# Examining the Efficacy of a Kindergarten Mathematics Intervention by Group Size and Initial Skill: Implications for Practice and Policy


## Authors

Ben Clarke, Christian T. Doabler, Jessica Turtura, Keith Smolkowski, Derek B. Kosty, Marah Sutherland, Evangeline Kurtz-Nelson, Hank Fien, Scott K. Baker



## Publication History

- Published online August 06, 2020

## Full Reference

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. doi: 10.1086/710041

## Founding Source

This research was supported by the ROOTS Project, grant number **R324A120304**, and funded by the US Department of Education, Institute of Education Sciences. The opinions expressed are those of the authors and do not represent the views of the Institute or the US Department of Education.

Abstract

This study examined whether the efficacy of a 50 lesson mathematics intervention program focused on whole number concepts for at-risk kindergarten students, ROOTS, differed by group size and whether initial skill moderated intervention effects by group size. The study utilized a randomized block design with at-risk students (n = 1,251) within classrooms (n = 138) randomly assigned to one of two treatment conditions (a small group of two or five students) or control condition. Proximal and distal measures were collected in the fall (pretest), spring (posttest) and winter of first grade (follow-up). Results indicated that students who received ROOTS performed better at posttest than control students (Hedges' g from 0.09 to 0.81), that impact did not vary by group size, and that initial skill moderated the impact of ROOTS compared to control student outcomes but not likely differences in group size.

Keywords: Mathematics, Learning Disabilities, Response to Intervention

# Examining the Efficacy of a Kindergarten Mathematics Intervention by Group Size and Initial Skill: Implications for Practice and Policy

Over the past two decades, the importance of mathematics learning has garnered increased interest at the national level (National Research Council, 2001) resulting in federal initiatives aimed at improving the mathematics instruction provided to the nation's students (Common Core State Standards Initiative, 2010; National Mathematics Advisory Panel, 2008). Despite continued focus and efforts to improve mathematics learning, National Assessment of Educational Performance (NAEP) indicate stagnating levels of achievement with only 40 percent of students being classified as at or above proficient. Of even greater concern are the significantly lower levels of performance for students from low SES backgrounds, minorities, and English Language Learners (ELL) resulting in substantive and persistent achievement gaps (NAEP, 2017) at a time when state and federal policy and initiatives are aimed at ensuring all learners have access to STEM and STEM related opportunities (National Conference of State Legislatures, 2019; The White House Office of Science and Technology Policy, 2018).

Difficulties in mathematics are relatively stable with deficits as early as kindergarten entry fostering long term difficulty across elementary school and impacting access to and success with higher order mathematics including rational number systems and prealgebra (Duncan et al., 2007; National Mathematics Advisory Panel, 2008). Yet there is some evidence that is trajectories are altered during this period there can be a significant impact on long term outcomes (Morgan, Farkas, & Wu, 2009) with one analyses indicating that growth in mathematics in kindergarten and first grade predicted high mathematics outcomes more strongly than initial skill (Watts, Duncan, Siegler, & Davis-Kean, 2014). Given these findings, several researchers have developed, evaluated, and found positive impacts for intervention curricula targeting early number sense and foundational whole number concepts with the goal of preventing later mathematics difficulty (e.g. Clarke et al., 2014; Dyson, Jordan, & Glutting, 2013; L. S. Fuchs et al., 2005). The interventions programs developed and evaluated as part of this emerging research base are designed to be delivered in small groups of at-risk students within a Response to

Intervention (RTI) or Multi-Tier Systems of Support (MTSS) service delivery framework.

First conceptualized as a mechanism to identify students with specific learning disabilities (IDEA, 2004) , RTI has morphed into a widely adopted instructional service delivery system or MTSS designed to address the learning needs of all students not just those with learning disabilities (Balu et al., 2015) . While several principles form the foundation of MTSS practice including early identification and prevention (Clarke, Doabler, & Nelson, 2014), fundamentally the system rests upon the idea of providing more intensive instruction as students progress through a series of instructional tiers (National Center on Response to Intervention, 2010). Although variations exist, most MTSS models consist of three tiers with core instruction in general education classroom serving as the first tier, small group supplemental instruction serving as the second tier, and some form of more intensive instruction forming the basis of the third tier (Gersten et al., 2009; NASDSE, 2005).

The work done in mathematics intervention research to date largely fit in fit within standard MTSS service delivery systems as Tier 2 small interventions. An overview of best practices in mathematics noted the dearth of Tier 3 research and variations to Tier 2 programs (Gersten et al., 2009). Not surprisingly subsequent calls have advocated for building upon the existing research base by examining variations in how programs are delivered and gaining greater insight into what works, for whom, and under what conditions (Miller, Vaughn, & Freund, 2014). Because MTSS systems rests on a series of cascading tiers in which theoretically the intensity of services is increased as students move from tier to tier with each subsequent tier serving students with greater risk and thus requiring greater intensity (Codding & Lane, 2015), it is logical to extend mathematics intervention research by examining variables related to the intensity of instruction and the degree of student need.

While intensifying instruction can take many forms, one popular mechanism to increase intensity is to decrease group size in order to provide a more individualized learning experience. A limited number of studies have investigated the impact of manipulating group size on student reading outcomes. Results from those studies have been mixed. A meta-analysis on the impact of

group size by Wanzek and Vaughn (2007) found larger effect sizes for smaller groups. Vaughn and colleagues (2003) conducted a study in which instructional content was kept constant across small groups of varying sizes. Results indicated stronger effects for two small groups (1:1 and 1:3 teacher to student ratio) when contrasted with a small group with a 1:10 teacher to student ratio. However, the two smaller groups did not differ. Similar work in mathematics has not been conducted. A review of the literature found zero studies investigating a systematic manipulation of group size.

