Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Testing the Immediate and Long-Term Efficacy of a Tier 2 Kindergarten Mathematics Intervention

Ben Clarke, Christian Doabler, Keith Smolkowski, Evangeline Kurtz Nelson, Hank Fien, Scott K. Baker & Derek Kosty

To cite this article: Ben Clarke, Christian Doabler, Keith Smolkowski, Evangeline Kurtz Nelson, Hank Fien, Scott K. Baker & Derek Kosty (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, DOI: 10.1080/19345747.2015.1116034

To link to this article: https://doi.org/10.1080/19345747.2015.1116034

Accepted author version posted online: 06 Jan 2016.
Published online: 20 Jun 2016.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. 4, 607–634
http://dx.doi.org/10.1080/19345747.2015.1116034

Testing the Immediate and Long-Term Efficacy of a Tier 2 Kindergarten Mathematics Intervention

Ben Clarke[a], Christian Doabler[a], Keith Smolkowski[b], Evangeline Kurtz Nelson[a], Hank Fien[a], Scott K. Baker[c], and Derek Kosty[b]

ABSTRACT

This study examined the efficacy of a kindergarten mathematics intervention program, ROOTS, focused on developing whole-number understanding in the areas of counting and cardinality and operations and algebraic thinking for students at risk in mathematics. The study utilized a randomized block design with students within classrooms randomly assigned to treatment or control conditions. Measures of mathematics achievement were collected in the fall (pretest) and spring (posttest) in kindergarten and in the winter of first grade (delayed posttest). Significant differences between conditions favoring treatment students were found on four of six measures at posttest. Treatment students reduced the achievement gap with their not-at-risk peers. No effect was found on follow-up first-grade achievement scores. Implications for Tier 2 mathematics instruction in a Response to Intervention model are discussed.

KEYWORDS

intervention
mathematics
RTI

Although a significant amount of controversy surrounds the adoption and implementation of the Common Core State Standards (Common Core State Standards Initiative, 2010a), it is widely agreed that the new standards have greatly increased the expectations for the content that students are expected to acquire and accelerated the timeline on which that acquisition occurs. In part, these rigorous expectations are driven by the need to remain competitive in a globalized economy where job growth in science, technology, engineering, and mathematics fields is expected to outpace overall job growth at roughly a 3:1 ratio, and the array of opportunities within multiple settings will be increasingly dependent upon a fundamental understanding of mathematics (National Science Board, 2008). However, results from the 2013 National Assessment for Educational Progress (NAEP) indicate that relatively few students are equipped to meet these new standards, with only 42% of fourth graders at or above proficient in mathematics and 17% below basic. Results are even more disconcerting for students with learning disabilities, with only 18% of students at or above proficient and 45% scoring in the below-basic category (National Center for Education Statistics, 2013).

One promising approach to meeting the learning needs of students at risk for or with mathematics learning disabilities (MLD) is the provision of instruction through a tiered

CONTACT Ben Clarke mathctl@uoregon.edu University of Oregon, Center on Teaching and Learning, 5292 University of Oregon, Eugene, OR 97403, USA.
[a]University of Oregon, Eugene, Oregon, USA
[b]Oregon Research Institute, Eugene, Oregon, USA
[c]Southern Methodist University, Dallas, Texas, USA

system of instructional support typically referred to as a multitier system of support (MTSS) or response to intervention (RTI) (Fuchs & Vaughn, 2012). Originally conceptualized as a potential mechanism or component of a comprehensive evaluation to determine LD eligibility for special education services (Individuals With Disabilities Education Act, 2004; Vaughn & Fuchs, 2003), RTI provides support to students through levels or tiers of support with greater intensity being provided as the complexity and intractability of a student's learning problems increase (Burns & Vanderheyden, 2006; Fuchs, Fuchs & Zumeta, 2008). As part of this general movement toward tiers of levels of service delivery, an increased emphasis has been placed on the early intervention and prevention of learning disabilities before they develop and become a significant impediment to student learning of more advanced academic content (National Association of State Directors of Special Education, 2006).

The importance of early intervention and prevention is particularly relevant to the development of mathematics understanding. Longitudinal research indicates that students who perform poorly in mathematics in the early elementary grades are likely to continue to struggle throughout elementary school (Bodovski & Farkas, 2007; Duncan et al., 2007; Hanich, Jordan, Kaplan, & Dick, 2001; Morgan, Farkas, & Wu, 2009) and that differences between students with MLD and their on-track peers widens over time (Morgan, Farkas, & Wu, 2011). Using a nationally representative sample of students from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), Morgan, Farkas, Hillemeier, and Maczuga (2014) found that students experiencing low mathematics achievement experienced persistent difficulty in mathematics throughout elementary and middle school by a multiplicative factor of almost 17 and that this relationship was significantly stronger than other kindergarten predictors (e.g., cognitive delays). In tandem with additional findings from Morgan and colleagues (2011) that fifth-grade students with an MLD profile in kindergarten had achievement gaps of two standard deviation units with fifth-grade students who did not demonstrate an MLD profile in kindergarten, these findings argue strongly for focused efforts in kindergarten aimed at the prevention of long-term difficulties in mathematics.

Ideally, prevention efforts, and in particular those in the early elementary grades, would begin through the delivery of a research-based core curricular program (Fuchs, Fuchs, & Compton, 2013). Emerging evidence suggests that effective core programs can positively impact mathematics outcomes, including the achievement of at-risk students. For example, in a comparison of four first-grade curricula, Agodini and Harris (2010) found differential program impact on the achievement of students in the lower third of the study sample. Similarly, Clarke, Smolkowski, and colleagues (2011) found a differential impact for students identified as at risk on a test of mathematics achievement when investigating the efficacy of a core kindergarten mathematics curriculum. However, even with the provision of generally effective core programs, achievement gaps have not been fully closed. Based on limited research on core programs, a lack of emphasis on mathematics relative to reading and social skills development (La Paro et al., 2009), and the lack of design elements embedded in core programs specifically targeting the learning needs of at-risk students (Doabler, Fien, Nelson-Walker, & Baker, 2012; Sood & Jittendra, 2007), it is likely that some students, especially those with or at risk for MLD, will need additional and intensive instructional support.

In kindergarten, prevention efforts in mathematics often center on the development of number sense. Although difficult to operationally define (Berch, 2005), number sense allows students to connect their initial understandings of mathematical concepts to numerical relationships (Gersten & Chard, 1999). However, many students fail to develop this informal

understanding of number prior to school entry (Klein, Starkey, Clements, Sarama, & Iyer, 2008), and thus have difficulties in accessing the formal mathematics taught during the kindergarten year (Jordan, Kaplan, Olah, & Locuniak, 2006). Thus, a focus on number sense and its connection to whole-number concepts and procedures offers a promising area for intervention development. Although the depth of intervention development in mathematics pales in comparison to efforts in reading (Gersten et al., 2007), and the use of rigorous research designs to evaluate efficacy is limited (Dyson, Jordan, & Glutting, 2013), there are a growing number of intervention programs targeting kindergarten students at risk in mathematics that have been rigorously evaluated (e.g., Clarke et al., 2016; Dyson et al., 2013; Fuchs et al., 2005; Gersten et al., 2015). Common elements across intervention programs include a dual focus on critical whole-number concepts (Gersten, Beckmann et al., 2009; National Mathematics Advisory Panel, 2008) and the integration of systematic and explicit instructional design elements specific to the learning needs of at-risk students (Baker, Gersten, & Lee, 2002; Gersten, Chard, et al., 2009; Kroesbergen & Van Luit, 2003). There exists a compelling need to continue the expansion of the research base on early mathematics interventions (Gersten, Chard, et al., 2009).

Purpose of the Study

This article examines the effects of a kindergarten intervention curriculum, ROOTS, on student achievement. The study will add to the existing base on Tier 2 kindergarten intervention curricula. We investigate the following a priori research questions:

1. Does the ROOTS curriculum produce greater achievement gains at the end of kindergarten than standard district mathematics instruction?
2. Is the achievement gap for ROOTS students reduced by ROOTS students making greater gains than their not-at-risk peers?
3. Does the ROOTS curriculum produce greater achievement gains at the middle of first grade than standard district mathematics instruction?

Due to its dual focus on critical whole-number content and the use of validated instructional design principles, we hypothesize that ROOTS will produce greater achievement gains at the end of kindergarten for ROOTS students than standard district mathematics instruction for control students and that those gains will enable ROOTS students to reduce the gap with their not-at-risk peers. Based on the persistent lack of findings regarding the long-term impact of mathematics interventions (Starkey & Klein, 2008), we also hypothesize that the impact of ROOTS will fade by the middle of first grade.

