

Clarke, B., Doabler, C. T., Kosty, D., Kurtz Nelson, E., Smolkowski, K., Fien, H., & Turtura, J. (2017). Testing the efficacy of a kindergarten mathematics intervention by small group size. *AERA open*, 3(2), 1–16. doi: 10.1177/2332858417706899

Grant/Contract Number: R324A120304

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.

Testing the Efficacy of a Kindergarten Mathematics Intervention by Small Group Size

Ben Clarke

University of Oregon

Christian T. Doabler

University of Texas at Austin

Derek Kosty

Oregon Research Institute

Evangeline Kurtz Nelson

University of Oregon

Keith Smolkowski

Oregon Research Institute

Hank Fien

Jessica Turtura

University of Oregon

This study used a randomized controlled trial design to investigate the ROOTS curriculum, a 50-lesson kindergarten mathematics intervention. Ten ROOTS-eligible students per classroom ($n = 60$) were randomly assigned to one of three conditions: a ROOTS five-student group, a ROOTS two-student group, and a no-treatment control group. Two primary research questions were investigated as part of this study: What was the overall impact of the treatment (the ROOTS intervention) as compared with the control (business as usual)? Was there a differential impact on student outcomes between the two treatment conditions (two- vs. five-student group)? Initial analyses for the first research question indicated a significant impact on three outcomes and positive but nonsignificant impacts on three additional measures. Results for the second research question, comparing the two- and five-student groups, indicated negligible and nonsignificant differences. Implications for practice are discussed.

Keywords: *response to intervention, mathematics, at risk, number sense*

THE importance of a successful start in mathematics is garnering increased attention (Frye et al., 2013), supported by a growing research base highlighting the critical role that early mathematics plays in the development of mathematical thinking and long-term success in mathematics. Analyses of the Early Childhood Longitudinal Study data set indicate that poor performance at school entry is strongly related to poor performance in later grades (Morgan, Farkas, & Wu, 2009) and that the relationship between early and later mathematics achievement is stronger than the relationship between early and later literacy skills (Duncan et al., 2007). In addition, Morgan, Farkas, Hillemeier, and Maczuga (2016) found that students who exited kindergarten with low mathematics achievement experienced persistent difficulty in mathematics through later elementary and middle school at a factor of 17 times that of their not-at-risk peers and that

early mathematics achievement was a significantly stronger predictor than a number of other variables (e.g., cognitive variables) associated with mathematics achievement. Unless such difficulties are addressed early, long-term persistent struggles with mathematics are likely to occur and be increasingly resistant to change (Geary, 1993; Jordan, Kaplan, & Hanich, 2002).

The rising concern over early mathematics and its role in long-term development occurs as greater expectations around mathematics are codified, through initiatives such as the Common Core State Standards (CCSS; CCSS Initiative, 2010), and used to guide service delivery in schools. However, while expectations are increasing, national achievement data continue to show consistent and worrisome patterns of performance. Data from the 2015 National Assessment of Educational Progress indicate that relatively



few students are likely to meet these new higher standards, with only 40% of students being classified as at or above proficiency and 18% as below basic. Data are even more concerning for minority students, low socioeconomic status students, and students with disabilities, as relatively few (16%–26%) are classified as at or above proficiency. In addition, National Assessment of Educational Progress data show that after positive gains from 1990 through 2007, fourth-grade achievement levels have largely stagnated.

Approaches to Address Low Mathematics Achievement

Solutions to address systematic low achievement in mathematics are confounded by the challenge that schools face in providing additional time for mathematics instruction at the early elementary grades. Data indicate that a relatively small amount of time during the school day is allocated to mathematics instruction (La Paro et al., 2009) and that time is likely spent toward core or Tier 1 instruction without added support for at-risk students. In addition, schools are not likely to have developed the same institutionalized supports around early mathematics as are in place for beginning reading instruction. Such supports include mandated time blocks of instruction, research-validated or research-based materials for core and intervention instruction, screening systems to identify at-risk students, progress-monitoring systems to measure growth over time, professional development on best practices, and implementation support provided by reading specialists or coaches (Balu et al., 2015; Clarke, Baker, & Chard, 2008).

The aforementioned supports are components of a response to intervention (RTI) service delivery model. Initially conceptualized and operationalized as an alternate mechanism to determine eligibility for special education services (Individuals with Disabilities Education Act of 2004) by differentiating whether a student's lack of growth was due to poor instruction or disability (Fuchs, Fuchs, & Hollenbeck, 2007), aspects of RTI models have been adopted in general to support the academic and behavioral growth of all students (Fuchs & Vaughn, 2012; Vaughn & Swanson, 2015) through a multitier system of support (MTSS) service delivery model. Although multitier models vary, most have three tiers of support, with Tier 1 consisting of core instruction, Tier 2 including the use of standard protocol interventions typically delivered in small groups, and Tier 3 focused on individualized problem solving and more intensive degrees of support (e.g., one-on-one tutoring). MTSS service delivery models are increasingly common in the area of reading, with fewer applications in the area of mathematics. A self-reported survey of mathematics practices detailed that only one-third of schools provided multitier systems of support in first grade. In contrast from the same grade, 71% of schools reported providing full implementation of MTSS models in the area of reading (Balu et al., 2015).

A research base on early mathematics practices for use in MTSS in the early elementary grades is emerging and includes key elements, such as screening for mathematics difficulty in kindergarten (Fuchs et al., 2010; Smolkowski & Gunn, 2012) and a number of research studies investigating the efficacy of kindergarten intervention programs (Clarke, Doabler, Smolkowski, Kurtz Nelson, et al., 2016; Dyson, Jordan, & Glutting, 2013; Fuchs et al., 2005; Sood & Jitendra, 2013). A number of key elements exist across these intervention programs. As called for by experts, the focus of each program is on the development of number sense (Berch, 2005; Gersten & Chard, 1999) and whole number understanding (Clarke, Baker, & Fien, 2009; National Mathematics Advisory Panel, 2008). In addition, each program employs a systematic and explicit instructional framework (Archer & Hughes, 2011; Coyne, Kame'enui, & Carnine, 2011) that includes instructional design elements and features shown to be effective for at-risk learners (Baker, Gersten, & Lee, 2002; Gersten et al., 2009; Kroesbergen & Van Luit, 2003).

Despite broader advances made by the field in designing interventions based on these key elements and rigorous studies of their efficacy, there have been calls to examine more finite questions related to intervention effectiveness (Gersten, 2016; Ochsendorf, 2016) to help refine the implementation of multitier models to better meet the needs of all learners (Miller, Vaughn, & Freund, 2014). One potential mechanism for such investigations is examining the treatment of instructional intensity of intervention services.

Treatment Intensity and Small Group Instruction

Warren, Fey, and Yoder (2007) theorized that treatment intensity functioned as a generalized variable that offered a key mechanism by which to optimize intervention effects, and they specified a framework for quantifying treatment intensity of interventions by examining variables related to dose (i.e., teaching episode), dose form, dose frequency, and duration, resulting in a metric of cumulative intervention intensity. A small cluster of studies in the area of print awareness have examined finite manipulations of treatment intensity to more fully understand intervention impacts (e.g., Breit-Smith, Justice, McGinty, & Kaderavek, 2009; Ezell, Justice, & Parsons, 2000; Justice & Ezell, 2000, 2002; Justice, Kaderavek, Fan, Sofka, & Hunt, 2009; Justice, McGinty, Piasta, Kaderavek, & Fan, 2010; Lovelace & Stewart, 2007; McGinty, Breit-Smith, Fan, Justice, & Kaderavek, 2011). However, despite calls for such investigations, studies that isolate variables of treatment intensity are still relatively limited (Coddling & Lane, 2015) in other academic areas, including mathematics. Given the research base on systematic and explicit instruction in mathematics as a cornerstone of programs designed for at-risk students (Baker et al., 2002; Gersten et al., 2009; Kroesbergen & Van Luit, 2003), the behaviors associated with this approach

(e.g., teacher models, practice opportunities, feedback) are logical mechanisms to manipulate to increase treatment intensity.