A second key consideration when examining the conditions under which an intervention is effective is the role of initial skill (D. Fuchs & Fuchs, 2019). In part the importance of initial skill is best understood within the findings that there are a subset of students who exhibit persistent MLD (Geary, 2011) and fail to respond to generally efficacious interventions (L. S. Fuchs, Fuchs, & Compton, 2013). One potential proposed approach to address non-response is to identify variables, such as initial skill, and screen students likely to exhibit non-response to Tier 2 programs directly into a more intensive Tier 3 instructional setting (Al Otaiba et al., 2014; D. Fuchs & Fuchs, 2017). Researchers have begun to explore the relationship between initial skill and intervention response in mathematics with mixed findings to date. For example, Fuchs and colleagues in a series of studies have found intervention response to not vary by initial skill for a first grade mathematics intervention targeting whole number (L. S. Fuchs, Fuchs, & Gilbert, 2019) and a fourth grade intervention targeting fractions (L. S. Fuchs, Sterba, Fuchs, & Malone, 2016). Contrasting findings included response to an early numeracy intervention by initial early numeracy skill (Toll & Van Luit, 2013) and for a fraction word problem intervention by reasoning ability (L. S. Fuchs, Malone, et al., 2016) with greater response found for students with greater initial skill. Given the range of findings and the types of mathematics content covered within the interventions, drawing conclusions from this emerging research area is tenuous. In addition, across both reading and mathematics intervention research, we found no studies exploring the relationship between group size, initial skill, and response variation. Exploring this complex relationship is critical in light of how service delivery models are

constructed. For example, a student with a moderate degree of risk may be theorized to gain the same benefit from an intervention regardless of the group size in which it is delivered whereas a student with severe risk may only benefit from a more intensive intervention experience. As such additional research is called for (Fuchs & Fuchs, 2019) that examines the complex interaction between moderation variables, like initial skill, and instructional variations, like group size.

Our four-year efficacy trial funded by the Institute of Education Sciences (Clarke, Doabler, Fien, Baker, & Smolkowski, 2012)was designed to address this gap and to 1) examine the general efficacy of a kindergarten mathematics intervention, ROOTS, on at-risk student mathematics outcomes 2) investigate questions related to intervention impact by group size 3) investigate questions related to intervention impact by initial skill and lastly 4) explore the relationship between group size, initial skill, and intervention impact. The study included two treatment conditions in which students received ROOTS in either a 2 student small group or a 5 student small group. Previous examinations of the ROOTS intervention program revealed positive impacts on student achievement (Clarke, Doabler, Smolkowski, Baker, et al., 2016; Doabler et al., 2016), similar impacts by group size (Clarke et al., 2017), and that initial mathematics skill did not moderate student outcomes (Clarke et al., 2019 ). However, the studies were underpowered for examining secondary exploratory research questions related to group size, initial skill, and the role of initial skill in moderating outcomes by group size. Thus, the work presented here is the first to utilize the full ROOTS data set to investigate a comprehensive range of research questions directly relevant for expanding the research literature on effective mathematics intervention within MTSS service delivery systems.

Three research questions were examined as part of this work:

**Research Questions**

1. What was the overall impact of the treatment, ROOTS intervention, compared to control, business as usual?
2. Was there a differential impact on student outcomes between the two treatment conditions (i.e., ROOTS large group versus ROOTS small group)?

3. Did students benefit differentially from the ROOTS intervention by initial skills, as measured by pretest variables? And does that relationship vary by group size?

## Method

This study presents the results of analyses on data collected during the federally funded ROOTS Efficacy Project (Clarke et al., 2012 ). Implementation of the ROOTS intervention occurred across three school years (2012-2015) at two different research sites: Oregon and Massachusetts. A partially nested randomized controlled trial was employed (Baldwin et al., 2011), randomly assigning kindergarten students within classrooms to one of three conditions: (2:1 ROOTS group, 5:1 ROOTS group, and a no-treatment control condition).

### Participants

**Districts and Schools.** Twenty-three schools from four Oregon school districts and two Massachusetts school districts participated. The two Massachusetts districts were located in close proximity to Boston. Three of the Oregon districts were located in rural and suburban areas of western Oregon, while one district was located near Portland. Across the six districts, student enrollment ranged from 2,736 to 39,002. A total of 23 schools participated. Within these schools, 0%-12% of students were American Indian or Native Alaskan, 0%-16% were Asian, 0%-16% were Black, 0%-83% were Hispanic, 0%-2% were Native Hawaiian or Pacific Islander, 9%-92% were White, and 0%-15% were more than one race. Within these same schools, 8%-25% of students received special education services, 0%-69% were English language learners, and 17%-87% were eligible for free or reduced lunch.

**Classrooms and Teachers.** A total of 138 kindergarten classrooms participated in the study, with the majority (57%) providing half-day kindergarten programs. The 138 classrooms were taught by 75 certified kindergarten teachers, of which 48 teachers participated for two consecutive years in the ROOTS Efficacy Project. Among the 75 participating teachers, 70 provided demographic information. All teachers identified as female, 88.6% as White, and 4.3% as Asian American/Pacific Islander. The remaining 7.1% of teachers identified as another

race/ethnicity or declined to respond. Teachers had an average of 15.2 years of teaching experience (*SD* = 9.1). The majority of teachers (78.6%) had a master's degree in education, and 58.6% had taken algebra at the college or graduate level.

**Students.** All students with parental consent from the 138 classrooms were screened in the late fall of their kindergarten year. The screening process included the Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke, Gersten, Dimino, & Rolfhus, 2011) and the Number Sense Brief (NSB; Jordan et al., 2010), which are standardized measures of early mathematics proficiency. Students were eligible for the ROOTS intervention and thus considered at risk for MD if they received an NSB score of 20 or less and an ASPENS' composite score in the strategic or intensive ranges. After being determined eligible for the ROOTS intervention, students' NSB and ASPENS scores were separately converted into standard scores and then combined to form an overall composite score for each at risk student. All data management were conducted by project's independent evaluator. Composite scores within each classroom were then rank ordered, and the 10 ROOTS-eligible students with the lowest composite scores were randomly assigned to one of three conditions: (a) 2:1 ROOTS group, (b) 5:1 ROOTS group, or (c) a no-treatment control condition. Control students received business as usual instruction and continued to receive core instruction. The small group size of five students was utilized based on the common group sizes used in intervention settings at Tier 2. The group size of two was selected instead of a group size of one due to potential attrition at the group level (i.e. if one student leaves from a "group" of one) and representing the lower bound of typical small group size. The five and two student groups allowed us to contrast typical small group instruction with a more intensive instructional format. Of the 138 classrooms included in this study, 105 had at least 10 students who met the ROOTS eligibility criteria. When classrooms did not have 10 students who met the eligibility criteria, the project's independent evaluator applied a cross-class grouping procedure, which consisted of combining classrooms for to create a virtual ROOTS classroom. For example, in the project's first year, at-risk students from two classrooms were combined and thus provided one 2:1 ROOTS group, a 5:1 ROOTS group, and a control group.