Method

Research Design and Context

Efficacy of the ROOTS intervention was examined in a randomized controlled trial that utilized a partially nested design, with students nested within interventionists and interventionists nested within classrooms. This study analyzed data collected during Year 1 of a four-year efficacy trial funded by the Institute of Education Science (IES). Blocking on classrooms, the 10–12 lowest performing students from each participating kindergarten classroom were randomly assigned to one of three conditions: (a) a ROOTS instructional group with a

2:1 student–teacher ratio, (b) a ROOTS instructional group with a 5:1 student–teacher ratio, or (c) a no-treatment control condition. Students randomly assigned to the two treatment groups received the ROOTS intervention in addition to district-approved core mathematics instruction. Students in the control condition received district-approved core mathematics instruction only. In total, 58 ROOTS intervention groups were conducted in Year 1. The unit of analysis for this study was instructional groups.

Although a primary aim of the four-year efficacy trial is to identify potentially important differences between the two treatment groups (i.e., 2:1 ROOTS vs. 5:1 ROOTS), the Year 1 data set by itself is underpowered to conduct such an investigation. Therefore, in this study, students in the ROOTS groups were combined to compare their gains in important mathematics outcomes relative to students in the no-treatment control condition. Student-level mathematics achievement data were collected during the kindergarten year at the intervention's pretest (T1) and posttest (T2) times, and at a follow-up (T3) approximately six months into the students' first-grade year.

The study took place in 37 kindergarten classrooms from four school districts in Oregon. One school district was located in the metropolitan area of Portland, while the remaining three districts were located in suburban and rural areas of western Oregon. Across the four districts, student enrollment ranged from 2,736 to 38,557 students. Within the 14 participating schools, between 8%–23% of students received special education services, 5%–68% were English learners, 17%–86% were eligible for free or reduced-price lunch, 0%–12% were American Indian or Native Alaskan, 0%–16% Asian, 0%–9% were Black, 0%–74% were Hispanic, 0%–2% were Native Hawaiian or Pacific Islander, 46%–92% were White, and 0%–15% were more than one race. Schools targeted for recruitment across the four districts were primarily those that received Title 1 funding.

Participants

Classrooms

A total of 37 kindergarten classrooms participated in the first year of the efficacy trial. Of these classrooms, 33 provided a half-day kindergarten program and four provided a full-day program. All classrooms operated five days per week and provided mathematics instruction in English. Average class size was 23.0 students ($SD = 6.0$).

Twenty-eight certified teachers taught the 37 classrooms. Nine of the teachers taught two half-day classrooms (i.e., AM and PM), and all teachers participated for the duration of the study. Most of the teachers were female (98%) and had, on average, 17.21 years of teaching experience and 8.94 years of experience teaching at the kindergarten level. The majority of the teachers held a graduate degree in education (85%) and 65% had completed college-level coursework in Algebra. Of the 28 teachers, 81% identified themselves as White, 12% as Asian American/Pacific Islander, and 4% as representing another ethnic group.

Interventionists

District-employed instructional assistants and interventionists hired specifically for the study taught the 58 ROOTS intervention groups. Observation data collected in a previous investigation of ROOTS showed that interventionists with these types of backgrounds were able to deliver the ROOTS curriculum with a high degree of implementation fidelity and quality (Clarke et al., 2016). Most interventionists in this study were

female (98%) and 61% held a bachelor's degree or higher. Interventionists had an average of eight years of experience working in schools and approximately 20% held current teaching licenses. Among the interventionists, 94% identified themselves as White, 3% as Hispanic, and 3% as representing another ethnic group.

Criteria for Participation

The research team applied a three-step process to identify students who were at risk for mathematics difficulties and, in turn, might qualify for the ROOTS intervention. First, within the 37 participating classrooms, all kindergarten students with parental consent were screened in late fall of their kindergarten school year. Screening measures included two standardized assessments of early mathematics: Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke, Gersten, Dimino, & Rolfhus, 2011) and the Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2008). Students qualified for the intervention if they scored 20 or less on the NSB (Jordan et al., 2008) and had a composite score on the ASPENS that placed in the *strategic* or *intensive* range (Clarke, Gersten, et al., 2011). These thresholds were selected because prior research suggests that students who score in these ranges at the start of kindergarten are at risk for developing long-term mathematics difficulties (Clarke, Smolkowski, et al., 2011; Jordan et al., 2008).

Second, prior to random assignment, students' ASPENS and NSB scores were separately converted into standard scores and then combined to form an overall composite standard score. Third, students' composite standard scores were rank-ordered and the lowest 10 students that met our established a priori criteria were considered eligible for random assignment. Students were then randomly assigned to one of three conditions: (a) a ROOTS two-student group, (b) a ROOTS five-student group, or (c) a no-treatment control condition. Group sizes were selected to represent the lower (i.e. two-student groups) and upper (i.e. five-student groups) bound of small groups commonly utilized in schools. In classrooms with strong histories of high student mobility, the 11th and 12th ranked students who met the inclusion criteria were selected as alternates and randomized to one of the two treatment conditions. Of the 37 classrooms, 23 had a sufficient number of eligible students to participate as a ROOTS classroom and 14 had an insufficient number of eligible students. In situations where classrooms had insufficient numbers of students to form the ROOTS intervention and control groups, classrooms were combined. This cross-class grouping procedure was applied six times. In four instances using eight total classrooms, sets of two classrooms were combined creating four ROOTS classrooms, and in two instances using six classrooms, sets of three classrooms were combined creating two ROOTS classrooms. Collectively, 29 "ROOTS" classrooms participated ($23 + 4 + 2$).

Students

A total of 850 kindergarten students were screened in late fall of 2012 to determine eligibility for the ROOTS intervention. Of the 850 kindergarten students, 290 met the inclusion criteria and were randomly assigned to the two-student ROOTS condition ($n = 58$), five-student ROOTS condition ($n = 145$), or the no-treatment control condition ($n = 87$). As discussed, students in the ROOTS groups were combined in this study to draw comparisons of their relative performance to students in the no-treatment control condition. Table 1 shows student demographic information for all students screened in the fall of 2012 differentiated by ROOTS eligibility and intervention condition.

Table 1. Descriptive statistics for student characteristics by condition.

Student characteristic	ROOTS-eligible students		Typically achieving students ineligible for ROOTS
	ROOTS	Control	
Age at pretest, *M* (*SD*)	5.2 (0.4)	5.2 (0.4)	5.4 (0.5)
Male	45%	41%	51%
Race			
American Indian/Alaskan Native	4%	3%	2%
Asian	2%	5%	6%
Black	4%	5%	2%
Native Hawaiian/Pacific Islander	0%	0%	0%
White	56%	60%	65%
More than one race	3%	2%	4%
Hispanic	33%	32%	16%
Limited English proficiency	32%	31%	13%
SPED eligible	13%	9%	5%

Note. The sample included 203 students in the ROOTS condition, 87 in the control condition, and 560 typically achieving students who were not eligible for ROOTS.

Procedures

ROOTS Intervention

ROOTS is a 50-lesson, Tier 2 kindergarten intervention program designed to build students' proficiency with critical concepts and skills of whole numbers. In the current study, the intervention was delivered in 20-minute, small-group sessions, five days per week for approximately 10 weeks. The intervention was scheduled for delivery at a time that did not conflict with students' core mathematics and reading instruction. ROOTS instruction began in early December and continued through March. The intervention start date was selected because it allowed adequate opportunity for students to respond to core (Tier 1) mathematics instruction and minimized false positive errors during the screening process (i.e., identifying typically achieving students as in need of the intervention).

The primary aim of ROOTS is to support students who struggle with mathematics in developing procedural fluency with and conceptual understanding of whole-number concepts and skills identified in the Common Core State Standards for Mathematics (CCSS-M, 2010b). ROOTS is based on two key principles: (a) focused whole-number content and (b) the use of validated instructional design principles. We hypothesize that the integration of these two principles in the ROOTS curriculum impacts student development of mathematics in two interdependent knowledge forms: (a) conceptual understanding and (b) procedural fluency. We consider conceptual understanding and procedural fluency as interdependent constructs that develop in tandem and reciprocally (Rittle-Johnson, Siegler, & Alibali, 2001; Wu, 1999) and that these in turn impact overall student achievement in mathematics. Specifically, the intervention prioritizes topics from two kindergarten domains in the CCSS-M: (a) counting and cardinality and (b) operations and algebraic thinking. Table 2 shows the CCSS-M standards addressed by lesson in ROOTS. The intense focus on whole numbers aligns with calls from mathematicians and expert panels to support all students, particularly at-risk learners, in developing robust and lasting number sense (Gersten, Beckmann, et al., 2009; National Research Council, 2009). ROOTS provides in-depth instruction in whole-number concepts by strategically linking the informal mathematical knowledge students

Table 2. Common core state standards addressed by ROOTS.