A number of studies have begun to look at the impact of small groups as a proxy for treatment intensity. Small group instruction is considered crucial to the provision of interventions because it represents a mechanism for individualizing and intensifying instruction (Fien et al., 2011; Gersten et al., 2008), and the use of smaller groups is generally accepted practice within MTSS as a mechanism to increase intensity within and across tiers of instruction (Baker, Fien, & Baker, 2010; Denton et al., 2013).

Meta-analyses of studies on reading interventions show positive effects for small group instruction. Wanzek and Vaughn (2007) conducted a meta-analysis examining a range of variables for a K–3 reading intervention including group size ranging from one on one to small groups of eight students. Results indicated that smaller instructional groups were associated with greater effect sizes. In addition, Elbaum, Vaughn, Hughes, and Moody (2000) found significant positive effects for 1:1 reading interventions for at-risk elementary students. While such comparisons investigate the overall efficacy of small group instruction or 1:1 instruction, additional work has examined contrasting small group composition. Research indicates that variations in small group size are associated with key academic variables, including academic engaged time (Thurlow, Ysseldyke, Wotruba, & Algozzine, 1993). Vaughn, Thompson, Kouzekanani, and Dickson (2003) conducted a study in which three small group sizes were contrasted, holding all other variables constant (i.e., each group size used the same intervention program), and they found a significant impact for smaller group sizes (1:1 and 1:3) at posttest and follow-up when compared with a larger group (1:10) but no significant differences between 1:1 and 1:3. A study by Vaughn, Cirino, et al. (2010) also investigated manipulating group size (1:5 and 1:12–15) and found positive but nonsignificant results on reading outcomes for seventh and eighth students.

Within mathematics, relatively few investigations of varying group size have been conducted. B. R. Bryant et al. (2016) examined the impact of a Tier 3 intervention that was developed by systematically modifying a Tier 2 intervention to provide a more intensive instructional experience. Due to modifying multiple aspects of the intervention (e.g., dose, instructional design features) with group size, attributing cause to any single variable was not possible. However, results were generally effective, with 75% of participants showing performance after the intervention that would no longer indicate the need Tier 3 services (i.e., >25th percentile on a distal measure). Such results thus indicate some promise for considering group size as a variable by which to increase intervention intensity in mathematics. Note that their work was across two studies and not a direct investigation of group size as an independent variable within a single

study and that the results did not include an analysis of whether students maintained their gains (i.e., remained >25th percentile) at later time points. Despite mixed results related to group size and student outcomes, the interest in mechanisms to increase treatment intensity (Coddling & Lane, 2015) within multitier systems of service delivery suggests that continued investigation into the role of group size and student outcomes is warranted.

Purpose and Research Questions

To date, we have found no studies in mathematics that manipulated small group size. The purpose of our research was to conduct an investigation of an early mathematics intervention delivered in different group size formats. The study used a randomized controlled trial design (blocking on classrooms) to investigate the ROOTS intervention in 69 kindergarten classrooms with approximately 10 eligible students per classroom. ROOTS is a 50-lesson Tier 2 kindergarten intervention curriculum. The goal of ROOTS is to support students' conceptual understanding of and procedural fluency with critical whole number concepts. ROOTS is fully aligned to the kindergarten CCSS in the area of number and operations (CCSS Initiative, 2010). The research team randomly assigned these 10 students to one of three conditions (student:teacher ratio): a ROOTS large group (5:1), a ROOTS small group (2:1), and a no-treatment control group. Overall efficacy of the ROOTS intervention has been investigated (Clarke, Doabler, Smolkowski, Baker, et al., 2016; Clarke, Doabler, Smolkowski, Kurtz Nelson, et al., 2016) and found to positively affect student math outcomes. Two primary and one secondary research question were investigated as part of this study:

Research Question 1: What was the overall impact of the treatment (ROOTS intervention) as compared with the control (business as usual)?

Research Question 2: Was there a differential impact on teacher and student behaviors between the two treatment conditions (ROOTS large group vs. ROOTS small group)?

Research Question 3: Was there a differential impact on student outcomes between the two treatment conditions (ROOTS large group vs. ROOTS small group)?

We hypothesized that ROOTS would have a positive impact on student achievement when compared with the control condition. In addition, we hypothesized that there would be a significant difference in treatment intensity between ROOTS small group and ROOTS large group as measured by critical teacher and student behavior, and that there would be a differential impact on student outcomes for the ROOTS small group compared to the ROOTS large group.

Method

Participants

Schools. Fourteen elementary schools from four Oregon school districts participated in the present study. Three were located in rural and suburban areas of western Oregon and one in the Portland metropolitan area. Student enrollment ranged from 2,736 to 39,002. Schools targeted for recruitment received Title I funding. Within these 14 schools, 0%–12% of students were American Indian or Native Alaskan; 0%–16%, Asian; 0%–9%, Black; 0%–74%, Hispanic; 0%–2%, Native Hawaiian or Pacific Islander; 19%–92%, White; and 0%–15%, more than one race. Within these same schools, 8%–25% of students received special education services; 5%–69%, English language learners; and 17%–87%, eligible for free or reduced-price lunch. School and district enrollment and demographics did not change significantly from Year 1 to Year 2.

Classrooms. Sixty-nine classrooms participated in the study ($n = 37$ classrooms in Year 1, $n = 32$ classrooms in Year 2). Each year represents a separate sample. In this study, the samples from each year were combined. Of the 69 classrooms, 63 offered a half-day kindergarten program, and 6 offered a full-day program. All classrooms provided mathematics instruction in English and operated 5 days per week. Across both years of the study, classrooms had an average of 25.06 students ($SD = 5.60$).

The 69 classrooms were taught by 31 teachers. Of the 31 teachers, 20 participated in both years of the study. Nine Year 1 teachers and seven Year 2 teachers taught two participating half-day classrooms (a.m. and p.m.). All were certified kindergarten teachers and participated for the full duration of the ROOTS study. Of the 31 teachers, 100% identified as female, 84% as White, and 10% as Asian American/Pacific Islander. One teacher identified as representing another ethnic group, and one teacher declined to provide ethnicity information. Teachers had an average of 16.45 years of teaching experience and 8.81 years of kindergarten teaching experience; 87% of teachers had a master's degree in education; and 68% of teachers had completed an algebra course at the college level.

Criteria for participation. In each participating classroom, all students with parental consent were screened in the late fall of their kindergarten year. The screening process included the Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke, Gersten, Dimino, & Rolhus, 2011) and the Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2008), which are standardized measures of early mathematics proficiency. Students were eligible for the ROOTS intervention and thus considered at risk for mathematics difficulties if they received an NSB score ≤ 20

and an ASPENS composite score in the *strategic* or *intensive* ranges.

Once students were determined eligible for the ROOTS intervention, the project's independent evaluator separately converted students' NSB and ASPENS scores into standard scores and then combined the two standard scores to form an overall composite score for each student. Composite scores within each classroom were then rank ordered, and the 10 lowest ROOTS-eligible students were randomly assigned to a two-student ROOTS intervention group (2:1), a five-student ROOTS intervention group (5:1), or a no-treatment control condition. Out of the 69 participating classrooms, 53 had at least 10 students who met ROOTS eligibility criteria. Fourteen classrooms in Year 1 and two classrooms in Year 2 had <10 ROOTS-eligible students, and in these instances classrooms were combined to create virtual ROOTS "classrooms." The cross-class grouping procedure was applied seven times, with five sets of two classrooms combined to create five ROOTS classrooms and two sets of three classrooms combined to make two ROOTS classrooms. After these procedures were applied, a total of 60 ROOTS classrooms participated in this study.