After these procedures were applied, a total of 255 ROOTS groups were formed: 129 of the 2:1 groups, 126 of the 5:1 groups. From the 138 classrooms, a total of 3,130 kindergarten students were screened for MD and, in turn, ROOTS eligibility. Of 3,130 students, 1,251 met eligibility criteria and were randomly assigned to the 2:1 group condition ($n$ = 258), the 5:1 group condition ($n$ = 622), or the no-treatment control condition ($n$ = 371).

**ROOTS Interventionists**

The ROOTS intervention was delivered by district employees and interventionists hired specifically for the efficacy trial. The majority of interventionists (93.5%) identified as female (93.5%) and White (76.1%), with 12.0% identifying as Hispanic. The remaining 11.9% identified as another race/ethnicity or declined to respond. Almost all interventionists (92.3%) had previous experience providing small group instruction, and 60.5% had a bachelor's degree or higher. About half of interventionists (56.5%) had taken an algebra course at the college or graduate level. On average, interventionists had 10.4 years of teaching experience ($SD$ = 8.6) and 22.0% had a current teaching license or certification.

The ROOTS interventionists participated in two five-hour professional development workshops that were delivered by project staff with a background in mathematics education, including one of the curriculum developers. The initial workshop focused on mathematics content covered through Lesson 25, effective instructional practices (e.g., eliciting group and individual responses, providing academic feedback), and strategies for small-group management (e.g., instructional pacing, setting group expectations, etc.). The second workshop covered content from Lesson 26 to 50 and reviewed instructional and management strategies. During each workshop, project staff modeled lesson delivery and provided time for interventionists to practice lessons and receive feedback on their use of instructional practices. All interventionists also received between two and four coaching visits from ROOTS coaches during intervention implementation to boost implementation fidelity and enhance instructional quality. The coaching visits consisted of direct observations of lesson delivery followed by feedback on instructional quality (e.g., use of effective instructional practices and group management strategies) and

fidelity of intervention implementation (e.g., correctly using mathematical models, following the teacher scripting).

**ROOTS.** ROOTS is a 50-lesson, Tier 2 mathematics program designed to build students' proficiency in whole number concepts and skills. The ROOTS intervention was delivered in 20-minute small group sessions (2:1 or 5:1) 5 days per week for approximately 10 weeks. Instruction for all students began in the late fall and ended in the spring, and this start date was selected to provide students with the opportunity to respond to initial core mathematics instruction and to therefore minimize the identification of typically-achieving students. ROOTS was designed to supplement core mathematics instruction and thus was delivered at times that did not conflict with students' core instruction in mathematics.

ROOTS content emphasizes the Counting and Cardinality, Operations and Algebraic Thinking, and Number and Operations in Base Ten strands of the CCSS-M (2010), following recommendations from expert panels to focus on critical whole number concepts and skills (Gersten et al., 2009). The scope and sequence of ROOTS is organized in tracks, with skills built and revisited across multiple lessons. Approximately four to six brief activities are included in each lesson, providing students with practice on multiple skills each day and frequent cumulative review. For example, in Lesson 30, students participate in a daily warm-up including identifying numbers and counting using a "Nifty Fifty" number chart, rational count to 11, work with base ten rods and cubes to build number models, and complete a daily "Math Practice" activity that includes several review problems. Across the curriculum, students are introduced to number names and the count sequence to 100, but an intense focus is placed on numbers 0-20 given the challenges that at-risk students frequently encounter with understanding teen numbers (National Research Council, 2001).

ROOTS employs mathematical models to build students' conceptual understanding of abstract mathematical concepts, with lessons following the concrete-representational-abstract sequence (Agrawal & Morin, 2016). Initially, students work with manipulative models such as finger models, teddy bear counters, and base ten blocks to represent numbers. As students

progress, these materials are faded out and visual representations such as ten frames, tally marks, and number lines are used. Last, students work solely with numerals having built understanding of their meaning, relations among numerals, and place value.

The ROOTS instructional approach is drawn from principles of explicit and systematic mathematics instruction (Coyne, Kame'enui, & Carnine, 2011; Gersten et al., 2009 ) including explicit teacher modeling of new concepts, guided and deliberate teacher-led practice, and corrective or confirmatory academic feedback. Lessons are fully scripted, enabling instructors to use precise and consistent mathematical language within and across lessons, and to ensure that student-teacher interactions, such as student response opportunities, are high-quality. Frequent opportunities for students to verbalize their mathematical thinking and reasoning are also embedded throughout the program's lessons. For example, when building models of teen numbers using base ten blocks and cubes, students are asked how many ten sticks and cubes they would use to represent a given numeral and to explain their reasoning.

**Implementation Fidelity.** Fidelity of ROOTS implementation was measured via direct observations by trained research staff. Each ROOTS group was 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 provided teacher scripting, and (c) used the prescribed mathematics models for that lesson. Observers also recorded whether the interventionist taught the number of activities prescribed in the lesson. Interventionists were observed to meet instructional objectives ($M$ = 3.49, $SD$ = 0.69), follow scripting ($M$ = 3.31, $SD$ = 0.75), and use prescribed models ($M$ = 3.61, $SD$ = 0.64). Interventionists also taught the majority of prescribed activities ($M$ = 4.14 out of 5 activities per lesson, $SD$ = 0.77). Interclass correlation coefficients (ICCs) for these fidelity ratings across observers were as follows: .82 for number of activities taught, .70 for meeting instructional objectives, .75 for following teacher scripting, and .70 for using prescribed mathematics models. Per guidelines proposed by Landis and Koch (1977) , these ICCs indicate substantial agreement across observers. There were no differences in ICCs by group size.