	Common Core state standards	ROOTS lessons 1–10	11–20	21–30	31–40	41–50
Counting and Cardinality						
	Know number names and the count sequence.					
1	Count to 100 by ones and by tens.	1 to 5	1 to 8	1 to 20	1 to 20	1 to 20
2	Count forward beginning from a given number within the known sequence (instead of having to begin at 1).		✓	✓	✓	✓
3	Write numbers from 0 to 20. Represent a number of objects with a written numeral 0–20 (with 0 representing a count of no objects).	✓	✓	✓	✓	✓
	Count to tell the number of objects.					
4	Understand the relationship between numbers and quantities; connect counting to cardinality.	✓	✓	✓	✓	✓
4.a	When counting objects, say the number names in the standard order, pairing each object with one and only one number name and each number name with one and only one object.		✓	✓	✓	✓
4.b	Understand that the last number name said tells the number of objects counted. The number of objects is the same regardless of their arrangement or the order in which they were counted.	✓	✓	✓	✓	✓
4.c	Understand that each successive number name refers to a quantity that is one larger.		✓	✓	✓	✓
5	Count to answer "how many?" questions about as many as 20 things arranged in a line, a rectangular array, or a circle, or as many as 10 things in a scattered configuration; given a number from 1–20, count out that many objects.	✓	✓	✓	✓	✓
	Compare numbers.					
6	Identify whether the number of objects in one group is greater than, less than, or equal to the number of objects in another group, e.g., by using matching and counting strategies (groups up to 10 numbers).	✓	✓	✓	✓	✓
7	Compare two numbers between 1 and 10 presented as written numerals.		✓	✓	✓	✓
Operations and Algebraic Thinking						
	Understand addition as putting together and adding to, and understand subtraction as taking apart and taking from.					
1	Represent addition and subtraction with objects, fingers, mental images, drawings, sounds (e.g., claps), acting out situations, verbal explanations, expressions, or equations.		✓	✓	✓	✓
2	Solve addition and subtraction word problems, and add and subtract within 10, e.g., by using objects or drawings to represent the problem.		✓	✓	✓	✓
3	Decompose numbers less than or equal to 10 into pairs in more than one way, for example, by using objects or drawings, and record each decomposition by a drawing or equation (e.g., $5 = 2 + 3$ and $5 = 4 + 1$).		✓	✓	✓	✓

(*Continued on next page*)

Table 2. (*Continued*).

	Common Core state standards	ROOTS lessons				
		1–10	11–20	21–30	31–40	41–50
4	For any number from 1 to 9, find the number that makes 10 when added to the given number, for example, by using objects or drawings, and record the answer with a drawing or equation.					✓
5	Fluently add and subtract within 5.		✓	✓	✓	✓
Number and Operations in Base Ten						
	Work with numbers 11–19 to gain foundations for place value.					
1	Compose and decompose numbers from 11 to 19 into ten ones and some further ones, for example, by using objects or drawings, and record each composition or decomposition by a drawing or equation (e.g., $18 = 10 + 8$); understand that these numbers are composed of ten ones and one, two, three, four, five, six, seven, eight, or nine ones.			✓	✓	✓

acquire prior to school entry with the formal mathematical knowledge developed in their kindergarten year.

For example, as shown in Table 2, the intervention's first two weeks (lessons 1–10) prioritize numbers 1 to 5. With these initial numbers, students learn the concepts of cardinality, less than, greater than, and equal to. Students also lean how to rational count, write numbers, and use math models (e.g., counting blocks) to represent numbers. Then, as students gain a deeper understanding of numbers, the intervention begins to focus on place-value concepts, providing students with deliberate opportunities to work with teen numbers (lessons 21–30). To minimize student confusion and reduce the potential for misconceptions, ROOTS systematically introduces teen numbers one at a time and interweaves them with previously learned numbers.

The ROOTS intervention is grounded in an *explicit* and *systematic* framework of mathematics instruction. A growing body of rigorous experimental research has reported the beneficial impact of this instructional approach for students with mathematics difficulties (Baker et al., 2002; Clarke et al. 2016; Doabler, Strand Cary, et al., 2012; Gersten, Beckmann, et al., 2009). Central to the ROOTS intervention are explicit instructional design and delivery principles that have been empirically validated to accelerate the mathematics learning of at-risk learners. Specifically, ROOTS includes scripted guidelines for interventionists to facilitate four essential features of explicit mathematics instruction: (a) teacher modeling, (b) deliberate practice, (c) visual representations of mathematics, and (d) academic feedback. Teacher modeling is defined as overt explanations and demonstrations of critical math content (Archer & Hughes, 2011; Doabler & Fien, 2013). For example, in ROOTS, interventionists might demonstrate the concept of cardinality by overtly counting objects to determine how many in a countable set. Interventionists would model a range of examples that depict counting objects in different configurations (e.g., line, array, circle, and scattered). Deliberate practice in ROOTS is designed to promote high rates of learning success and eventual learner independence. Research suggests that when practice opportunities are systematically planned and scaffolded, they permit students to acquire new knowledge, retain previously learned material, and connect existing background knowledge with new and more sophisticated content (Doabler et al., 2015; Doabler, Clarke, & Fien, 2012; Gersten, Beckmann, et al., 2009). Visual representations of mathematics help deepen students' conceptual understanding. In ROOTS we utilized a number of representations, including number lines and counting and place-value blocks, to aid in developing understanding of key concepts. Academic feedback is defined as an interventionist providing informational feedback to correct a student mistake (Archer & Hughes, 2011). When interventionists deliver academic feedback they are able to address student misconceptions and potential knowledge gaps. In ROOTS, academic feedback is designed to be immediate and specific to the student error. For example, if a student makes a mistake in determining how many objects are in a countable set, interventionists are expected to model how to ascertain the cardinality of the set.

When implemented as designed, ROOTS lessons assist interventionists in deeply engaging students in important whole-number concepts and skills. One instructional interaction at the forefront of the intervention is mathematical discourse or student mathematics verbalizations (Doabler et al., 2015). ROOTS facilitates structured opportunities for struggling learners to verbalize their mathematical thinking and discuss their solution methods for solving whole-number problems. For example, an interventionist will have pairs of students explain how they solved "add to" word problems with the total unknown (e.g., "Three cubes

are on the table. Two more cubes are placed on the table. How many cubes are now on the table?"). In this example, the students might state how they used three counting cubes to show the initial quantity and then included two more cubes to show the increase. Next, the students might explain how they derived their final answer (e.g., "We combined the cubes and counted them altogether. There are five cubes."). The interventionist will conclude the interaction by having students state the completed addition equation (e.g., "Three cubes plus two cubes equals five cubes"). Verbalizations are also directed to individuals so that interventionists can gauge whether particular learners are grasping key concepts, such as the relationship between a bundle of ten and ten single ones.

Professional Development

Participating interventionists attended two five-hour professional development workshops focused on the ROOTS intervention. The initial workshop targeted the instructional objectives of Lessons 1–25, whole-number concepts and skills identified in the CCSS-M (Common Core State Standards Initiative, 2010b), small-group management techniques, and instructional practices that have been empirically validated to increase student mathematics achievement, such as provision of student mathematics verbalizations. The second workshop had a similar agenda but focused on the mathematical content prioritized in the intervention's final 25 lessons. Interventionists were given opportunities throughout both workshops to practice with sample lessons and receive feedback on their instructional delivery from key research staff.

To bolster implementation fidelity and enhance the quality of Tier 2 mathematics instruction, each ROOTS group received between three and four in-class coaching visits. Five former educators, who were knowledgeable in the science of early mathematics development and instruction, served as coaches during the study. Coaching visits facilitated an observation-feedback loop that began with direct observation of intervention implementation followed by a debriefing on the quality of instructional delivery and implementation fidelity.