Students. A total of 1,550 kindergarten students were screened for ROOTS eligibility. Of these students, 592 met eligibility criteria and were randomly assigned within each of the 60 classrooms to the two-student group condition ($n = 120$), the five-student group condition ($n = 295$), or the no-treatment control condition ($n = 177$). Student demographic information for all ROOTS eligible students is presented in Table 1.

Interventionists. ROOTS intervention groups were taught by district-employed instructional assistants and by interventionists hired specifically for this study. Among the interventionists, 89% identified as female, 93% as White, 4% as Hispanic, and 2% as another ethnicity. Most interventionists had previous experience providing small group instruction (93%) and had a bachelor's degree or higher (58%). Interventionists had an average of 8 years of teaching experience; 20% had a current teaching license; and 63% had taken an algebra course at the college level.

Procedures

Intervention. ROOTS is a Tier 2 kindergarten program that consists of 50 lessons designed to build students' whole number proficiency. The ROOTS intervention was delivered in 20-min small group sessions (two or five students) 5 days per week for approximately 10 weeks. For all students, instruction began in late fall and ended in the spring. The late fall start date was selected to provide students with

TABLE 1
Descriptive Statistics for Student Characteristics by Condition

Student characteristic	ROOTS, % ^a		Control, % ^a
	Two-student group	Five-student group	
Age at pretest, <i>M</i> (<i>SD</i>)	5.2 (0.4)	5.3 (0.4)	5.2 (0.4)
Male	50	53	49
Race			
American Indian/ Alaskan Native	3	3	3
Asian	3	3	4
Black	4	3	4
Native Hawaiian/ Pacific Islander	1	1	0
White	60	52	54
More than one race	2	4	1
Hispanic	27	25	28
Limited English proficiency	28	23	26
SPED eligible	11	10	9

Note. The sample included 120 students in the two-student ROOTS group condition and 295 students in the five-student ROOTS group condition, and 177 students in the control condition. SPED = special education.

^aValues are presented as percentages unless noted otherwise.

opportunities to respond to core mathematics instruction and therefore minimize the identification of typically achieving students during the screening process. To ensure that ROOTS students received ROOTS instruction and core mathematics instruction, ROOTS occurred at times that did not conflict with core whole-class mathematics instruction.

ROOTS instruction is aligned with CCSS for mathematics (CCSS Initiative, 2010) and recommendations from expert panels to focus intensively on whole number concepts and skills (Gersten et al., 2009). Specifically, ROOTS instruction emphasizes concepts from the Counting & Cardinality and Operations & Algebraic Thinking domains of the CCSS for mathematics to promote robust whole number sense for struggling students. The ROOTS instructional approach is drawn from principles of explicit and systematic mathematics instruction (Coyne et al., 2011; Gersten et al., 2009). In this way, lessons include explicit teacher modeling, deliberate practice, visual representations of mathematics, and academic feedback. ROOTS also provides frequent opportunities for students to verbalize their mathematical thinking and discuss problem-solving methods. For more information on ROOTS, see Clarke, Doabler, Smolkowski, Kurtz Nelson, et al. (2016).

Professional development. All interventionists participated in two 5-hr professional development workshops delivered

by project staff. The first workshop focused on the instructional objectives and content of Lessons 1–25, whole number concepts and skills, empirically validated instructional practices in mathematics, and small group management techniques. The second workshop focused on the mathematics content emphasized in Lessons 26–50. Workshops provided opportunities for interventionists to practice and receive feedback on lesson delivery from instructional coaches and project staff. To promote implementation fidelity and enhance the quality of instruction, all interventionists received between two and four coaching visits from ROOTS coaches during intervention implementation. ROOTS coaches were former educators with specialized knowledge and training in the science of early mathematics instruction and effective small group instructional practices. Coaching visits consisted of direct observations of lesson delivery, followed by feedback on instructional quality and fidelity of intervention implementation.

Control condition. Core (Tier 1) mathematics instruction delivered in the kindergarten classroom served as the control condition or counterfactual in this study, as all participating treatment and control students received daily core mathematics instruction. For treatment students, ROOTS instruction was provided in addition to core mathematics instruction. The control condition was documented through teacher surveys and direct observations of core instruction. Observation and survey data reflected that teachers used a variety of published and teacher-developed mathematics programs during core instruction. The majority of teachers reported using *Everyday Math* as part of core instruction, and additional published programs included *Houghton Mifflin*, *Bridges in Mathematics*, *Saxon Math*, *Investigations*, and *Engage New York*.

Teachers reported that they provided an average of 31.32 min of daily mathematics instruction ($SD = 9.88$). Survey data also identified that all teachers included mathematics topics during calendar time. All teachers reported that counting and cardinality was incorporated into core mathematics instruction, and 97% of teachers reported that core mathematics instruction addressed operations and algebraic thinking as well as numbers and operations in base 10. Sixty-five percent of teachers noted that knowing number names and the count sequence was their first priority when teaching whole number concepts and skills, while 29% stated that counting to tell the number of objects was the primary instructional priority. All teachers reported that they provided whole group and teacher-led mathematics instruction, and the majority of teachers reported that they provided opportunities for peer or group work, independent student work, and math centers. Seventy-seven percent of teachers also reported providing small group mathematics instruction, and 65% of teachers stated that they provided individual mathematics instruction. Finally,

teachers stated that they regularly incorporated explicit instructional practices into their core math instruction, such as demonstrations of mathematics concepts, guided practice, and opportunities for students to verbalize their mathematical thinking.

Information about the control condition was also gathered from direct observations or core mathematics instruction by trained project staff. Direct observations were conducted in each participating classroom. All observations indicated that ROOTS materials were not used during core instruction, and no evidence of treatment diffusion during core mathematics instruction was identified. Nearly all observations (98%) documented some form of teacher-led instruction, while other instructional formats were observed less frequently, including peer learning, independent student learning, mathematics centers, small group or 1:1 instruction, and instruction via technology. The majority of observations documented instruction on counting (83%) and operations and algebraic thinking (66%). Observations also showed clear evidence of the following principles of explicit and systematic instruction (Gersten et al., 2009): demonstrations of mathematics content, opportunities for group and individual verbalization of mathematics thinking, guided and independent practice opportunities, mathematics representations, and academic feedback. Observations indicated that teachers were less likely to provide scaffolded instruction for struggling students and written mathematics practice for all students.

Fidelity of implementation. Fidelity of implementation was measured via direct observations by trained research staff. Each ROOTS group was observed three times during the course of the intervention. On a 4-point scale (4 = *all*, 3 = *most*, 2 = *some*, 1 = *none*), observers rated the extent to which the interventionist (a) met the lesson's instructional objectives, (b) followed the provided teacher scripting, (c) used the prescribed mathematics models for that lesson, and (d) taught the number of prescribed activities. For example, an interventionist received a rating of 3 for prescribed activities if she taught four of the five activities in an observed lesson. Observations indicated that interventionists delivered the majority of prescribed activities ($M = 4.03$ out of 5 activities, $SD = 0.87$). Interventionists were also observed to meet mathematics objectives ($M = 3.43$, $SD = 0.74$), follow teacher scripting ($M = 3.20$, $SD = 0.77$), and use prescribed mathematics models ($M = 3.58$, $SD = 0.67$). Intraclass correlation coefficients (ICCs) were calculated across observers for these items. ICCs for individual fidelity ratings indicated moderate to nearly perfect agreement: .92 for number of activities delivered, .72 for met mathematics objectives, .72 for followed teacher scripting, and .59 for used prescribed mathematics models. Landis and Koch (1977) characterize ICCs of .41 to .60 as moderate, .61 to .80 as substantial, and .81 to 1.00 as nearly perfect.

Measures

Students were administered five measures of whole number sense at pretest (T_1) and posttest (T_2). These measures included a proximal assessment of whole number understanding that measured skills taught during ROOTS, two distal measures of whole number sense, and a set of curriculum-based measures of discrete skills related to early number sense. In addition, a distal outcome measure was administered 6 months into students' first-grade year (T_3). Trained research staff administered all student measures, and interscorer reliability criteria $\geq .95$ were met for all assessment.