**Outcome Measures**

All treatment and control students were administered five measures of whole number understanding at pretest and posttest. Students were also administered one distal measure of mathematics achievement at posttest only. Trained research staff administered all student measures. Inter-scorer reliability criteria were met for all assessments (i.e., > 95% agreement).

**ROOTS Assessment of Early Numeracy Skills** (RAENS; Doabler, Clarke, & Fien, 2012) is a researcher-developed, individually administered measure that consists of 32 items. Items assess aspects of counting and cardinality, number operations, and the base-10 system. In an untimed setting, students are asked to count and compare groups of objects, write, order, and compare numbers, label visual models (e.g. ten-frames), and write and solve single digit addition expressions and equations. RAENS' predictive validity ranges from .68 to .83 for the TEMA-3 and the NSB. Inter-rater scoring agreement is reported at 100% (Clarke, Doabler, Smolkowski, Kurtz Nelson, et al., 2016).

**Oral Counting – Early Numeracy Curriculum-Based Measurement** (Clarke & Shinn, 2004). This curriculum-based measure has students orally count in English for one minute and the discontinue rule applies after the first counting error. The highest correct number counted represents a student's score. Test-retest reliability and alternate-form reliability are reported at above .80, concurrent validity is reported as ranging from .49 to .70, and predictive validity with standardized measures of mathematics ranging from .46 to .72.

**Assessing Student Proficiency in Early Number Sense** (ASPENS; Gersten et al., 2012 ) is a set of three curriculum-based measures validated for screening and progress monitoring in kindergarten mathematics. Each 1-minute fluency-based measure assesses an important aspect of early numeracy proficiency, including number identification, magnitude comparison, and missing number. Test-retest reliabilities of kindergarten ASPENS measures are in the moderate to high range (.74 to .85). Predictive validity of fall scores on the kindergarten ASPENS measures with spring scores on the TerraNova 3 is reported as ranging from .45 to .52.

**Number Sense Brief Screener** (NSB; Jordan, Glutting, & Ramineni, 2008) is an

individually administered measure with 33 items that assess counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems and number combinations. NSB has a coefficient alpha of .84.

**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 whole number understanding for children ranging in age from 3 to 8 years 11 months. Alternate-form and test-retest reliabilities of the TEMA-3 are .97 and .93, respectively. The TEMA-3 has concurrent validity with other mathematics measures ranging from .54 to .91.

**The Stanford Achievement Test-Tenth Edition** (SAT-10; Harcourt Educational Measurement, 2002) . The SAT-10 measure is a group administered, standardized, norm referenced test with two mathematics subtests, Problem Solving and Procedures. The kindergarten version of the SAT-10 is the Stanford Early Achievement Test (SESAT). The SAT-10 is a standardized achievement test with adequate and well-reported validity ($r = .67$) and reliability ($r = .93$). Student total and subtest scores are typically reported; however, detailed student reports are also available which note whether the student is below, at, or above average for specific skill clusters.

**Statistical Analysis**

The study design called for the randomization of individual students to receive ROOTS, nested within ROOTS groups, or a nonnested control condition. We conducted multiple sets of analyses to address our four research questions. First, we examined overall effects of the ROOTS intervention on math achievement using a mixed model Time × Condition analysis (Murray, 1998) designed to account for students either nested within small groups for intervention or nonnested control students (Baldwin, Bauer, Stice, & Rohde, 2011; Bauer, Sterba, & Hallfors, 2008). The ROOTS groups, but not the unclustered controls, required a group-level variance estimate.

The analytic model accounted for the potential heterogeneity of residual variances across

conditions (Roberts & Roberts, 2005). Because the residual variances may have differed between clustered intervention students and unclustered control students, we tested the assumption of homoscedasticity of residuals and reported results of the most appropriate model for each outcome measure. We tested whether the homoscedastic and heteroscedastic models could be assumed equivalent with a likelihood ratio test and reported the simpler model if we were able to accept the equivalence of the two models. Because this tests the noninferiority of the simpler model when compared to the more complex model, we reversed the null and alternative hypotheses and, hence, the Type I and Type II error rates, $\alpha$ and $\beta$, which is common among equivalence or noninferiority trials (e.g. Dasgupta, Lawson, & Wilson, 2010; Piaggio et al., 2006). For this reason, and the limited statistical power to detect differences in variance structures (Kromrey & Dickinson, 1996), we set $\alpha = .20$ as our Type I error rate and reported the more complex model unless we were relatively certain the two were equivalent.

The partially nested Time × Condition analysis tested for differences between conditions on gains in outcomes from the fall (T1) to spring (T2) of kindergarten and is described in greater detail by Clarke, Doabler, Smolkowski, Kurtz Nelson, et al. (2016) and Doabler et al. (2016). The statistical model included time, coded 0 at T1 and 1 at T2, condition, coded 0 for control and 1 for ROOTS, and the interaction between the two. 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 the SESAT and SAT10 available only at posttest we used the analysis of covariance approach described by Bauer et al. (2008) and Baldwin et al. (2011). We used Satterthwaite approximation to determine the degrees of freedom in tests of effect estimates.

To test for ROOTS group-size differences, we compared group sizes coded 0 for large groups and 1 for small groups among students in intervention groups. We conducted a nested Time × Group Size analysis (Murray, 1998) to account for the dependence of students clustered within small groups. Because all students were nested within small groups, these models did not requires the partially nested analysis described above. The analysis tested for differences

between conditions on gains in outcomes from the fall to spring of kindergarten. We used a mixed-model analysis of covariance to test the SESAT, measured only at posttest, and the SAT10, measured at follow-up. We also tested pretest TEMA-3 scores as a moderator to determine if group-size differences depended on pretest math skill. We tested an extended set of mixed-models to account for students clustered within classrooms, which produced similar results.

We also examined whether initial math achievement based on TEMA-3 scores or group size predicted differential response to the ROOTS intervention compared to control as well as differential response by group size. We expanded the statistical models above to include the predictor of differential response and its interaction with condition, time, and the Time × Condition term, resulting in a three-way interaction, all corresponding two-way interactions, and individual (conditional) effects. The three-way Time × Condition × Pretest interaction provided an estimate of whether condition effects varied by initial math achievement. Condition represented ROOTS versus control for moderation of impact and group size for moderation of group-size differences.