Control Condition

Core (Tier 1) mathematics instruction delivered in the 37 kindergarten classrooms served as the control condition. All students, including treatment and control students, received core mathematics instruction. As discussed, the ROOTS intervention occurred outside of, and in addition to, core mathematics instruction. To document the instructional practices and mathematics content employed in the control condition, research staff administered two surveys and conducted one direct observation of core mathematics instruction during the intervention time period. Observation data suggested that classroom teachers used a variety of published and teacher-developed mathematics programs. These materials were found to vary within and across participating schools. Several of the programs were leading sellers in the U.S. elementary textbook market, including Everyday Mathematics, Bridges in Mathematics, Investigations, Saxon Math, and Houghton Mifflin.

Surveys revealed that core mathematics instruction was delivered approximately 30 minutes per day, four to five days per week. Instruction occurred through a variety of different mediums, including learning centers, small group activities, and whole-class-delivered instruction. The instructional focus varied, with some teachers focusing more on whole-number concepts and others focusing on particular aspects of geometry and measurement. However, teachers reported that operations and algebraic thinking and geometry were the

primary mathematics domains of the CCSS-M targeted during core instruction. Instructionally, some classroom teachers were found to use explicit instructional practices, such as teacher modeling, structured practice opportunities for students, and corrective feedback. Survey data and observations of core math instruction indicated no evidence of treatment diffusion or contamination of ROOTS instruction.

Fidelity of Implementation

Trained research staff observed each ROOTS group three times across the intervention time period. Observations were used to gauge implementation fidelity or the extent to which interventionists implemented the intervention as intended. The fidelity measure focused on four features of implementation adherence, including the extent to which ROOTS interventionists: (a) delivered the prescribed number of activities in the observed lesson, (b) met the observed lesson's instructional objectives, (c) followed the teacher scripting, and (d) used the prescribed mathematics models. To measure the number of activities taught during an observation occasion, research staff used a copy of the ROOTS program to follow along with the small-group instruction and coded if an activity had been taught (1 = taught) or not taught (0 = not taught). On average, interventionists taught the majority of the prescribed activities ($M = 4.2$ out of 5, $SD = 0.5$). The final three features were measured using a 4-point scale (4 = all, 3 = most, 2 = some, 1 = none). Results also indicated that interventionists met the instructional objectives ($M = 3.5$, $SD = 0.5$), adhered to the prescribed teacher scripting ($M = 3.3$, $SD = 0.6$), and used the ROOTS mathematical models ($M = 3.6$, $SD = 0.5$). The ICC for the aggregate fidelity score was .60. The individual fidelity items were (1) met the math objectives (ICC = .58), (2) adherence to teacher scripting (ICC = .62), and (3) used the prescribed math models (ICC = .45). These ICCs suggest moderate to substantial agreement between observers (Landis & Koch, 1977).

Intervention dosage served as another metric of implementation fidelity. In this study, dosage was operationally defined as the amount of the intervention received by the participating students. Of the 58 intervention groups, 57 completed 97% or more of the prescribed lessons, while one group completed 88% of the lessons. Dosage also represented the degree to which interventionists delivered ROOTS at the prescribed intervention frequency (i.e., sessions per week) and duration (i.e., session length). As intended, students in all treatment groups received ROOTS at a frequency of five sessions per week.

Measures

Students were assessed at pretest (T1) and posttest (T2) on measures of foundational aspects of number sense and whole-number understanding. The measurement net included a measure of whole-number understanding considered proximal to the ROOTS intervention, a set of early mathematics curriculum-based measures that focused on discrete skills of number sense, and two distal outcome measures of students' procedural and conceptual knowledge of whole numbers. A third distal outcome measure was administered as a follow-up assessment (T3) approximately six months into students' first-grade year. Trained staff administered all student measures, with data collection meeting acceptable reliability criteria (i.e., implementation fidelity of .95 or higher).

ROOTS Assessment of Early Numeracy Skills (RAENS)

To assess proximal outcomes in early numeracy, a researcher-developed instrument, RAENS (Doabler et al., 2012) was administered at pretest and posttest time periods. RAENS is an individually administered assessment consisting 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 with widely used measures of mathematics achievement including the TEMA and the NSB. Inter-rater scoring agreement is reported at 100% (Clarke, Doabler, Smolkowski, Fien, & Baker, 2014).

Assessing Student Proficiency in Early Number Sense (ASPENS)

ASPENS (Clarke, Gersten, et al., 2011) is a set of three curriculum-based measures validated for screening and progress monitoring in kindergarten mathematics (Clarke, Gersten, et al., 2011). 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 (Clarke, Gersten, et al., 2011).

Number Sense Brief (NSB) Screen

The NSB (Jordan et al., 2008) is an individually administered test with 33 items that assess counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations. Authors report a coefficient alpha of .84 at the beginning of first grade.

Test of Early Mathematics Ability-Third Edition (TEMA-3)

The Test of Early Mathematics Ability-Third Edition (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 math outcome measures, the TEMA-3 manual reports coefficients ranging from .54 to .91.

The Stanford Achievement Test-Tenth Edition (SAT-10)

The SAT-10 measure (Harcourt Educational Measurement, 2002) 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 that note whether the student is below, at, or above average for specific skill clusters. All participating students, including students who did not meet criteria on the screening measures, were administered the SESAT at posttest (T_2) and the SAT-10 midway through their first-grade year (T_3).

Statistical Analysis

We assessed intervention effects on each of the primary outcomes with a mixed model (multilevel) time × condition analysis (Murray, 1998) designed to account for students partially nested within small groups (Baldwin, Bauer, Stice, & Rohde, 2011; Bauer, Sterba, & Hallfors, 2008). The study design called for the randomization of individual students to receive ROOTS, nested within small 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 ROOTS groups required a group-level variance while the unclustered controls did not. Furthermore, because the residual variances may have differed between conditions, we tested the assumption of homoscedasticity of residuals. We next describe the details of the analysis strategy.

Baldwin et al. (2011) and Bauer et al. (2008) presented a mixed-model analysis of variance approach to account for the different variance structures between conditions and tests for heteroscedastic residual variances, and we expand their approach to a time × condition analysis. The analysis tests for differences between conditions on gains in outcomes from the fall (T_1) to spring (T_2) of kindergarten. The basic model includes time, T, coded 0 at T_1 and 1 at T_2, condition, C, coded 0 for control and 1 for ROOTS, and the interaction between the two:

$$Y_{ij} = \pi_{0j} + \pi_{1j}C_j + \pi_{2j}T_{ij} + \pi_{3j}T_{ij}C_j + e_{ij} \sim \mathrm{N}\left(0,\ \sigma^2\right) \quad (1)$$

$$\pi_{0j} = \beta_{00} \quad (2)$$

$$\pi_{1j} = \beta_{10} \quad (3)$$

$$\pi_{2j} = \beta_{20} \quad (4)$$

$$\pi_{3j} = \beta_{30} + r_{3j} \sim \mathrm{N}\left(0,\ \tau^2\right) \quad (5)$$

For Level 1, the first equation, Y_{ij} represents a score for individual i within small-group cluster j, and the model includes condition, C_j, time, T_{ij}, and their interaction as predictors. While the e_{ij} are distributed N(0, σ^2), the time × condition analysis decomposes the individual-level variance, σ^2, into a variance for the assessments, σ_s^2, and covariance between the T_1 and T_2 assessments, r_s^2, where σ_s^2 and r_s^2 sum to σ^2 (Murray, 1998). To minimize unnecessary complexity, however, we focus on σ^2 for the following discussion but present both the variance and covariance terms in the results. Because individual students were assigned to condition and only partially nested, the model differs in two ways from models used for fully clustered randomized trials (FCRT). First, the "cluster" j has a unique value for each group in the intervention condition and a unique value for each individual student in the control condition. Second, where condition typically resides at the cluster level in FCRT, in this partially clustered trial we include condition at the individual level.

The Level 2 equations predict scores with (a) an intercept, π_{0j}, which represents the pretest control-group mean; (b) the difference between conditions at pretest, π_{1j}; (c) the slope for control students, π_{2j}; and (d) the difference between conditions on the slope, π_{3j}. The intercept, condition effect at pretest, and slope for control students do not require cluster variances because they represent either unclustered control-group effects or differences between conditions at pretest, before students were clustered. The model included a cluster-level variance, r_{3j}, for the time × condition effect to account for the posttest

clustering that occurs only in the intervention condition. The following composite equation is obtained by substituting the Level 2 equations into the Level 1 equation, in which r_{3j} is defined only for the intervention condition ($C_j = 1$):

$$Y_{ij} = \beta_{00} + \beta_{10}C_j + \beta_{20}T_{ij} + \beta_{30}T_{ij}C_j + \left(T_{ij}C_j r_{3j} + e_{ij}\right) \tag{6}$$

Due to the unbalanced nesting structure, the residual variance may differ by condition, so we fit two models. The homoscedastic model shown in Equation 6 assumed a single residual error term. The second model assumed heteroscedastic residual variances as follows:

$$\mathrm{V}(eij \mid Cj = 0) = \sigma_0^2 \tag{7}$$

$$\mathrm{V}(e_{ij} \mid C_j = 1) = \sigma_1^2 \tag{8}$$

Both σ_0^2 and σ_1^2 decompose into a variance and covariance in the analyses, as described above, and we report both in the results.