ROOTS Assessment of Early Numeracy Skills (RAENS; Doabler, Clarke, & Fien, 2012) is a researcher-developed instrument that was administered at T_1 and T_2 . RAENS is individually administered and consists of 32 items assessing 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., 10-frames); and write and solve single-digit addition expressions and equations. The predictive validity RAENS ranges from .68 to .83 for the Test of Early Mathematics Ability—Third Edition (TEMA-3) and the NSB. Interrater scoring agreement is reported at 100% (Clarke, Doabler, Smolkowski, Kurtz Nelson, et al., 2016).

Oral Counting—Early Numeracy Curriculum-Based Measurement (Clarke & Shinn, 2004) is a curriculum-based measure that requires students to orally count in English for 1 min. Oral counting scores have predictive validity, with spring criteria ranging from .46 to .72, as well as high interscorer (.99) and test-retest (.78) reliability.

ASPENS (Clarke et al., 2011) is a set of three curriculum-based measures validated for screening and progress monitoring in kindergarten mathematics. Each 1-min fluency-based measure assesses an important aspect of early numeracy proficiency, including number identification, magnitude comparison, and missing number identification. Test-retest reliabilities of kindergarten ASPENS measures are in the moderate to high range (.74–.85). Predictive validity of fall scores on the kindergarten ASPENS measures, with spring scores on the TerraNova 3, ranges from .45 to .52.

NSB is an individually administered measure with 33 items that assess counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations. Jordan and colleagues (Jordan et al., 2008; Jordan, Glutting, Ramineni, & Watkins, 2010) report a coefficient alpha for the NSB of .84 at the beginning of first grade and high levels of diagnostic accuracy, as measured by receiver operating characteristics.

TEMA-3 (Pro-Ed, 2007) is a standardized, norm-referenced, individually administered measure of beginning mathematics ability. The TEMA-3 assesses whole number understanding, including counting and basic calculations,

for children ranging in age from 3 years to 8 years 11 months. The TEMA-3 reports alternate-form and test-retest reliabilities of .97 and .82–.93, respectively. The TEMA-3 manual reports concurrent validity with other mathematics measures ranging from .54 to .91.

Stanford Achievement Test–Tenth Edition (SAT-10; Harcourt Educational Measurement, 2002) and the Stanford Early School Achievement Test (SESAT) are group-administered, standardized, norm-referenced measures. Both measures are multiple choice and have two mathematics subtests: Problem Solving and Procedures. The SESAT is administered in the kindergarten year and the SAT-10 in first grade. The SAT-10 is a standardized achievement test with adequate and well-reported validity ($r = .67$) and reliability ($r = .93$). All treatment and control students were administered the SESAT at posttest (T_2) and the SAT-10 midway through their first-grade year (T_3).

Observations

To gain information about instructional interactions within the two- and five-student ROOTS groups, the Classroom Observations of Student-Teacher Interactions–Mathematics (COSTI-M; Doabler et al., 2015) measure was used during direct observations of ROOTS instruction. This observation measure is a modified version of the Smolkowski and Gunn (2012) early literacy observation instrument that was designed to document the frequency of explicit student-teacher instructional interactions that occur during kindergarten mathematics instruction. Observers used the COSTI-M to collect data on the frequency of teacher models, guided practice, unguided practice, individual practice, and group practice. Teacher models represented teachers clearly explaining and overtly demonstrating mathematical concepts, procedures, and skills. For example, teacher models might include a teacher describing the attributes of 3-dimensional shapes or showing students how to graph data with a bar graph. Guided practice was operationally defined as an opportunity for one student or multiple students to practice a mathematical concept, definition, procedure, strategy, fact, or task with varying levels of concurrent instructional support (physical or verbal) from the teacher. For example, guided practice might include a student or group of students counting with the teacher or tracing a number while directed by the teacher. Unguided practice was defined as an opportunity for one student or multiple students to independently practice a mathematical concept, definition, procedure, strategy, fact, or task without teacher support. Unguided practice might include a student identifying a numeral independently or answering a number combination (e.g., $5 + 1 =$) on one's own. Individual practice opportunities were defined as any practice opportunity (guided or unguided) provided to one student, while group practice opportunities were defined as any practice opportunity provided to two or more

students. One student counting out loud or identifying a numeral would be documented as an individual practice opportunity, while two or more students counting together or writing numerals would be recorded as a group practice opportunity.

Each of these variables is considered an indicator of instructional intensity, with more frequent instructional interactions indicating higher instructional intensity. Mean rates of these behaviors were calculated by dividing the frequency of each behavior during an observed lesson by the number of minutes in the observation. In addition, an “all practice” variable was calculated by summing all observed practice opportunities (i.e., guided, unguided, individual, and group practice) and dividing that total by the number of minutes in the observation.

Each ROOTS group was directly observed three times by trained observers. Observers completed a 6-hr training focused on direct observation procedures and use of the observation instrument. Prior to completing independent observations, observers were required to complete a video checkout in which they coded a 5-min video of small group kindergarten mathematics instruction. Next, observers completed a real-time checkout with a primary observer during a ROOTS observation. On both checkouts, observers were required to meet interobserver reliability standards $\geq .85$. Interobserver reliability ICCs for COSTI-M variables were as follows: .73 for teacher models, .91 for all practice, .94 for all individual practice, .96 for all group practice, .59 for all guided practice, and .72 for all unguided practice. These ICCs indicate nearly perfect agreement for all practice, all individual practice, and all group practice; moderate agreement for all guided practice; and substantial agreement for teacher models and all unguided practice (Landis & Koch, 1977). ICCs were also calculated across the three observations within each ROOTS group to provide an estimate of stability. Stability ICCs for COSTI-M variables were as follows: .06 for teacher models, .37 for all practice, .13 for all individual practice, .32 for all group practice, .40 for all guided practice, and .21 for all unguided practice. These ICCs represent moderate to low stability, indicating that rates of instructional interactions generally differed across observations.

Statistical Analysis

Analyses were conducted to address three research questions. First, we assessed overall ROOTS intervention effects, with two- and five-student ROOTS groups as the intervention condition, on student outcomes using a mixed model (multilevel) Time \times 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 two- or five-student ROOTS groups, or a non-nested 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 among conditions, we tested the assumption of homoscedasticity of residuals. The analysis tested for differences among conditions on gains in outcomes from the fall (T_1) to spring (T_2) of kindergarten and is described in detail by Clarke, Doabler, Smolkowski, Kurtz Nelson, et al. (2016) and Doabler, Clarke, Kosty, et al. (2016). The statistical model included time, coded 0 at T_1 and 1 at T_2 ; condition, coded 0 for control and 1 for ROOTS; and the interaction between the two. These models test for net differences among 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).

Second, we tested whether two- and five-student ROOTS groups experienced differential rates of observed instructional interactions using independent-samples t tests.

Third, we examined the effects of the two- versus five-student ROOTS group size on student outcomes using a fully nested mixed-model (multilevel) Time \times Condition analysis (Murray, 1998) to account for the intraclass correlation associated with students nested within ROOTS groups. Similar to the first set of analyses, the model included time, coded 0 at T_1 and 1 at T_2 ; condition, coded 0 for five-student ROOTS groups and 1 for two-student ROOTS groups; and the interaction between them. Mixed analysis of covariance models were used to analyze the SESAT and the SAT-10 measured at one time point.

Model estimation. We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute Inc., 2009) using restricted maximum likelihood, generally recommended for multilevel models (Hox, 2002). Maximum likelihood estimation for the Time \times 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). 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 families moving).