Finally, we examined variability of the condition effect by classroom. We tested an additional set of mixed-models that extended those discussed above to account for students clustered within classrooms. Results were similar to those without the classroom level, and condition effects did not vary by classroom. We therefore omitted these results.

**Model Estimation**. We fit the aforementioned statistical models to our data using SAS PROC MIXED version 14.2 (SAS Institute, 2016) and restricted maximum likelihood estimation. Maximum likelihood estimation with all available data produces potentially unbiased results even in the face of substantial missing data, provided the missing data were missing at random (Schafer & Graham, 2002), although nonrandom missingness "is often not sufficient to affect the internal validity of an experimental study to any practical extent" (Graham, 2009 , p. 568). In the present study, we did not believe that 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 families 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. Multilevel regression methods are also quite robust to violations of normality (e.g., Hannan & Murray, 1996).

**Effect Sizes and Multiple Tests**. To interpret results, we computed effect sizes, the Hedges' *g* for continuous measures and Cox's *d* for dichotomous measures, for pretest differences and model results using What Works Clearinghouse procedures (WWC, 2017). We also corrected for multiple tests with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) and reported the original *p*-values as well as the Benjamini-Hochberg adjusted *p*-values for each outcome. We adjusted *p*-values separately within the set of analyses for each research question.

## Results

The efficacy of the ROOTS intervention—the comparison between ROOTS and control groups—has been tested for subgroups of the full sample in Clarke et al. (2017), Clarke, Doabler, Smolkowski, Kurtz Nelson, et al. (2016), Doabler et al. (2018), and Clarke, Doabler, Smolkowski, Baker, et al. (2016). Clarke et al. (2017); Clarke et al. (2019) also compared group sizes with subgroups. Herein we summarize the demographic information, math measures, and present results for the comparison between intervention conditions for the full sample. We then present results for the comparison between small and large groups, condition differences moderated by initial skill, and group-size differences moderated by initial skill.

### Descriptive Results and Baseline Equivalence

Demographic characteristics were reported in Table 1. ROOTS and control groups did not meaningfully differ on proportions of students who were male, White, or Hispanic ($d < 0.05$)

or on their mean age ($g = 0.02$). We did not test other race categories because all groups represented less than 5% of the total sample. Approximately 8% of the sample was designated as special education, with 8% in the intervention sample and 9% in the control sample ($d = 0.11$). Cox's *d*, however, is very sensitive to differences between groups when the base rate is less than .10 or greater than .90. For example, had we observed one special education student among the 880 intervention students and one special education student out of 371 control students, the combination would have produced a Cox's *d* of −0.52. This would indicate a substantial pretest difference, yet such small rates would not meaningfully affect the results.

Table 2 reports the descriptive statistics by assessment time and intervention condition, ROOTS and control, for each outcome measure except the SAT10, collected only at follow-up (ROOTS $M = 496.8$, $SD = 28.1$, $N = 463$; Control $M = 495.2$, $SD = 25.3$, $N = 198$). The sample size for the SAT10 was smaller because it was not collected for the fourth cohort of students.

ROOTS group sizes did not notably differ on proportions of students who were male, White, or Hispanic ($d < 0.05$). We did not test other race categories because all groups represented less than 5% of the total sample. Students in small groups were slightly younger ($M = 5.25$ years, $SD = 0.44$) than large groups ($M = 5.28$, $SD = 0.45$; $g = 0.07$). Approximately 8% of the sample was designated as special education, with 9% in small groups and 7% in large groups ($d = 0.07$).

Table 3 reports the descriptive statistics by assessment time and group size for each outcome measure except the SAT10, collected at follow-up (small group $M = 494.06$, $SD = 28.66$, $N = 138$; large group $M = 497.89$, $SD = 27.87$, $N = 325$). The SAT10 was not collected for the fourth cohort of students.

**Attrition**

**Condition differences**. The overall rate of missingness was less than 6.6% for the measures available at pretest, and the difference in rates of missingness between conditions was below 4.0% for the 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 for *differential attrition effects* to identify potential threats to internal validity. To do so, we conducted a mixed-model analysis of variance designed to test whether attrition differentially affected condition differences for outcome variables. Specifically, the analyses tested the association between pretest measures and (a) study condition (ROOTS versus control or small versus large group sizes), (b) attrition status, and (c) the interaction between the two (Graham & Donaldson, 1993). At pretest, the intervention groups had not yet been formed, so this analysis did not account for the partially nested structure found at posttest. We found no interactions between attrition and condition that predicted baseline outcomes that were large enough to suggest that attrition threatened internal validity ($p > .17$).

**Group-size differences**. The overall rate of missingness at posttest was 7.7% for the measures available at pretest, and the difference in rates of missingness between group sizes was below 3.5% for pretest measures. From an analysis of differential attrition effects, we found little evidence that attrition threatened internal validity ($p > .10$).

**Differences between ROOTS and Control Students**

Tables 4 and 5 presents the results of the partially nested analyses that compared ROOTS students in small groups to unclustered control students at posttest or follow-up. The bottom two rows of the table show the likelihood ratio test results that compared homoscedastic residuals with heteroscedastic residuals, and the tables report a different number of variances depending on the results. The data fit the homoscedastic model that assumed equivalent residual and pre-post covariance estimates between conditions for oral counting, TEMA-3, and SESAT. The data fit the heteroscedastic model for the RAENS, ASPENS, NSB, and SAT10. Although the variance structures differed between these models, the estimates of condition effects and related statistics were similar for both models.

The *g* and *p* values in Tables 4 and 5 represent the test of condition differences, and we also provided Benjamini-Hochberg adjusted *p* values. Students in the ROOTS condition improved from fall to spring at a greater rate than students in the control condition on the RAENS ($g = 0.81$, 95% CI [0.69, 0.93]), ASPENS (0.49, [0.38, 0.60]), NSB (0.18, [0.06, 0.30]),

TEMA-3 (0.23, [0.14, 0.32]), and SESAT (0.23, [0.12, 0.33]). The data did not support differences between conditions on oral counting at posttest (0.09, [−0.03, 0.21]) or the SAT10 at follow-up (−0.02, [−0.17, 0.12]).

