–23. https://doi.org/10.1177/001440291107800101

Butler, F. M., Miller, S. P., Crehan, K., Babbitt, B., & Pierce, T. (2003). Fraction instruction for students with mathematics disabilities: Comparing two teaching sequences [https://doi.org/10.1111/1540-5826.00066]. *Learning Disabilities Research & Practice, 18*(2), 99–111. https://doi.org/https://doi.org/10.1111/1540-5826.00066

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher, 47*(9), 594–605. https://doi.org/10.3102/0013189X18788047

Clarke, B., Cil, G., Smolkowski, K., Sutherland, M., Turtura, J., Doabler, C. T., Fien, H., & Baker, S. K. (in press). Examining the Cost Effectiveness of a Kindergarten Mathematics: Implications for Practice and Policy. *School Psychology Review*.

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., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the Immediate and Long-Term Efficacy of a Tier 2 Kindergarten Mathematics Intervention. *Journal of Research on Educational Effectiveness, 9*(4), 607–634. https://doi.org/10.1080/19345747.2015.1116034

Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012-2016). *A randomized control trial of a tier 2 kindergarten mathematics intervention* (Project No R324A120304, awarded $3,338,552). USDE; Institute of Education Sciences; Special Education Research, CFDA No. 84.324A http://ies.ed.gov/funding/grantsearch/details.asp?ID=1327.

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), 1–16. https://doi.org/10.1177/2332858417706899

Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D. B., Baker, S. K., Fien, H., & Smolkowski, K. (2014). Preliminary evaluation of a tier-2 mathematics intervention for first grade students: Utilizing a theory of change to guide formative evaluation activities. *School Psychology Review, 43*, 160–177. https://doi.org/10.1080/02796015.2014.12087442

Clarke, B., Doabler, C. T., Turtura, J., Smolkowski, K., Kosty, D., 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/10.1086/710041

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (ASPENS)* [Measurement instrument]. Cambium Learning Group, Sopris Learning.

Codding, R. S., & Lane, K. L. (2015). A Spotlight on Treatment Intensity: An Important and Often Overlooked Component of Intervention Inquiry [journal article]. *Journal of Behavioral Education, 24*, 1–10. https://doi.org/10.1007/s10864-014-9210-z

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

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2014). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education*, Advanced online publication. https://doi.org/10.1177/0741932514557271

Council of Chief State School Officers. (2010). *Common core state standards initiative: Designing common state assessment systems*. NGA. http://www.edweek.org/media/ngaccsso_assessmentdesignpaper.pdf

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education, 37*, 244–253. https://doi.org/10.1177/0741932516648463

DeFouw, E. R., Codding, R. S., Collier-Meek, M. A., & Gould, K. M. (2018). Examining Dimensions of Treatment Intensity and Treatment Fidelity in Mathematics Intervention Research for Students at Risk. *Remedial and Special Education, 40*(5), 298–312. https://doi.org/10.1177/0741932518774801

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the Association between Explicit Mathematics Instruction and Student Mathematics Achievement. *The Elementary School Journal, 115*(3), 303–333. https://doi.org/10.1086/679969

Doabler, C. T., & Clarke, B. (2012). *Quality of explicit mathematics instruction* [Unpublished observation instrument]. Eugene, OR:

Doabler, C. T., Clarke, B., Kosty, D., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2019). 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., Turtura, J. E., Sutherland, M., Maddox, S. A., & Smolkowski, K. (2021). Using Direct Observation to Document "Practice-Based

Evidence" of Evidence-Based Mathematics Instruction. *Journal of Learning Disabilities, 54*(1), 20–35. https://doi.org/10.1177/0022219420911375

Doabler, C. T., Clarke, B., Kosty, D. B., Baker, S. K., Smolkowski, K., & Fien, H. (2016). The effects of a core kindergarten mathematics program on the mathematics achievement of Spanish-Speaking English learners. *School Psychology Review, 45*(3), 343–361. https://doi.org/10.17105/SPR45-3.343-361

Doabler, C. T., Nelson, N. J., Stoolmiller, M. L., & Baker, S. K. (2015-2017). *Exploring Alterable Variables in Tier 1 and Tier 2 Instruction: A Collaboration Across Interdisciplinary Fields of Observational Research (Project CIFOR)* (Project No R305A150037, awarded $699,706). US Department of Education, Institute of Education Sciences, National Center of Education Research, Effective Teachers and Effective Teaching, CFDA No. 84.305A

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

Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics Content Coverage and Student Learning in Kindergarten. *Educational Researcher, 45*, 293–300. https://doi.org/10.3102/0013189x16656841

Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review, 20*, 695–723. https://doi.org/10.1177/0193841X9602000603

Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). *Teaching math to young children: A practice guide* (No. NCEE 2014-4005). National Center for Education and Regional Assistance (NCEE), Institute of Education Sciences,

US Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/early_math_pg_111313.pdf

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

Fuchs, L. S., Fuchs, D., & Malone, A. S. (2018). The Taxonomy of Intervention Intensity. *TEACHING Exceptional Children, 50*(4), 194–202. https://doi.org/10.1177/0040059918758166

Fuchs, L. S., Newman-Gonchar, R., Schumacher Robin, F., Dougherty, B., Bucka, N., Karp, K., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R. M., Jayanthi, M., Keating, B., & Morgan, S. T. (2021). *Assisting Students Struggling with Mathematics: Intervention in the Elementary Grades* (WWC No. 2021006). National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/WWC2021006-Math-PG.pdf

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-Intervention: A decade later. *Journal of Learning Disabilities, 45*, 195–203. https://doi.org/10.1177/0022219412442150

Gersten, R. M., Beckmann, S., Clarke, B., Foegen, A., March, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (Practice Guide Report No. NCEE 2009-4060). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_math_pg_042109.pdf

Gersten, R. M., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33*, 18–28. https://doi.org/10.1177/002246699903300102

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

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78*, 119–128.