The models assume independent and normally distributed observations. We addressed the first, more important assumption (Van Belle, 2008) by explicitly modeling the

multilevel nature of the data. The data in the present study also do not markedly deviate from normality; skewness and kurtosis fell with ± 2.0 for all measures except for oral counting, where kurtosis was 3.1. 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. Hedges's g , recommended by the What Works Clearinghouse (2014), represents an individual-level effect size comparable to Cohen's d (Cohen, 1988; Rosenthal & Rosnow, 2008).

Results

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

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 592 students. Attrition rates were approximately 11% for all outcomes measured at T_2 . Only 9% (52) of students were missing all posttest data. The proportion of students missing all posttest data did not differ between the ROOTS condition, with 10% (41) missing, and the control condition, with 6% (11) missing, $\chi^2_{(1)} = 2.08$, $p = .1492$. Although differential rates of attrition are undesirable, differential scores on mathematics tests present a far greater threat to validity, so we conducted an analysis to test whether student mathematics scores were differentially affected by attrition across conditions. We examined the effects of ROOTS condition (two- or five-student group), attrition status, and their interaction on T_1 scores for all five measures available at T_1 . We found no statistically significant interactions or evidence that mathematics scores were differentially affected by attrition across conditions.

Efficacy Effects for ROOTS

Table 3 presents the results of the partially nested statistical models comparing gains between nested ROOTS students and unclustered control students. The table presents the results of the homoscedastic model if it was deemed equivalent to the more complicated heteroscedastic model (ASPENS, TEMA, and RAENS). Otherwise, we provide results for the heteroscedastic model (NSB, oral counting). The bottom two rows of the table show the likelihood ratio

TABLE 2
Descriptive Statistics for Mathematics Measures by Condition and Assessment Time

Measure	Fall of kindergarten (T ₁)			Spring of kindergarten (T ₂)			First grade (T ₃)		
	ROOTS			ROOTS			ROOTS		
	Two-student group	Five-student group	Control	Two-student group	Five-student group	Control	Two-student group	Five-student group	Control
NSB									
<i>M</i>	12.43	12.53	11.97	19.06	19.23	18.26			
<i>SD</i>	3.96	3.68	3.48	4.62	4.79	5.26			
<i>n</i>	120	295	177	108	257	162			
ASPENS									
<i>M</i>	23.90	23.55	21.26	74.18	79.0	56.55			
<i>SD</i>	18.19	18.33	17.39	32.38	36.14	34.80			
<i>n</i>	119	292	175	108	257	162			
Oral counting									
<i>M</i>	21.09	24.08	20.66	43.50	44.82	38.80			
<i>SD</i>	13.83	16.22	14.13	22.55	24.04	21.26			
<i>n</i>	119	292	176	108	257	162			
TEMA									
<i>M</i>	16.79	17.55	16.45	25.25	26.56	23.25			
<i>SD</i>	6.33	7.41	6.46	6.96	7.94	8.13			
<i>n</i>	118	287	174	106	257	163			
RAENS									
<i>M</i>	11.13	11.61	10.92	22.72	23.19	17.81			
<i>SD</i>	5.42	5.73	5.60	5.65	6.13	6.83			
<i>n</i>	118	287	174	106	257	163			
SESAT total									
<i>M</i>				453.68	455.12	446.88			
<i>SD</i>				29.09	33.63	34.11			
<i>n</i>				108	258	163			
SAT-10 total									
<i>M</i>							494.91	497.2	494.98
<i>SD</i>							28.90	26.98	23.95
<i>n</i>							82	195	121

Note. The sample sizes represent students with a particular measure at each assessment period. The complete sample included 120 students in the two-student ROOTS group, 295 students in the five-student ROOTS group, and 177 students in the control condition. NSB = Number Sense Brief; ASPENS = Assessing Student Proficiency in Early Number Sense; TEMA = 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–Tenth Edition.

test results that compared homoscedastic residuals with heteroscedastic residuals. Although the variance structures differed between these models, the condition effect estimates and statistical significance values were very similar for the heteroscedastic and homoscedastic models.

The models in Table 3 tested fixed effects for differences among conditions at pretest (condition effect), gains across time, and the interaction between the two. We found no statistically significant differences at pretest ($p > .16$ 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 three dependent variables. Students in the ROOTS condition made

greater gains than control students on the ASPENS ($t = 6.41$, $df = 272$, $p < .0001$), TEMA standard scores ($t = 3.76$, $df = 263$, $p = .0002$), and RAENS ($t = 9.36$, $df = 315$, $p < .0001$). We did not detect statistically significant differences among conditions in gains on the NSB or oral counting or differences among conditions on the SESAT ($p = .1117$) or SAT-10 ($p = .1253$), both tested with the ASPENS and TEMA as pretest covariates. The Time \times Condition model estimated differences in gains among conditions of 0.4 for the NSB (Hedges's $g = 0.09$), 18.3 for the ASPENS ($g = 0.52$), 3.2 for oral counting ($g = 0.14$), 2.0 for the TEMA standard score ($g = 0.25$), and 4.7 for the RAENS ($g = 0.76$). The analysis of covariance model estimated differences between ROOTS

TABLE 3

Results From a Partially Nested Time \times Condition Analysis on Fall-to-Spring Gains in Math Comparing Intervention Students Nested Within ROOTS Groups and Unclustered Control Students

	NSB	ASPENS	Oral counting	TEMA	RAENS
Fixed effects ^a					
Intercept	11.97 ^{****} (0.33)	21.10 ^{****} (2.05)	20.66 ^{****} (1.35)	16.41 ^{****} (0.55)	10.91 ^{****} (0.45)
Time	6.19 ^{****} (0.31)	35.18 ^{****} (2.23)	18.00 ^{****} (1.64)	6.77 ^{****} (0.41)	6.86 ^{****} (0.40)
Condition	0.53 (0.41)	2.51 (2.55)	2.39 (1.71)	0.92 (0.67)	0.54 (0.54)
Time \times condition	0.43 (0.41)	18.30 ^{****} (2.85)	3.17 (2.07)	1.97 ^{***} (0.52)	4.73 ^{****} (0.50)
Variances					
Gains between ROOTS groups	1.75 [*] (0.69)	71.95 ^{**} (24.89)	25.69 [~] (13.60)	2.08 ^{**} (0.87)	1.52 [*] (0.72)
Covariance: pre-post		324.98 ^{****} (35.37)		39.41 ^{****} (2.86)	21.60 ^{****} (1.75)
Residual		337.19 ^{****} (25.25)		11.73 ^{****} (0.92)	11.73 ^{****} (0.90)
ROOTS					
Residual	8.05 ^{****} (0.66)		212.11 ^{****} (16.76)		
Covariance: pre-post	7.73 ^{****} (1.08)		144.48 ^{****} (22.25)		
Control					
Residual	6.14 ^{****} (1.11)		197.26 ^{****} (27.93)		
Covariance: pre-post	11.65 ^{****} (1.75)		96.72 ^{****} (25.50)		
Time \times condition					
Hedges's g	0.088	0.523	0.139	0.251	0.755
p values	.2974	<.0001	.1272	.0002	<.0001
df	225	272	254	263	315
Likelihood ratio, ^b χ^2	4.67	0.38	3.43	0.73	2.04
p values	.0969	.8258	.1802	.6958	.3612

Note. Table entries show parameter estimates with standard errors in parentheses. NSB = Number Sense Brief; ASPENS = Assessing Student Proficiency in Early Number Sense; TEMA = Test of Early Mathematics Ability–Third Edition; RAENS = ROOTS Assessment of Early Numeracy Skills.

^aTests of fixed effects (first four rows) accounted for small groups as the unit of analysis within the intervention condition (ROOTS) and unclustered individuals in the control condition. ^bThe likelihood ratio test compared homoscedastic residuals with heteroscedastic residuals with a criterion α of .20 ($df = 1$). $\sim p < .10$. $^* p < .05$. $^{**} p < .01$. $^{***} p < .001$. $^{****} p < .0001$.

and control conditions of 3.9 for the SESAT ($g = 0.12$) and 0.1 for the SAT-10 ($g < 0.01$).