**ROOTS Group Size Differences**

To examine the impact of group size, we compared group sizes among only students who received ROOTS. We found little evidence for differences between large and small groups for any variables except for the SESAT. The analyses produced a difference between groups sizes for the SESAT of 4.89 ($SE$ = 2.39, $g$ = 0.14, 95% CI [0.01, 0.27], $p$ = .0418) but with a Benjamini-Hochberg adjusted $p$-value of $p$ = .2926. See Tables 6 and 7 for all results.

**Differential Response to ROOTS versus Control based on Pretest Math Skill**

Tables 8 and 9 present tests of differential response to ROOTS as a function of pretest TEMA-3 scores. The tables use the same format as those for the main effects but with additional fixed effects. The TEMA-3 moderated condition effects for the RAENS, NSB, and TEMA-3. Students in the ROOTS condition outperformed those in the control condition on the RAENS across nearly all pretest TEMA-3 scores (those below the 98$^{th}$ sample percentile or a TEMA-3 score of 36). Conditions differed for 98% of the sample, and students with lower TEMA-3 scores at pretest appeared to benefit most from ROOTS on the RAENS. Students in the ROOTS condition outperformed those in the control condition—the confidence bounds excluded zero—on the NSB with pretest TEMA-3 scores below the 62$^{nd}$ sample percentile (62%, score of 19), and the TEMA-3 at posttest with pretest TEMA-3 scores below the 82$^{nd}$ sample percentile (24). Figure 1 depicts theses pretest moderation results.

We also explored other measures of pretest skill as moderators, which produced similar results. For example, the pretest value of each respective measure moderated the impact of ROOTS for the RAENS, NSB, and TEMA-3. For example, pretest NSB moderated the impact on posttest NSB. Similarly, the ASPENS composite also moderated the impact on the RAENS and NSB. Due to the redundancy in results and interpretation, we presented only the results when moderated by the pretest TEMA-3.

**Differential Response to Group-Size Differences based on Pretest Math Skill**

Finally, we tested whether initial TEMA-3 scores moderated the differences between group sizes. Pretest TEMA-3 scores moderated group size differences for the posttest TEMA-3 ($p$ = .0328,). Students with pretest TEMA-3 scores below the 16$^{th}$ sample percentile (9) performed better in small groups on the posttest TEMA-3. After correcting $p$-values with Benjamini-Hochberg procedure, however, group-size differences on the posttest TEMA-3 were not moderated by the initial TEMA-3 scores ($p$ = .2296). See Figure 2 depicts the relationship between baseline TEMA-3 scores and group-size differences, which chose that the confidence intervals included zero for the approximately 84% of the sample.

Analyses to examine differential response to group size differences based on initial ASPENS scores also demonstrated a moderation effect for the posttest ASPENS, NSB, and TEMA-3. The results were similar to those discussed above except that the confidence bounds included zero for 92% to 99% of the sample, and all Benjamini-Hochberg corrected $p$-values exceeded .068.

## Discussion

The results from this study add to a growing literature on the ROOTS intervention(Clarke, Doabler, Smolkowski, Kurtz Nelson, et al., 2016; Clarke et al., 2017; Clarke, Doabler, Smolkowski, Baker, et al., 2016; Doabler et al., 2016 ) with significant positive impacts across a range of proximal and distal outcome measures. The results from this study compliment research studies of other intervention programs targeting early mathematics content (e.g. Bryant & Bryant, 2008; Dyson et al., 2013; L. S. Fuchs et al., 2005; Gersten et al., 2012 ) shown to produce positive impacts on student mathematics achievement. Advancements in developing and studying the general efficacy of mathematics intervention programs are now allowing and leading the field to turn to questions that more fully investigate the conditions under which and for whom interventions work (Miller et al., 2014) including the types of questions investigated as part of this research study related to group size and initial skill.

Critically, results from this study indicated no difference in student outcomes whether the intervention was delivered to a two or five student small group (mean $g = 0.02$, range from −0.07 to 0.14). Across group sizes, students with lower initial skill received the greatest benefit from the intervention. Additionally, a trend was detected in which initial skill moderated outcomes by group size. That is, students with lower initial skills gained greater benefit from the smaller small group. However, this finding, found across initial TEMA and ASPENS scores, moved from significance to non-significance when multiple comparisons were considered.

Collectively, the results from the study have specific implications for how we should consider group size and student skill level when providing early mathematics intervention. The results shed light on a long standing, yet rarely tested, assumption that smaller groups sizes are "better" and in particular that as a student's educational need increases that services should be provided in groups of decreasing size. Generally speaking, we did not find significant differences across group sizes. This finding challenges the assumption that a smaller group will be more intensive and thus more impactful than a larger small group. However, there is some degree of nuance in this finding as we did detect a non-significant trend indicating for those students most at-risk there might have been a benefit to being in the smaller small group.

We hypothesized that group size was a critical proxy for intervention intensity because it smaller groups would enable a greater frequency of instructional interactions around critical mathematics content (Doabler et al., in press ). In this study, we utilized two direct observation instruments. One instrument, the Classroom Observations of Student-Teacher Interactions–Mathematics (COSTI-M; Doabler, Nelson, Stoolmiller, & Baker, 2015 ), measured the quantity of instructional interactions (i.e., overt teacher modeling, student practice opportunities, and academic feedback), whereas the second instrument, the Quality of Explicit Mathematics Instruction (QEMI; Doabler & Fien, 2013 ), measured the quality of such interactions. Results on our observation measures showed similar rates of teacher models and academic feedback across groups with significantly greater rates of individual practice in the two student small groups and significantly greater rates of group practice in the five student small group. Rating of overall

quality did not differ across group size. Thus, it could be hypothesized that for most students in the larger small group the intervention was of sufficient intensity and quality to meet their learning needs. However, given the trend noted related to initial skill and group size further research is needed to fully flesh out the complex relationship between group size (intervention intensity) and initial skill.