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 we test for equivalence, or the noninferiority, of the simpler model when compared to the more complex model, we must reverse the null and alternative hypotheses and, hence, the α and β values that represent Type I and Type II error rates as we might in equivalence or noninferiority trials (e.g., Dasgupta, Lawson, & Wilson, 2010; Piaggio, Elbourne, Altman, Pocock, & Evans, 2006). For this reason, as well as the low statistical power to detect differences between variance structures (Kromrey & Dickinson, 1996), we compare models with likelihood ratio test using $\alpha = .20$ as our criterion Type I error rate and, as a consequence, report the more complex model unless we are relatively certain the two are equivalent.

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, the SESAT available only at posttest and the SAT-10 collected as a follow-up measure in Grade 1, we used the analysis of covariance approach described by Bauer et al. (2008) and Baldwin et al. (2011). For the analysis of covariance approach, we also compared homoscedastic and heteroscedastic models with likelihood ratio tests. In all models, based on recommendations of Baldwin et al., we used Satterthwaite approximation to determine the degrees of freedom.

Because students were randomly assigned within classrooms and schools, we tested an additional set of models that extended those discussed above to account for clustering within classrooms or schools. Raudenbush and Sadoff (2008) have shown that "the between-school variance component ... does not influence [the variance of the treatment-effect estimator]: One of the key advantages of randomizing within schools is that school-level variance in the mean outcome is removed from the experimental error variance" (p. 146), and this result would generalize to the classroom-level in the present study. Consistent with their observation, the overall pattern of intervention results remained similar in all models, whether or not we included classroom or school levels in the model. Please see the appendix for test of condition-effect variability by school and classroom.

Model Estimation

We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute, 2009) using restricted maximum likelihood (REML), generally recommended for multilevel models (Hox, 2002). Maximum likelihood estimation for the time × condition analysis uses 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). In the present study, 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. The data in the present study also do not markedly deviate from normality; skewness and kurtosis fell with ± 2.0 for all measures except oral counting, where kurtosis was 3.3. Nonetheless, multilevel regression methods have also been found quite robust to violations of normality (e.g., Hannan & Murray, 1996).

Effect Sizes

To ease interpretation, we computed an effect size, Hedges's *g* (Hedges, 1981), for each fixed effect as described by the What Works Clearinghouse (2014). Hedges's *g* represents an individual-level effect size comparable to Cohen's *d* (Cohen, 1988; Rosenthal & Rosnow, 2008).

Results

Table 3 presents means, standard deviations, and sample sizes for the seven dependent variables by ROOTS eligibility, assessment time, and condition. Below we present results from tests of attrition effects, ROOTS intervention impact, and differences between students who received ROOTS compared to typically achieving peers who were ineligible for ROOTS.

Attrition

Student attrition was defined as students with data at T_1 but missing data at T_2, and we examined attrition with respect to the ROOTS-eligible sample of 290 students. Attrition rates varied between 10%, for the TEMA and RAENS, and 12% for NSB, ASPENS, and oral counting. Only 6.9% of students were missing all posttest data. The proportion of students missing all posttest data did not differ between conditions ($\chi^2_{(1)} = 0.26$, $p = .6131$). Although differential rates of attrition are undesirable, differential scores on math tests present a far greater threat to validity, so we conducted an analysis to test whether student math scores were differentially affected by attrition across conditions. We examined the effects of condition, attrition status, and their interaction on pretest scores for all five measures available at pretest. We found no statistically significant interactions for ASPENS, oral counting, or TEMA scores.

The analysis produced a statistically significant interaction for the pretest (T_1) NSB total score ($t = 2.22$, $df = 232$, $p = .0276$). Control students with NSB data at T_2 scored 3.4 points higher than those without T_2 data, while the students in the ROOTS condition averaged the

Table 3. Descriptive statistics for mathematics measures by condition and assessment time.

		ROOTS-eligible students								
		T_1		T_2		T_3		Typically achieving students ineligible for ROOTS		
Measure		ROOTS	Control	ROOTS	Control	ROOTS	Control	T_1	T_2	T_3
NSB	*M*	12.58	12.1	19.05	17.99			21.26	25.29	
	(SD)	(3.89)	(3.43)	(4.37)	(5.01)			(5.15)	(4.22)	
	N	203	87	177	79			504	437	
ASPENS	*M*	26.9	23.26	79.64	56.74			84.22	120.32	
	(SD)	(19.45)	(18.05)	(34.12)	(33.27)			(43.19)	(40.93)	
	N	201	87	177	79			502	436	
Oral Counting	*M*	24.32	23.23	46.79	39.67			48.5	66.19	
	(SD)	(15.32)	(15.85)	(23.96)	(19.59)			(25.37)	(23.17)	
	N	199	86	177	79			503	436	
TEMA-3	*M*	17.9	16.99	26.61	23.46			27.84	31.64	
	(SD)	(7.07)	(6.12)	(7.65)	(7.93)			(7.53)	(8.39)	
	N	201	86	179	82			85	11	
RAENS	*M*	12.05	11.29	23.12	17.9			19.74	28.45	
	(SD)	(5.75)	(5.72)	(5.96)	(6.86)			(6.69)	(1.57)	
	N	201	86	179	82			85	11	
SESAT Total	*M*			453.99	447.71				502.27	
	(SD)			(34.71)	(35.43)				(36.65)	
	N			178	78				426	
SAT-10 Total	*M*					492.41	492.89			528.68
	(SD)					(22.09)	(21.98)			(32.08)
	N					132	65			400

Note. The sample sizes represent students with a particular measure at each assessment period. The complete sample included 203 students in the ROOTS condition, 87 in the control condition, and 560 typically achieving students who were not eligible for ROOTS. NSB = Number Sense Brief; ASPENS = Assessing Student Proficiency in Early Number Sense; TEMA-3 = Test of Early Mathematics Ability-Third Edition; RAENS = ROOTS Assessment of Early Numeracy Skills; SESAT = Stanford Early School Achievement Test, SAT-10 = Stanford Achievement Test Series-Tenth Edition.

same score whether they had T_2 data or not. Alternately, students with T_2 data differed between conditions by just 0.3, while ROOTS students without T_2 data scored 3.8 points higher than control students missing T_2. We also found evidence of differential attrition for pretest (T_1) RAENS scores ($t = 2.10$, $df = 230$, $p = .0365$). Students with RAENS data at T_2 scored 4.6 points higher than those without T_2 data among control students; among students in the ROOTS condition, those with T_2 data scored 1.8 points lower than those without T_2 data. Among students with T_2 data, pretest means differed by 0.3 across conditions. For students without T_2 data, the treatment mean was 6.7 points higher than the control mean at pretest. As implied here and shown in Table 3, we have no pretest differences between conditions for students with posttest data. The analyses also incorporated all available data, further reducing the likelihood of bias (Graham, 2009). Nonetheless, results for the NSB total scores and the RAENS should be interpreted with caution.

Efficacy Effects for ROOTS

Table 4 presents the results of the statistical models. The table presents the results of the homoscedastic model if it was deemed equivalent to the more complicated heteroscedastic model (ASPENS, oral counting, TEMA, & RAENS). Otherwise, we provide results for the heteroscedastic model (NSB). The bottom two rows of the table show the likelihood ratio test results that compared homoscedastic residuals to heteroscedastic residuals. Although the variance structures differed between these models, the condition effect estimates and statistical significance values were very similar for both the heteroscedastic and homoscedastic models. We also tested models with an additional level for either classrooms or schools. The overall pattern of results remained very similar in these models as well, so we did not present the results from these models. Overall, the results suggest that the intervention effects were not particularly sensitive to the variance structures.