Rates of Instructional Interactions

Table 4 presents descriptive statistics for the observed rates of instructional interactions as well as results of independent-samples t tests comparing rates of instructional interactions by ROOTS group size. Compared with the five-student ROOTS groups, two-student ROOTS groups experienced higher rates of individual practice opportunities ($t = 4.25$, $p < .001$, $g = 0.78$) and lower rates of group practice opportunities ($t = -3.18$, $p = .002$, $g = -0.58$). We found no effects of ROOTS group size on the rate of teacher models ($p = .273$), guided practice ($p = .529$), unguided practice ($p = .131$), or all practice combined ($p = .309$).

Effects of the Two- Versus Five-Student ROOTS Group on Student Outcomes

Table 5 presents the results of the fully nested statistical models comparing gains between two- and five-student ROOTS groups. The models in Table 5 tested fixed effects

TABLE 4

Results of Independent-Samples t Tests Comparing Rates of Instructional Interactions by Size of ROOTS Group

	ROOTS, M (SD)		t	p	Hedges's g
	Two-student groups	Five-student groups			
All guided practice	0.6 (0.4)	0.7 (0.3)	-0.63	.529	-0.14
All unguided practice	3.5 (0.7)	3.3 (0.7)	1.52	.131	0.27
All individual practice	2.6 (0.8)	2.1 (0.6)	4.25	<.001	0.78
All group practice	1.4 (0.7)	1.8 (0.7)	-3.18	.002	-0.58
All practice	4.1 (0.9)	3.9 (0.8)	1.02	.309	0.18

Note. Group t tests were based on 60 two-student ROOTS groups and 59 five-student ROOTS groups ($df = 117$).

for differences among conditions at pretest (two-student ROOTS group effect), gains across time, and the interaction

TABLE 5

Results From a Fully Nested Time \times Condition Analyses on Fall-to-Spring Gains in Math Comparing Two- and Five-Student ROOTS Groups

	NSB	ASPENS	Oral counting	TEMA	RAENS
Fixed effects					
Intercept	12.53 ^{****} (0.33)	23.41 ^{****} (2.11)	24.02 ^{****} (1.44)	17.58 ^{****} (0.59)	11.60 ^{****} (0.42)
Time	6.62 ^{****} (0.27)	55.19 ^{****} (1.89)	20.49 ^{****} (1.29)	8.74 ^{****} (0.38)	11.51 ^{****} (0.36)
Two-student ROOTS group	-0.10 (0.54)	0.35 (3.45)	-3.01 (2.41)	-0.83 (0.95)	-0.50 (0.71)
Time \times Two-Student ROOTS Group	0.00 (0.49)	-4.81 (3.31)	1.96 (2.39)	-0.09 (0.65)	0.15 (0.62)
Variances					
ROOTS group intercept	3.52 ^{***} (0.92)	116.45 ^{**} (35.26)	56.93 ^{***} (16.59)	10.28 ^{***} (2.86)	3.38 [*] (1.48)
ROOTS group gains	0.20 (0.44)	23.72 (20.05)	-1.13 (9.50)	1.26 [~] (0.76)	0.99 (0.67)
Student	5.53 ^{****} (0.93)	246.79 ^{****} (38.44)	104.56 ^{****} (20.15)	29.96 ^{****} (3.03)	17.16 ^{****} (2.01)
Residual	8.77 ^{****} (0.77)	362.39 ^{****} (31.11)	226.43 ^{****} (18.97)	12.52 ^{****} (1.10)	12.17 ^{****} (1.06)
Time \times Two-Student ROOTS Group					
Hedges's g	0.000	-0.137	0.083	-0.012	0.026
p values	.9969	.1481	.4130	.8846	.8062
df	162	154	191	150	166

Note. Table entries show parameter estimates with standard errors in parentheses. NSB = Number Sense Brief; ASPENS = Assessing Student Proficiency in Early Number Sense; TEMA = Test of Early Mathematics Ability—Third Edition; RAENS = ROOTS Assessment of Early Numeracy Skills.

^aTests of fixed effects (first four rows) accounted for small groups as the unit of analysis within the two- and five-student ROOTS conditions.

[~] $p < .10$. ^{*} $p < .05$. ^{**} $p < .01$. ^{***} $p < .001$. ^{****} $p < .0001$.

between the two. We found no statistically significant differences at pretest ($p > .21$ for all measures), which suggested that students were similar in the fall of kindergarten. We found no statistically significant differences by ROOTS group size in gains from fall to spring ($p > .15$ for all measures). The Time \times Condition model estimated differences in gains between ROOTS group sizes of 0.0 for the NSB (Hedges's $g = 0.00$), -4.8 for the ASPENS ($g = -0.14$), 2.0 for oral counting ($g = 0.08$), -0.1 for the TEMA standard score ($g = -0.01$), and 0.15 for the RAENS ($g = 0.03$). The analysis of covariance model estimated differences between two- and five-student ROOTS groups of 1.0 for the SESAT ($g = 0.03$) and 0.5 for the SAT-10 ($g = 0.02$).

Discussion

As educators grapple with building and providing better services within RTI or MTSS frameworks, research examining intervention efficacy is crucial, as are research questions that focus on moderators and mediators of treatment impact (Miller et al., 2014), including those related to treatment intensity and the allocation of finite resources (Coddling & Lane, 2015). Our examination of the ROOTS intervention program found that overall results for the ROOTS program were effective, with a significant positive impact on 3 of 6 posttest measures and all measures with a positive effect size. Results for ROOTS program would be classified by the What Works Clearinghouse as having a “statistically significant positive impact.” Second, we found significant differences between the two- and five-student small groups, with

the two-student small group providing a higher rate of individual practices opportunities and the five-student group providing a higher rate of group practice opportunities. No differences were found on other teacher and student behaviors. Despite finding differences on one measure of treatment intensity favoring the two-student small group (the rate of individual practice opportunities), we did not detect significant differences on student achievement outcome measures between the ROOTS two- and five-student groups.

The first finding related to the efficacy of the ROOTS intervention adds to the corpus of research on this particular intervention program and the general body of research on whole number interventions targeting the early elementary grades (e.g., D. P. Bryant et al., 2011; Clarke et al., 2014; Fuchs et al., 2005). For districts or schools implementing RTI or MTSS, the research base enables them to select from a growing number of programs (<http://www.intensiveintervention.org/>) that, if implemented with fidelity, can reasonably be expected to positively affect student outcomes and function as a component of a framework to support early mathematics achievement. Our second and third research questions present a more complex context in which to interpret our findings. While results from the study indicated a more intensive intervention experience for students in the two-student small group, this did not translate into greater student achievement outcomes. Critically, two things should be considered when contextualizing the results from the present study. First, from a school- or resource-based perspective, the lack of significant differences between small groups has significant resource allocation implications.

Second, what do our lack of findings mean for understanding of treatment intensity, and how should that guide future research efforts? We address each of these areas in turn.

At a federal level, there is an increasing interest in considering cost when examining the efficacy of educational programs. For example, the Institute of Education Sciences' 2016 Special Education Grants Request for Applications requires a cost analysis section as part of the research plan for Goal 3 efficacy and replication grants: "The cost analysis should help schools and districts understand the monetary costs of implementing the intervention (e.g., expenditures for personnel, facilities, equipment, materials, training, and other relevant inputs), and "Intervention costs can be contrasted with the costs of comparison group practice to reflect the difference between them" (p. 65). Cost analyses do not include measures of benefit (Levin, 1983), and procedures for examining costs have become relatively standardized (Levin & McEwan, 2001), thus enabling them to provide a quantitative metric to help examine questions related to cost-benefit and an additional and important lens through which to view and guide practice. For example, findings by Elbaum et al. (2000) related to the benefits of 1:1 instruction could and should be contrasted with reading interventions targeting similar content but utilizing larger groups sizes (Gersten et al., 2008) and allow comparisons of programs providing similar benefits at varying costs (Keeney & Raiffa, 1993). In a similar vein, the work of Vaughn and colleagues (Vaughn, Cirino, et al., 2010; Vaughn et al., 2003) affords opportunities to integrate cost analyses and subsequent consideration of cost-benefit into the informal evaluation of an intervention program to complement the formal evaluation of whether or not a program is efficacious.