The research presented in this manuscript should be viewed in light of limitations of the study. Generalizability of findings should be weighed based on geographical and demographic variables and while the results presented here are from multiple cohorts in two distinct sites, additional replication studies are warranted (Coyne, Cook, & Therrien, 2016). A further limitation of the study is that we explored the questions of group size and initial skill within the context of one specific intervention program and thus drawing conclusions regarding other intervention programs and service delivery models is premature. We see this limitation as a spur to explore the specific conditions under which and for whom the ROOTS intervention works, but also as an imperative for the field to integrate more fully examinations associated with the provision of variations of intervention programs and educational services. Such research could focus contrasting treatments options systematically designed to vary on a key variable while holding other variables constant such as group size, as was done in this study.

Further exploration of variables related to implementation factors is also warranted (Fixsen, Blase, Metz, & van Dyke, 2013). Within our study, we included a degree of implementation support through professional development and coaching that is not typically when mathematics interventions are delivered. For example, it would be reasonable to hypothesize that the degree of coaching and support provided to the interventionists mitigated any differences between the two treatment conditions. Additional studies could explore coaching related questions in greater detail (e.g. provided limited support across conditions, or systematically varying coaching support with group size as a constant) or other variables of interest that would provide insight into ROOTS specifically and mathematics interventions generally. Such questions fit with the framework laid out by Onken and colleagues (2014) in

which early stage research focuses on implementation under ideal conditions (e.g. high degrees of implementation support) and later stages research shifting towards research under real world conditions. Investigations of ROOTS under real world conditions and specifically systematically manipulating the degree of implementation support provided through variables like coaching would provide valuable insight into the conditions under which ROOTS works.

Lastly, the work summarized in this manuscript and proposed for future research have implications for how MTSS are designed and how services within those systems are delivered. Choices in constructing MTSS (e.g. small group size, degree of coaching support) are typically evaluated exclusively in terms of benefit to students without consideration of costs. The cost aspect is important to consider in light of the relatively rare focus on examining the cost to benefit ratio in educational research (Levin & Belfield, 2015) despite increasing calls to consider the cost to benefit ratio of educational programs (Belfield & Bowden, 2019; IES RFP; US Dept. of Education, 2018). For example, if impact does not vary by group size, schools could decide to focus on serving a greater number of students by electing to provide intervention services in larger small groups. In the research described here on the ROOTS intervention, this would be deciding to deliver ROOTS in groups of five instead of two. Such a choice would allow schools to serve 150% more students. Conversely, schools could choose to serve the same number of students and reallocate resources to other efforts to improve student math achievement such as sustained professional development and coaching (Gersten, Taylor, Keys, Rolfhus, & Newman-Gonchar, 2014) to support effective mathematics teaching practices if research indicated a greater benefit from due to such support. Research that moves beyond determining an intervention's efficacy and towards understanding for whom, under what conditions and at what cost will move the field towards a greater understanding of how best to support the learning needs of all students in mathematics.

## References

Agrawal, J., & Morin, L. L. (2016). Evidence-based practices: Applications of concrete representational abstract framework across math concepts for students with mathematics disabilities. *Learning Disabilities Research & Practice, 31*(1), 34–44. doi: 10.1111/ldrp.12093

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in tier 1 or intervene immediately: A randomized experiment examining first-grade response to intervention in reading. *Exceptional Children, 81*, 11–27. doi: 10.1177/0014402914532234

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114. doi: 10.2307/271083

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

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. M. (2015). *Evaluation of response to intervention practices for elementary school reading* (NCEE Report No. 2016-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED560820.pdf

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. doi: 10.1080/00273170802034810

Belfield, C. R., & Bowden, A. B. (2019). Using resource and cost considerations to support educational evaluation: Six domains. *Educational Researcher, 48*(2), 120–127. doi: 10.3102/0013189X18814447

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. Retrieved from http://www.jstor.org/stable/2346101

Bryant, B. R., & Bryant, D. P. (2008). Introduction to the special series: Mathematics and learning disabilities. *Learning Disability Quarterly, 31*, 3–11. doi: 10.2307/30035521

Clarke, B., Doabler, C. T., & Nelson, N. J. (2014). Best practices in mathematics assessment and intervention with elementary students. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (6th ed., Vol. 1, pp. 219–232). Bethesda, MD: National Association of School Psychologists.

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

Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). *A randomized control trial of a tier 2 kindergarten mathematics intervention* (Project ROOTS). (USDE; Institute of Education Sciences; Special Education Research, CFDA Num: 84.324A, 2012-2016, Funding Number: R324A120304, awarded $3,338,552).

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. doi: 10.1177/2332858417706899

Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2016). Examining the efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of Learning Disabilities, 49*, 152–165. doi: 10.1177/0022219414538514

Clarke, B., Doabler, C. T., Smolkowski, K., Turtura, J., Kosty, D., Kurtz Nelson, E., . . . Baker, S. K. (2019). Exploring the relationship between initial mathematics skill and the impact of a kindergarten mathematics intervention on student mathematics outcomes. *Exceptional Children, 85*, 129–146. doi: 10.1177/0014402918799503

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.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (aspens)*. Longmont, CO: Cambium Learning Group, Sopris Learning.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234–248.

Codding, R. S., & Lane, K. L. (2015). A spotlight on treatment intensity: An important and often overlooked component of intervention inquiry. *Journal of Behavioral Education, 24*, 1–10. doi: 10.1007/s10864-014-9210-z

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

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. doi: 10.1177/0741932516648463

Coyne, M. D., Kame'enui, E. J., & Carnine, D. (2011). *Effective teaching strategies that accommodate diverse learners* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Dasgupta, A., Lawson, K. A., & Wilson, J. P. (2010). Evaluating equivalence and noninferiority trials. *American Journal of Health-System Pharmacy, 67*, 1337–1343. doi: 10.2146/ajhp090507

Doabler, C. T., Clarke, B., & Fien, H. (2012). *Roots assessment of early numeracy skills (raens)*. Unpublished measurement instrument. Center on Teaching and Learning, University of Oregon. Eugene, OR.