The models in Table 4 tested fixed effects for differences between conditions at pretest (condition effect), gains across time, and the interaction between the two. We found no statistically significant differences at pretest ($p > .30$ for all measures), which suggested that students were similar in the fall of kindergarten. We found statistically significant differences by condition in gains from fall to spring for four dependent variables. Students in the ROOTS condition made greater gains than control students on the ASPENS ($t = 5.20$, $df = 136$, $p < .0001$), oral counting ($t = 2.14$, $df = 132$, $p = .0333$), TEMA standard scores ($t = 3.35$, $df = 142$, $p = .0010$), and RAENS ($t = 6.84$, $df = 162$, $p < .0001$). We did not detect statistically significant differences between conditions in gains on the NSB nor differences between conditions on the SESAT or SAT-10; both tested with the ASPENS and TEMA as pretest covariates. The time × condition model estimated differences in gains between conditions of 0.75 for the NSB (Hedges's $g = .16$), 19.7 for the ASPENS ($g = .58$), 6.5 for oral counting ($g = .28$), 2.45 for the TEMA standard score ($g = .32$), and 4.7 for the RAENS ($g = .75$).

Closing the Gap

We hypothesized that students provided with ROOTS would make greater gains than students in the same classrooms who did not receive ROOTS because the intervention was designed to close the gap between lower performing students and higher performing students. We tested this question with a second set of partially clustered time × condition

Table 4. Results from a time × condition analysis on fall-to-spring gains in math with a partially nested model to account for intervention students nested within ROOTS groups.

		NSB	ASPENS	Oral counting	TEMA	RAENS
Fixed Effects	Intercept	12.10****	23.26****	23.26****	16.99****	11.29****
		(.46)	(2.87)	(2.87)	(.78)	(.64)
	Time	5.69****	32.87****	32.87****	6.29****	6.48****
		(.42)	(2.94)	(2.94)	(.57)	(.55)
	Condition	.49	3.61	3.61	.90	.76
		(.57)	(3.56)	(3.56)	(.95)	(.78)
	Time × Condition	.75	19.65****	19.65****	2.45**	4.68****
		(.59)	(3.78)	(3.78)	(.73)	(.68)
Variances	Gains Between	2.47*	60.45*	60.45*	2.10*	.97
	ROOTS Groups	(1.11)	(30.03)	(30.03)	(1.11)	(.92)
	Pre–Post Covariance		364.55****	364.55****	38.40****	23.23****
			(49.80)	(49.80)	(3.95)	(2.59)
	Residual		289.62****	289.62****	11.45****	11.55****
			(31.35)	(31.35)	(1.24)	(1.25)
	ROOTS Residual	7.00****				
		(.84)				
	Pre–Post Covariance	6.97****				
		(1.44)				
	Control Residual	4.57**				
		(1.59)				
	Pre–Post Covariance	11.29****				
		(2.41)				
Hedges's *g*	Time × Condition	.164	.580	.281	.317	.749
p values	Time × Condition	.2064	<.0001	.0339	.0010	<.0001
df	Time × Condition	99	136	132	142	162
Likelihood Ratio χ^2		3.24	0.43	2.50	1.23	1.21
p values		.1981	.8069	.2860	.5412	.5466

Note. Table entries show parameter estimates with standard errors in parentheses except for Hedges's *g* values, *p* values, and the degrees of freedom (*df*). Tests of fixed effects (first four rows) accounted for small groups as the unit of analysis within the intervention (ROOTS) condition and unclustered individuals in the control condition. Likelihood ratio test compared homoscedastic residuals to heteroscedastic residuals with a criterion α of .20 and one degree of freedom. NSB = Number Sense Brief; ASPENS = Assessing Student Proficiency in Early Number Sense; TEMA-3 = Test of Early Mathematics Ability-Third Edition; RAENS = ROOTS Assessment of Early Numeracy Skills.
$^{\sim}p < .10, ^{*}p < .05, ^{**}p < .01, ^{***}p < .001, ^{****}p < .0001$.

models that paralleled those used for the efficacy analyses but with students who received ROOTS, clustered in small groups, and students who did not qualify for ROOTS. This essentially amounts to swapping the control sample for the students who performed well enough that they were not eligible for ROOTS. The interpretation of these results, however, no longer benefits from the assumption of equivalence. In addition, the differences between conditions at pretest required tests of net gains because analysis of covariance models can introduce bias with nonequivalent groups (Allison, 1990; Jamieson, 1999; Oakes & Feldman, 2001). Hence, we did not test the SESAT or SAT-10 measures. Furthermore, the TEMA and RAENS were collected only for ROOTS-eligible students, and the NSB did not produce statistically significant differences between conditions, so we also excluded these measures.

To test this hypothesis, we analyzed the ASPENS and oral counting dependent measures, and for both measures, we report results from the heteroscedastic models (LRT $p < .0001$). The gains made by ROOTS intervention students exceeded gains by students who were not eligible to receive ROOTS by 17.6 on the ASPENS ($t = 6.19$, $df = 77$, $p < .0001$) and by 5.4 on the oral counting measure ($t = 2.48$, $df = 97$, $p = .0150$). This represents an effect of $g = .45$ for the ASPENS and $g = .23$ for oral counting. Thus, although students in ROOTS groups made gains of 19.7 over randomly assigned control students who were also eligible

for ROOTS, they also made gains of 17.6 more than more typically achieving students who had not qualified for ROOTS.

Discussion

The purpose of our study was to examine the effectiveness of the ROOTS intervention program. We investigated two questions related to the impact of ROOTS on student achievement. The first question compared treatment students to their at-risk control group peers and the second question examined whether treatment students had reduced the achievement gap with their not-at-risk peers. For the first research question, we found that ROOTS had a significant positive impact on student achievement, greater gain scores, on four out of five outcome measures at kindergarten posttest. ROOTS did not have a significant impact on one measure but still had a positive effect size. Overall findings for the ROOTS intervention would be classified as substantively and positively important based on the What Works Clearinghouse Standards (2011). Similar impacts have been observed for early mathematics intervention programs (e.g. Bryant et al., 2011; Clarke et al., 2016; Dyson et al., 2013; Fuchs et al., 2005; Gersten et al., 2015). Commonalities across intervention curricula suggest a set of emerging themes in the early mathematics intervention research base. Curricula shared a focus on building an understanding of number and number properties and operations and utilized a teacher-directed approach to instruction. All interventions were delivered in small groups and most included 25 to 60 lessons. Impacts on student achievement were in the moderate range. The second research question investigated whether ROOTS students reduced the achievement gap with their not-at-risk peers. On a set of early numeracy measures targeting key aspects of number sense, ROOTS students reduced the achievement gap at kindergarten posttest by making greater gains than their not-at-risk peers. Last, we found that by the middle of first grade (delayed posttest), there were no significant differences between treatment and control students on a distal measure of mathematics achievement.

Limitations and Future Research

Interpreting results from any study conducted with a specific and limited sample should be done with caution. The demographics of the students and teachers in the current sample are not reflective of a national sample. To that end, future research should emphasize additional studies of the ROOTS intervention in different geographical areas and with different demographic samples to increase the confidence that results found are not unique to the current study's sample. We currently have additional studies planned of the ROOTS intervention with multiple sites and cohorts of students. There is a growing recognition of the importance in educational research to engage in and value replication studies, and future research planned will address this critical stage of research (Cook, 2014).

The study utilized a partially nested design, and it should be noted that partially nested trials have weakened internal validity compared to fully clustered or unclustered RCTs. The clustered subjects in the intervention condition and unclustered subjects in the control condition do not necessarily represent the potential outcomes for each other solely in terms of the intervention (Bauer et al., 2008). Although we have neither a theoretical rationale nor empirical evidence that simply clustering kindergarten students in small groups would lead to improved math outcomes, we cannot rule out such an effect because the difference in

clustering across conditions has been confounded with the intervention delivery. On the other hand, the external validity may be stronger. Ungrouped control students represent the most appropriate contrast condition because they reflect the experience of students at risk for math difficulties in the many schools that do not provide tiered math instruction in kindergarten.

The measurement net utilized in the study was designed to provide coverage of both proximal and distal outcomes. However, two of the distal measures, Stanford Achievement Tests, included coverage of items not taught in the ROOTS intervention program related to measurement and geometry. The lack of alignment may have contributed to nonsignificant impact findings on those measures. It should also be noted that in examining whether ROOTS students reduced the gap with their not-at-risk peers, the outcome measure used was fluency based, and not-at-risk students may have reached a ceiling of performance that effectively may have made it appear as if ROOTS students were reducing the achievement gap.

An ongoing concern in educational research (Starkey & Klein, 2008) and a pattern found in the results reported here is the limited impact on long-term student achievement. Coupled with increased concern regarding nonresponders to research-based standard Tier 2 interventions, there is interest in the field in exploring how to make treatments more effective for all students and for long-term sustained impacts on achievement. Miller, Vaughn, and Freund (2014) provide an overview of avenues that offer promise for exploring, including examining the intensity of treatment, exploring the role of learner characteristics including executive functioning skills, and developing greater precision in utilizing screening to identify potential nonresponders.