In a district setting where monetary resources are likely capped (e.g., a set amount exists to provide Tier 2 intervention services), schools can serve as many students as possible with the available dollars. Using a ROOTS grouping size of five would allow schools to serve 150% more students than if a two-student grouping was selected. In real terms, this is a significant difference and critical in the ability to implement a multitier model. For example, in a large-scale Institute of Education Sciences–funded efficacy trial (Clarke, Doabler, Fien, Baker, & Smolkowski, 2012), we found that approximately 70% of students entered kindergarten with some degree of risk in mathematics that, in a multitier model, would warrant additional Tier 2 services. Utilizing a cost-benefit framework allows schools to evaluate equivalent positive results between the two- and five-student small groups from a resource standpoint.

Limitations and Future Research

From a treatment intensity perspective, a number of factors are important to consider that relate to limitations of the current research and directions for future research. Although

we describe our five-student small group as less intense, that should not be conflated with considering the group to lack intensity. That is, the experience of students in the five-student small group would be, by almost any analysis, an intensive educational experience. The ROOTS program was designed to incorporate effective instructional design elements for at-risk learners (Archer & Hughes, 2011), and the resulting instructional experience for students included high levels of teacher models and demonstrations, opportunities to respond, and academic feedback—all variables with a demonstrated positive relationship with student outcomes (Baker et al., 2002; Gersten et al., 2009; Kroesbergen & Van Luit, 2003). The experience of the five-student small group included high rates of individual practice compared with typical instruction with high rates of group practice. Note that the selection of group size for the study was driven by recommendations for group size (Gersten et al., 2009) and by practical research design considerations (e.g., potential attrition with one-student small groups). Thus, we are not able to make statements regarding contrasts related to other group sizes.

For programs that consider these instructional design features a priori in their design and development phases, it may be that a threshold effect exists wherein after a certain base rate of critical teacher and student behavior is reached, the value of providing additional opportunities to engage in those behaviors is limited (Doabler et al., 2017). If a threshold effect exists, that would mean that structuring groups, including reducing group size, in ways to increase behaviors thought to theoretically underlie the intervention may not result in hypothesized higher outcomes. Note that in this investigation, students in the five-student group experienced significantly higher rates of group practice. Well-designed instruction with group practice built into its architecture may enable group practice to elicit the same benefit as individual practice. Future research should continue to examine the role of critical teacher and student behaviors and their interaction with group size and student outcomes in a variety of contexts. Such designs should occur with programs designed to include critical instructional principles at high rates but also those designed with different parameters. The inclusion criterion to identify the study sample was not focused exclusively on identifying a high-risk sample (e.g., students with mathematics learning disabilities) but included a broader range of mathematics abilities. Relatedly, we did not investigate whether the impact of group sizes was mediated by initial skill status. For example, it may be reasonable to hypothesize that a student with relatively low risk (within the at-risk sample) gains equal benefit from either group size but a student with relatively high risk may gain differential benefit from the smaller small group.

It is also vital to examine how we defined and operationalized treatment intensity. Our operationalization of treatment intensity focuses on a narrow, albeit critical, set of behaviors

and did not attempt to account for the quality of those specific behaviors. For example, while we captured the rate of teacher models, we did not analyze the overall quality of those models. Future research should examine overall quality of behaviors hypothesized to influence student outcomes. In addition, there is significant interest in the role of teacher content and pedagogical knowledge (Garet et al., 2016; Woodward, 2016). Our measurement net did not include measures of teacher knowledge. In a systematically designed program, like ROOTS and similar early mathematics intervention programs, the content knowledge of the instructor may play a vital role. A potential hypothesis is that a teacher with greater content knowledge would offer more sophisticated mathematics models and academic feedback and thus provide students with a more conceptually rich academic experience leading to greater mathematics outcomes. If such relationships are discovered in future studies, links to targeted professional development would be worth exploring despite mixed results from studies targeting that area (Gersten, Taylor, Keys, Rolffhus, & Newman-Gonchar, 2014).

Last, while we defined treatment intensity within the scope of one lesson (i.e., the teacher and student behaviors that occur within the lesson), treatment intensity can also be thought of as the scope of the intervention and the amount of content covered. Emerging evidence suggests that kindergarten mathematics content is often overfocused on basic concepts associated with smaller gains for almost all students when a focus on advanced content has been positively associated with student learning (Engel, Claessens, & Finch, 2013; Engel, Claessens, Watts, & Farkas, 2016). While an approach of this nature is complicated in the context of working with at-risk or Tier 2 students, the point speaks to a broader issue. If we are to reduce achievement gaps and reset the foundation of students' mathematical understanding such that they are able to acquire new material at the same rate of their peers, attempts to push the envelope in terms of the content covered in mathematics interventions is necessary. Doing so would require a significant rethinking of the traditional intervention model of providing interventions of limited duration and depth that are designed to build understanding that has already been mastered by same-grade peers.

Conclusion

Systematic examinations of intervention of delivery options are of particular interest within multitier models (Al Otaiba, Kim, Wanzek, Petscher, & Wagner, 2014; Vaughn, Denton, & Fletcher, 2010) and resource-limited environments. While work in mathematics is just beginning, it has the potential to inform the field regarding best practices in intervention (Miller et al., 2014; Vaughn & Swanson, 2015) as we strive to better understand, study, and implement models of support that address the learning needs of all students in acquiring mathematics knowledge.

Acknowledgments

The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education. An independent external evaluator and coauthor of this publication completed the research analysis described herein.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: 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

Funding

This research was supported by the ROOTS Project (Grant R324A120304), funded by the Institute of Education Sciences, U.S. Department of Education.

References

- Al Otaiba, S., Kim, Y.-S., Wanzek, J., Petscher, Y., & Wagner, R. K. (2014). Long term effects of first grade multi-tier intervention. *Journal of Research on Educational Effectiveness, 7*(3), 250–267. doi:10.1080/19345747.2014.906692
- Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114. doi:10.2307/271083
- Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford Press.
- Baker, S. K., Fien, H., & Baker, D. L. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of Tier 1 and Tier 2 instructional supports. *Focus on Exceptional Children, 42*(9), 1–20.
- 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. Retrieved from <http://www.jstor.org/stable/1002308>
- Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*, 149–165. doi:10.1037/a0023464
- Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. M. (2015). *Evaluation of response to intervention practices for elementary school reading* (NCEE No. 2016-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from <http://files.eric.ed.gov/fulltext/ED560820.pdf>
- Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43*, 210–236. doi:10.1080/00273170802034810
- 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
- Breit-Smith, A., Justice, L. M., McGinty, A. S., & Kaderavek, J. (2009). How often and how much? Intensity of print