Griffin, S., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 24–49). MIT Press.

Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli?: A Monte Carlo Comparison of the Performance of the Linear Mixed-Model and the Logistic Mixed-Model Analyses in Simulated Community Trials With a Dichotomous Outcome Variable at the Individual Level. *Evaluation Review, 20*, 338–352. https://doi.org/10.1177/0193841x9602000306

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128. https://doi.org/10.3102/10769986006002107

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Lawrence Erlbaum Associates.

Jamieson, J. (1999). Dealing with baseline differences: two principles and two dilemmas. *International Journal of Psychophysiology, 31*, 155–161. https://doi.org/10.1016/S0167-8760(98)00048-8

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The Importance of Number Sense to Mathematics Achievement in First and Third Grades. *Learning and individual differences, 20*, 82–88. https://doi.org/10.1016/j.lindif.2009.07.004

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Levin, H. M., & Belfield, C. (2015). Guiding the Development and Use of Cost-Effectiveness Analysis in Education. *Journal of Research on Educational Effectiveness, 8*(3), 400–418. https://doi.org/10.1080/19345747.2014.915604

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2 ed.). John Wiley & Sons.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304–316. https://doi.org/10.3102/0013189x14545513

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

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Who Is At Risk for Persistent Mathematics Difficulties in the United States? *Journal of Learning Disabilities, 49*, 305–319. https://doi.org/10.1177/0022219414553849

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306–321. https://doi.org/10.1177/0022219408331037

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. US Department of Education.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Mathematics Learning Study Committee.

National Science Foundation, The Institute of Education Sciences, & U.S. Department of Education. (2018). *Companion guidelines on replication & reproducibility in education research: A supplement to the common guidelines for education research and development*. National Science Foundation, The Institute of Education Sciences, U.S. Department of Education,. https://ies.ed.gov/pdf/CompanionGuidelinesReplicationReproducibility.pdf

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

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2*, 152–162. https://doi.org/10.1191/1740774505cn076oa

Rolfhus, E., Gersten, R., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2012). *An evaluation of "Number Rockets": A tier-2 intervention for grade 1 students at risk for difficulties in mathematics* (Final Report No. NCEE 2012-4007). National Center for

Education Evaluation and Regional Assistance. http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED529429

SAS Institute. (2016). *Base SAS® 9.4 procedures guide: Statistical procedures* (5th ed.). SAS Institute, Inc.

Saxe, G. B., Guberman, S. R., Gearhart, M., Gelman, R., Massey, C. M., & Rogoff, B. (1987). Social Processes in Early Number Development. *Monographs of the Society for Research in Child Development, 52*(2), i–162. https://doi.org/10.2307/1166071

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27*, 316–328. https://doi.org/10.1016/j.ecresq.2011.09.004

Sood, S., & Jitendra, A. K. (2013). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of Learning Disabilities, 46*, 328–346. https://doi.org/10.1177/0022219411422380

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*, 99–120. https://doi.org/10.1016/j.ecresq.2004.01.002

Van Belle, G. (2008). *Statistical rules of thumb* (2 ed.). Jon Wiley & sons.

Vaughn, S., Cirino, P. T., Wanzek, J., Wexler, J., Fletcher, J. M., Denton, C. D., Barth, A., Romain, M., & Francis, D. J. (2010). Response to Intervention for middle school students with reading difficulties: Effects of a primary and secondary intervention. *School Psychology Review, 39*, 3–21.

Vaughn, S., Linan-Thompson, S., Kouzekanani, K., Pedrotty Bryant, D., Dickson, S. V., & Blozis, S. A. (2003). Reading instruction grouping for students with reading difficulties. *Remedial and Special Education, 24*, 301–315. https://doi.org/10.1177/07419325030240050501

Vaughn, S., & Swanson, E. A. (2015). Special Education Research Advances Knowledge in Education. *Exceptional Children, 82*, 11–24. https://doi.org/10.1177/0014402915598781

What Works Clearinghouse. (2017). *Procedures Handbook version 4.0*. Institute of Education Science, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an Adaptive Game Intervention on Accessing Number Sense in Low-Socioeconomic-Status Kindergarten Children [https://doi.org/10.1111/j.1751-228X.2009.01075.x]. *Mind, Brain, and Education, 3*(4), 224–234. https://doi.org/https://doi.org/10.1111/j.1751-228X.2009.01075.x

Witzel, B., & Clarke, B. (2015). Focus on Inclusive Education: Benefits of Using a Multi-tiered System of Supports to Improve Inclusive Practices. *Childhood Education, 91*, 215–219. https://doi.org/10.1080/00094056.2015.1047315