The findings in this study are from the first-year first cohort of students of a four-year project funded by the Institute of Education Sciences (Clarke, Doabler, et al., 2012). Due to limited power, we focused on examining only one of our primary research questions, the impact of ROOTS on student achievement. However, a feature of the research is the use of differing ROOTS group sizes to examine the role of treatment, or instructional, intensity. As part of our rationale for examining the treatment intensity of ROOTS, we built on a theoretical framework specified by Warren, Fey, and Yoder (2007) that considers treatment intensity as "a general variable that may be a key to optimizing intervention effects" (p. 70). Theoretically, small-group math instruction offers a more effective and more intensive method for engaging at-risk students in instructional interactions around important math content. A growing body of research in both reading (Connor et al., 2009; Smolkowski & Gunn, 2012) and mathematics (Clements, Agodini, & Harris, 2013; Doabler et al., 2015; Morgan, Farkas, & Maczuga, 2015) has begun to illustrate the importance of frequent and high-quality instructional interactions that center on critical content. Although we do not report on findings in this article due to insufficient power, by systematically manipulating group size as part of this multiyear program of research, we will have an opportunity to investigate the role of instructional intensity at a molecular level and begin to explore and generate findings related to the relationship between instructional intensity and student achievement.

We see work examining instructional intensity as fitting within the vein of research that extends what we know about intervention programs and service delivery within tiered systems of support. Work in the area of reading has already begun to examine

variations of tiered systems of support. For example, Al Otaiba, Kim, Wanzek, Petscher, and Wagner (2014) examined the impact on achievement of students identified as at risk, who were immediately provided more intensive instructional services, in comparison to their peers, who were assigned to a traditional model where they first received Tier 1 services. We view as critical investigations that also mirror the delivery of services both within a school year with students moving across tiers based on responsiveness (e.g., O' Conner, 2000; O'Conner, Harty, & Fulmer, 2005), linked tiers of service delivery (e.g., Fien et al., 2015; Fuchs, Fuchs, & Vaughn, 2008), and examining the impact of multiple years of intervention services (e.g. Vaughn et al., 2014) because such research offers insights into how schools may structure intervention services to maximize impact within the constraints of finite resources.

Conclusion

The field of early mathematics research has begun to generate evidence on effective Tier 2 intervention programs, and our understanding of what constitutes standard elements of an effective intervention program—focused whole-number content and an explicit and systematic teaching approach—has grown. We consider it logical that the next steps include focused instructional interactions and intervention programs that fit within tiered-service delivery models in order to advance the field and ensure that all students are provided with effective mathematics instruction.

Conflict of Interest

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Scott Baker, Ben Clarke, and Hank Fien are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain ROOTS-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services. An independent external evaluator and coauthor of this publication completed the research analysis described in the article.

Funding

This research was supported by the ROOTS Project, Grant No. R324A120304, funded by the U.S. 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 U.S. Department of Education.

ARTICLE HISTORY

Received 1 May 2015
Revised 29 September 2015
Accepted 19 October 2015

EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

References

Abelson, R. P. (2012). *Statistics as principled argument*. New York, NY: Taylor and Francis.

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness*, *3*, 199–253. doi:10.1080/19345741003770693

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

Al Otaiba, S., Kim, Y.-S., Wanzek, J., Petscher, Y., & Wagner, R. K. (2014). Long-term effects of first-grade multitier intervention. *Journal of Research on Educational Effectiveness*, *7*, 250–267. doi:10.1080/19345747.2014.906692

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford Press.

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.

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

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

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, *38*, 333–339. doi:10.1177/00222194050380040901

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *Elementary School Journal*, *108*, 115–130. doi:10.1086/525550

Bryant, D. P., Bryant, B., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. M. (2011). Early numeracy intervention program for first-grade students with mathematics difficulties. *Exceptional Children*, *78*, 7–23. doi:10.1177/001440291107800101

Burns, M. K., & Vanderheyden, A. M. (2006). Response to intervention to assess learning disabilities: Introduction to the special series. *Assessment for Effective Instruction, 32*(3), 3–5.

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).

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., Fien, H., & Baker, S. K. (2014, September). *Testing the immediate and long-term efficacy of a Tier 2 kindergarten mathematics intervention*. Paper presented at the Annual Fall Society for Research on Educational Effectiveness Conference, Washington, DC. Retrieved from https://www.sree.org/conferences/2014f/program/downloads/abstracts/1287_2.pdf

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., Smolkowski, K., Baker, S. K., Fien, H., Doabler, C. T., & Chard, D. (2011). The impact of a comprehensive Tier 1 core kindergarten program on the achievement of students at risk in mathematics. *The Elementary School Journal*, *111*, 561–584. doi:10.1086/659033

Clements, D. H., Agodini, R., & Harris, B. (2013). *Instructional practices and student math achievement: Correlations from a study of math curricula* (NCEE Evaluation Brief No. 2013-4020). Retrieved from http://ies.ed.gov/ncee/pubs/20134020/pdf/20134020.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Common Core State Standards Initiative. (2010a). *Common core standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/wp-content/uploads/ELA_Standards.pdf

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

Connor, C. M., Morrison, F. J., Fishman, B., Ponitz, C., Glasney, S., Underwood, P., . . . Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*, 85–99. doi:10.3102/0013189X09332373

Cook, B. G. (2014). A call for examining replication and bias in special education research. *Remedial and Special Education*, *35*, 233–246. doi:10.1177/0741932514528995

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., 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*, 303–333. doi:10.1086/679969

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., & 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., Fien, H., Nelson, N. J., & Baker, S. K. (2012). Evaluating three elementary mathematics programs for presence of eight research-based instructional design principles. *Learning Disability Quarterly*, *35*, 200–211. doi:10.1177/0731948712438557

Doabler, C. T., Strand Cary, M., Jungjohann, K., Clarke, B., Fien, H., Baker, S. K., & Chard, D. J. (2012). Enhancing core mathematics instruction for students at risk for mathematics disabilities. *Teaching Exceptional Children*, *44*(4), 48–57. doi:10.1177/004005991204400405

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

Feng, Z., Diehr, P., Peterson, A., & McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health*, *22*, 167–189. doi:10.1146/annurev.publhealth.22.1.167

Fien, H., Smith, J. L. M., Smolkowski, K., Baker, S. K., Nelson, N. J., & Chaparro, E. A. (2015). An examination of the efficacy of a multitiered intervention on early reading outcomes for first grade students at risk for reading difficulties. *Journal of Learning Disabilities*, *48*, 602–621. doi:10.1177/0022219414521664

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, D., Fuchs, L. S., & Vaughn, S. (2008). *Response to intervention: A framework for reading educators*. Newark, DE: IRA.

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). Response to intervention: A strategy for the prevention and identification of learning disabilities. In E. L. Grigorenko (Ed.), *Educating individuals with disabilities: IDEIA 2004 and beyond* (pp. 115–135). New York, NY: Springer.

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

Gersten, R. M., Baker, S. K., Shanahan, T., Linan-Thompson, S., Collins, P., & Scarcella, R. (2007). *Effective literacy and English language instruction for English learners in the elementary grades* (Practice Guide No. NCEE 2007-4011). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/20074011.pdf

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). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/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. doi:10.1177/002246699903300102

Gersten, R. M., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*, 1202–1242. doi:10.3102/0034654309334431

Gersten, R. M., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*, 516–546. doi:10.3102/0002831214565787

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

Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology*, *93*, 615–627. doi:10.1037/0022-0663.93.3.615

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-Tenth edition [SAT-10]*. San Antonio, TX: Author.

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

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

Hox, J. J., & Maas, C. J. M. (2002). Sample sizes for multilevel modeling. In J. Blasius, J. J. Hox, E. de Leeuw, & P. Schimdt (Eds.), *Social science methodology in the new millennium. Proceedings of the fifth international conference on logic and methodology* (Second expanded ed., pp. 1–18). Cologne, Germany: Opladen, RG: Leske + Budrich Verlag (CD-ROM).

Individuals with Disabilities Education Act of 2004, public law 108-446 subpart 614(6)(b) (2004).