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*, 92–110. doi: 10.1177/0014402916660084

Doabler, C. T., & Fien, H. (2013). Explicit mathematics instruction: What teachers can do for teaching students with mathematics difficulties. *Intervention in School and Clinic, 48*, 276–285. doi: 10.1177/1053451212473151

Doabler, C. T., Nelson, N. J., Stoolmiller, M. L., & Baker, S. K. (2015). *Exploring alterable variables in Tier 1 and Tier 2 instruction: A collaboration across interdisciplinary fields of observational research (project cifor)*. (US Department of Education, Institute of Education Sciences, National Center of Education Research, Effective Teachers and Effective Teaching, CFDA Num: 84.305A, 2015-2017, Funding Number: R305A150037, awarded $699,706).

Doabler, C. T., Smith, J. L. M., Nelson, N. J., Clarke, B., Berg, T., & Fien, H. (2018). A guide for evaluating the mathematics programs used by special education teachers. *Intervention in School and Clinic, 54*(2), 97–105. doi: 10.1177/1053451218765253

Doabler, C. T., Smith, J. L. M., Nelson, N. J., Clarke, B., Berg, T., & Fien, H. (in press). A guide for evaluating the mathematics programs used by special education teachers. *Intervention in School and Clinic*.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi: 10.1037/0012-1649.43.6.1428

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. doi: 10.1177/0022219411410233

Fixsen, D., Blase, K., Metz, A., & van Dyke, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children, 79*, 213–230. Retrieved from http://cec.metapress.com/content/J47T21524330Q807

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

Fuchs, D., & Fuchs, L. S. (2017). Critique of the national evaluation of response to intervention: A case for simpler frameworks. *Exceptional Children, 83*, 255–268. doi: doi:10.1177/0014402917693580

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children, 85*(2), 126–128. doi: 10.1177/0014402918811924

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. doi: 10.1037/0022-0663.97.3.493

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2013). Intervention effects for students with comorbid forms of learning disability: Understanding the needs of nonresponders. *Journal of Learning Disabilities, 46*, 534–548. doi: 10.1177/0022219412468889

Fuchs, L. S., Fuchs, D., & Gilbert, J. K. (2019). Does the severity of students' pre-intervention math deficits affect responsiveness to generally effective first-grade intervention? *Exceptional Children, 85*(2), 147–162. doi: 10.1177/0014402918782628

Fuchs, L. S., Malone, A. S., Schumacher, R. F., Namkung, J., Hamlett, C. L., Jordan, N. C., . . . Changas, P. (2016). Supported self-explaining during fraction intervention. *Journal of Educational Psychology, 108*, 493–508. doi: 10.1037/edu0000073

Fuchs, L. S., Sterba, S. K., Fuchs, D., & Malone, A. S. (2016). Does evidence-based fractions intervention address the needs of very low-performing students? *Journal of Research on Educational Effectiveness, 9*, 662–677. doi: 10.1080/19345747.2015.1123336

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47*, 1539–1552. doi: 10.1037/a0025510

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 Report No. NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_math_pg_042109.pdf

Gersten, R. M., Clarke, B., Jordan, N., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78*, 423–445. Retrieved from http://cec.metapress.com/content/B75U2072576416T7

Gersten, R. M., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches* Report No. REL 2014-010). Washington, DC: US Department of Education, Institute of Education Sciences, Regional Educational Laboratory Southeast, National Center for Education Evaluation and Regional Assistance. Retrieved from http://files.eric.ed.gov/fulltext/ED544681.pdf

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability- third edition (tema-3)*. Austin, TX: ProEd.

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

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. doi: 10.1177/0193841x9602000306

Harcourt Educational Measurement. (2002). *Stanford achievement test [sat-10]*. San Antonio, TX: Author.

Institute of Education Sciences. (2018). *Rfa: Request for applications. Science, technology, engineering, and mathematics* (CFDA 84.324A Washington, DC: U.S. Department of Education. Retrieved from https://ies.ed.gov/ncser/projects/program.asp?ProgID=30

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

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

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. doi: 10.1016/j.lindif.2009.07.004

Kromrey, J. D., & Dickinson, W. B. (1996). Detecting unit of analysis problems in nested designs: Statistical power and type i error rates of the f test for groups-within-treatments effects. *Educational and Psychological Measurement, 56*, 215–231. doi: 10.1177/0013164496056002003

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. doi: 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. doi: 10.1080/19345747.2014.915604

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

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. doi: 10.1080/19345747.2014.927251

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. doi: 10.1177/0022219408331037

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

National Association of State Directors of Special Education. (2005). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: National Association of State Directors of Special Education. Retrieved from http://www.nasdse.org/Publications/tabid/577/product/43/Default.aspx

National Center on Response to Intervention. (2010). *Essential components of RTI - a closer look at response to intervention*. Washington, DC: U.S. Office of Special Education Programs; National Center on Response to Intervention. Retrieved from http://www.rti4success.org/

National Conference of State Legislatures. (2019). *Early STEM education*. Washington, DC: National Conference of State Legislatures. Retrieved from http://www.ncsl.org/research/education/early-stem-education.aspx

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: US Department of Education

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

Onken, L. S., Carroll, K. M., Shoham, V., Cuthbert, B. N., & Riddle, M. (2014). Reenvisioning clinical science: Unifying the discipline to improve the public health. *Clinical psychological science : a journal of the Association for Psychological Science, 2*(1), 22–34. doi: 10.1177/2167702613497932

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. W., & f, C. G. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the consort statement. *Journal of the American Medical Association, 295*(10), 1152–1160. doi: 10.1001/jama.295.10.1152

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

SAS Institute. (2016). Sas/stat® (Version 14.2) [ user's guide]. Cary, NC: SAS Institute. Retrieved from http://support.sas.com/documentation/index.html

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. doi: 10.1037//1082-989X.7.2.147

The White House Office of Science and Technology Policy. (2018). *Summary of the 2018 white house state- federal stem education summit*. Washington, DC: White House. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2018/06/Summary-of-the-2018-White-House-State-Federal-STEM-Education-Summit.pdf

Toll, S. W. M., & Van Luit, J. E. H. (2013). Accelerating the early numeracy development of kindergartners with limited working memory skills through remedial education. *Research in Developmental Disabilities, 34*(2), 745–755. doi: https://doi.org/10.1016/j.ridd.2012.09.003

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

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue:Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360. doi: 10.3102/0013189x14553660

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