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., Kaplan, D., Olah, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at-risk for mathematics difficulities. *Child Development*, *77*, 153–177. doi:10.1111/j.1467-8624.2006.00862.x

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness*, *1*, 155–178. doi:10.1080/19345740802114533

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling* (1st ed.). London, England: Sage Publications.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial & Special Education*, *24*, 97–114. doi:10.1177/07419325030240020501

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

La Paro, K. M., Hamre, B. K., Locasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D., . . . Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development*, *20*, 657–692. doi:10.1080/10409280802541965

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

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis, 46*(3), 427–440. doi:10.1016/j.csda.2003.08.006

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., Hillemeier, M. M., & Maczuga, S. (2014). Who is at risk for persistent mathematics difficulties in the United States? *Journal of Learning Disabilities*. Advanced online publication. doi:10.1177/0022219414553849

Morgan, P. L., Farkas, G., & Maczuga, S. (2015). Which instructional practices most help first-grade students with and without mathematics difficulties? *Educational Evaluation and Policy Analysis*, *37*, 184–205. doi:10.3102/0162373714536608

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

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

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

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? *Evaluation Review*, *20*, 313–337. doi:10.1177/0193841X9602000305

National Association of State Directors of Special Education. (2006). *Response to intervention: Policy considerations and implementation*. Retrieved from http://www.nasdse.org/publications-t577/response-to-intervention-policy-considerations-an.aspx

National Center for Education Statistics. (2013). *The nation's record card: A first look: 2013 mathematics and reading* (Report No. NCES 2014-451). Retrieved from http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014451.pdf

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. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Retrieved from http://www.nap.edu/catalog.php?record_id=12519

National Science Board. (2008). *Science and engineering indicators 2008* (Report No. NSB 08-01A). Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/statistics/seind08/pdf/volume2.pdf

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest–posttest designs: The impact of change-score versus ancova models. *Evaluation Review*, *25*, 3–28. doi:10.1177/0193841x0102500101

O'Connor, R. E. (2000). Increasing the intensity of intervention in kindergarten and first grade. *Learning Disabilities Research and Practice*, *15*, 43–54. doi:10.1207/SLDRP1501_5

O'Connor, R. E., Harty, K. R., & Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities*, *38*, 532–538. doi:10.1177/00222194050380060901

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., & Evans, S. J. W. (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

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*, 138–154. doi:10.1080/19345740801982104
Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*, 346–362. doi:10.1037/0022-0663.93.2.346
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
Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). Boston, MA: McGraw-Hill.
SAS Institute. (2009). *SAS/STAT ®9.2 user's guide*. Cary, NC: Author.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
Smolkowski, K., & Cummings, K. D. (2016). Evaluation of the DIBELS (sixth edition) diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. *Journal of Psychoeducational Assessment*, *34*, 103–118. doi:10.1177/0734282915589017
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. doi:10.1016/j.ecresq.2011.09.004
Sood, S., & Jitendra, A. K. (2007). A comparative analysis of number sense instruction in reform-based and traditional mathematics textbooks. *Journal of Special Education*, *41*, 145–157. doi:10.1177/00224669070410030101
Starkey, P., & Klein, A. (2008). Sociocultural influences on young children's mathematical knowledge. In O. N. Saracho & B. Spodek (Eds.), *Contemporary perspectives on mathematics in early childhood education* (pp. 253–276). Charlotte, NC: Information Age Publishing.
van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). Hoboken, NJ: Jon Wiley & Sons.
Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice*, *18*, 137–146. doi:10.1111/1540-5826.00070
Vaughn, S., Roberts, G., Wexler, J., Vaughn, M. G., Fall, A.-M., & Schnakenberg, J. B. (2014). High school students with reading comprehension difficulties: Results of a randomized control trial of a two-year reading intervention. *Journal of Learning Disabilities*. Advance online publication. doi:10.1177/0022219413515511
Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 70–77. doi:10.1002/mrdd.20139
What Works Clearinghouse. (2011). *Procedures and standards handbook, Version 2.1*. Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx
What Works Clearinghouse. (2014). *Procedures and standards handbook, Version 3.0*. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19
Wu, H.-H. (1999). Basic skills versus conceptual understanding. *American Educator, 23*(3), 14–19.

Appendix

Treatment Effect Heterogeneity

Approach

Tests of treatment effect variability have recently become of interest in educational research, so we explored whether the ROOTS effects varied by classroom or school. To test the condition-effect variation, we fit models that included school-level random effects for the intercept, time, condition, and their interaction (time × condition). The models required all

three to avoid including random school- or classroom-level variation associated with general student growth or assignment to condition within the estimate for cluster-level treatment-effect variability.

In our exploratory analyses, some models produced negative variance estimates. Negative variances often result from computational limitations for estimates very near zero and do not represent problems with the models or analysis procedures (Kreft & de Leeuw, 1998; Singer & Willett, 2003). In our experience, they are more likely to occur when estimating variances for a small number of units because variance estimation in such situations is poor. Forcing nonnegative estimates can lead to depressed Type I error rates and reduced statistical power (Murray, 1998; Murray, Hannan, & Baker, 1996). In some cases, complex models produce sufficiently negative or unstable variance estimates that these parameters must be constrained to at or above zero. In the present analyses, this was the case for the ASPENS, oral counting, and RAENS models.

Results

We report the treatment effect variation in terms of total variance of the dependent variable—the total variance of the pretest and posttest assessments. To do so, we standardized each dependent variable and provide both the proportion of variance unexplained by the fixed effects (e.g., time, condition, interactions) and the proportion of variability explained by the school-level variance of the condition effect. We also provide the p value provided by the Wald test for covariance parameters in SAS PROC MIXED.

For the NSB, fixed effects in Table 4 accounted for 37.6% of the overall variability, leaving 62.4% of the variability unexplained by time, condition, and their interaction. The school-level variability of the treatment estimator (time × condition) represented −0.4% of the total variability, 95% CI [−3.0%, 2.2%], $p = .7526$. The classroom-level variability of the intervention estimate accounted for −0.9% of the total variability, 95% CI [−3.1%, 1.4%], $p = .4510$.

Fixed effects in the analysis of TEMA scores accounted for 25.3% of the total variability, leaving 74.7% unexplained. Of the total variance for the TEMA, the school-level variability of the treatment effect accounted for 2.5% of the total TEMA variance, 95% CI [−1.7%, 6.8%], $p = .2461$. The classroom-level variance of the intervention effect accounted for 3.1% of the total TEMA variance, 95% CI [−1.4%, 7.7%], $p = 1718$.

As noted above, the analyses of school- and classroom-level variability of the intervention effect required that variances be constrained to zero or greater for the ASPENS, oral counting, and RAENS. In all but two cases, the variability of the intervention effect was fixed at zero, so no test of the variance was provided. The classroom-level variability of the intervention effect for Oral Counting was estimated at $6.56\% \times 10^{-39}$, but it was not possible to conduct a statistical test of the variance. For the RAENS, the school-level variability of the intervention effect was estimated at 0.1%, 95% CI [0.0%, $9.84\% \times 10^{72}$], $p = .4424$.

Limitations

Interpretation of these variance estimates requires several notes of caution. First, we had no theoretical reason to expect school- or classroom-level variability among the intervention effects. Conversely, because any measurable quantity collected in school settings will vary to some degree, intervention effects will always vary by school and classroom. Whether that variability achieves statistical significance, however, depends on a number of factors, such as

the size of the variance, sample sizes at each level, various types of error (e.g., measurement, sampling), model complexity, and so on.

Second, hypothesis tests are designed to identify systematic factors—a signal—among the total (random plus systematic) variability—the signal plus noise (Abelson, 2012; Smolkowski & Cummings, 2016). The tests of treatment effect variability determine whether variability exceeds an arbitrary threshold that produces $p < .05$ under an untenable assumption that treatment effects have zero variance among schools and classrooms. As noted above, all measurable quantities vary to some degree in educational settings. The hypothesis tests therefore assume an indefensible null hypothesis of no variability. Furthermore, they offer no information about whether the variability may be systematic or random. We therefore prefer tests of moderation that attempt to identify systematic variation associated with specific background characteristics (Abelson, 2012).

Third, variance estimates have been found to be imprecise in samples with few cases. This is why the degrees of freedom for cluster-level fixed effects must be reduced in multilevel models (Feng, Diehr, Peterson, & McLerran, 2001). With fewer than 100 clusters, standard errors are usually deflated, and variance components fail to approach nominal levels (Hox & Maas, 2002; Maas & Hox, 2004). The present analysis included just 14 schools and 37 classrooms. It is unlikely that the school- or classroom-level variances achieved in the present study will be replicable or, for that matter, come close to the variability expected among the population of schools and classrooms that might use the ROOTS intervention.