- referencing intervention. *Topics in Language Disorders*, 29, 360–369. doi:10.1097/TLD.0b013e3181c29db0
- Bryant, B. R., Bryant, D. P., Porterfield, J., Dennis, M. S., Falcomata, T., Valentine, C., . . . Bell, K. (2016). The effects of a Tier 3 intervention on the mathematics performance of second grade students with severe mathematics difficulties. *Journal of Learning Disabilities*, 49, 176–188. doi:10.1177/0022219414538516
- 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
- Clarke, B., Baker, S. K., & Chard, D. J. (2008). Best practices in mathematics assessment and intervention with elementary students. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology* (Vol. 5, pp. 453–464). Bethesda, MD: National Association of School Psychologists.
- Clarke, B., Baker, S. K., & Fien, H. (2009). *Foundations of mathematical understanding: Developing a strategic intervention on whole number concepts*. Eugene, OR: Center on Teaching and Learning.
- Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). *A randomized control trial of a Tier 2 kindergarten mathematics intervention*. Eugene: University of Oregon.
- Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2014). 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., 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., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness*, 9, 607–634. doi:10.1080/19345747.2015.1116034
- Clarke, B., Gersten, R. M., Dimino, J., & Rolffhus, E. (2011). *Assessing Student Proficiency in Early Number Sense (ASPENS)*. Longmont, CO: Cambium Learning Group.
- Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234–248.
- Codding, R. S., & Lane, K. L. (2015). A spotlight on treatment intensity: An important and often overlooked component of intervention inquiry. *Journal of Behavioral Education*, 24, 1–10. doi:10.1007/s10864-014-9210-z
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.
- Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. Retrieved from <http://www.coresstandards.org/the-standards/mathematics>
- Coyne, M. D., Kame'enui, E. J., & Carnine, D. (2011). *Effective teaching strategies that accommodate diverse learners* (4th ed.). Upper Saddle River, NJ: Pearson Education.
- Denton, C. A., Tolar, T. D., Fletcher, J. M., Barth, A. E., Vaughn, S., & Francis, D. J. (2013). Effects of Tier 3 intervention for students with persistent reading difficulties and characteristics of inadequate responders. *Journal of Educational Psychology*, 105, 633–648. doi:10.1037/a0032581
- Doabler, C. T., Baker, S. K., Kosty, D., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *Elementary School Journal*, 115, 303–333.
- Doabler, C. T., Clarke, B., & Fien, H. (2012). *ROOTS Assessment of Early Numeracy Skills (RAENS)* [Unpublished measurement instrument]. Eugene, OR: Center on Teaching and Learning.
- Doabler, C. T., Clarke, B., Kosty, D., Kurtz-Nelson, E., Fien, F., Smolkowski, K., & Baker, S. K. (2016). Testing the efficacy of a Tier-2 mathematics intervention: A conceptual replication study. *Exceptional Children*, 83, 92–110.
- Doabler, C. T., Clarke, B., Stoolmiller, M., Kosty, D., Fien, F., Smolkowski, K., & Baker, S. K. (2017). Treatment intensity and instructional interactions: Exploring the black box of a Tier 2 mathematics intervention. *Remedial and Special Education*. Advance online publication. doi:10.1177/0741932516654219
- 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
- Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, 92, 605–619. doi:10.1037/0022-0663.92.4.605
- Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, 35, 157–178. doi:10.3102/0162373712461850
- Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher*, 45, 293–300. doi:10.3102/0013189x16656841
- Ezell, H. K., Justice, L. M., & Parsons, D. (2000). Enhancing the emergent literacy skills of pre-schoolers with communication disorders: A pilot investigation. *Child Language Teaching and Therapy*, 16, 121–140. doi:10.1177/026565900001600202
- Fien, H., Santoro, L., Baker, S. K., Park, Y., Chard, D. J., Williams, S., & Haria, P. (2011). Enhancing teacher read alouds with small-group vocabulary instruction for students with low vocabulary in first-grade classrooms. *School Psychology Review*, 40, 307–318.
- Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). *Teaching math to young children: A practice guide* (Practice Guide No. NCEE 2014-4005). Washington, DC: National Center for Education and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/early_math_pg_111313.pdf
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513. doi:10.1037/0022-0663.97.3.493

- Fuchs, L. S., Fuchs, D., & Hollenbeck, K. N. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research & Practice, 22*, 13–24. doi:10.1111/j.1540-5826.2007.00227.x
- Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., & Bryant, J. D. (2010). The contributions of numerosity and domain-general abilities to school readiness. *Child Development, 81*, 1520–1533. doi:10.1111/j.1467-8624.2010.01489.x
- Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45*, 195–203. doi:10.1177/0022219412442150
- Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., . . . Wei, T. E. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE No. 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance.
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345–362. doi:10.1037/0033-2909.114.2.345
- Gersten, R. M. (2016). What we are learning about mathematics interventions and conducting research on mathematics interventions. *Journal of Research on Educational Effectiveness, 9*, 684–688. doi:10.1080/19345747.2016.1212631
- 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). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_math_pg_042109.pdf
- Gersten, R. M., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33*, 18–28. doi:10.1177/002246699903300102
- Gersten, R. M., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2008). *Mathematics instruction for students with learning disabilities or difficulty learning mathematics: A synthesis of the intervention research*. Portsmouth, NH: Center on Instruction.
- Gersten, R. M., Taylor, M. J., Keys, T. D., Rolffhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches* (No. REL 2014-010). Washington, DC: Regional Educational Laboratory Southeast. Retrieved from <http://files.eric.ed.gov/fulltext/ED544681.pdf>
- 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
- Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review, 20*, 338–352. doi:10.1177/0193841x9602000306
- Harcourt Educational Measurement. (2002). *Stanford Achievement Test (SAT-10)*. San Antonio, TX: Author.
- 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: Erlbaum.
- Individuals with Disabilities Education Act of 2004, Public Law 108-446, 34 C.F.R. § 300.8 (c)(10) (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., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39*, 181–195.
- Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology, 94*, 586–597. doi:10.1037/0022-0663.94.3.586
- Justice, L. M., & Ezell, H. K. (2000). Enhancing children's print and word awareness through home-based parent intervention. *American Journal of Speech-Language Pathology, 9*, 257–269. doi:10.1044/1058-0360.0903.257
- Justice, L. M., & Ezell, H. K. (2002). Use of storybook reading to increase print awareness in at-risk children. *American Journal of Speech-Language Pathology, 11*, 17–29. doi:10.1044/1058-0360(2002/003)
- Justice, L. M., Kaderavek, J. N., Fan, X., Sofka, A., & Hunt, A. (2009). Accelerating preschoolers' early literacy development through classroom-based teacher-child storybook reading and explicit print referencing. *Language, Speech, and Hearing Services in Schools, 40*, 67–85. doi:10.1044/0161-1461(2008/07-0098)
- Justice, L. M., McGinty, A. S., Piasta, S. B., Kaderavek, J. N., & Fan, X. (2010). Print-focused read-alouds in preschool classrooms: Intervention effectiveness and moderators of child outcomes. *Language, Speech, and Hearing Services in Schools, 41*, 504–520. doi:10.1044/0161-1461(2010/09-0056)
- Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge, UK: Cambridge University Press.
- 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
- 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
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. doi:10.2307/2529310
- Levin, H. M. (1983). *Cost-effectiveness: A primer* (Vol. 4). Beverly Hills, CA: Sage.

- Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications*. Thousand Oaks, CA: Sage. Retrieved from <https://books.google.com/books?id=HniLG23vYDwC>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Lovelace, S., & Stewart, S. R. (2007). Increasing print awareness in preschoolers with language impairment using non-evocative print referencing. *Language, Speech, and Hearing Services in Schools, 38*, 16–30. doi:10.1044/0161-1461(2007/003)
- McGinty, A. S., Breit-Smith, A., Fan, X., Justice, L. M., & Kaderavek, J. N. (2011). Does intensity matter? Preschoolers' print knowledge development within a classroom-based intervention. *Early Childhood Research Quarterly, 26*, 255–267. doi:10.1016/j.ecresq.2011.02.002
- 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. (2016). Who is at risk for persistent mathematics difficulties in the United States? *Journal of Learning Disabilities, 49*, 305–319. doi:10.1177/0022219414553849
- Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306–321. doi:10.1177/0022219408331037
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.
- National Assessment of Educational Progress. (2015). *The nation's report card: 2015 mathematics and reading assessments. National achievement level results, 8th grade*. Washington, DC: National Center for Education Statistics. Retrieved from http://www.nationreportcard.gov/reading_math_2015/#reading/acl?grade=8
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.
- Ochsendorf, R. (2016). Advancing understanding of mathematics development and intervention: Findings from NCSER-funded efficacy studies. *Journal of Research on Educational Effectiveness, 9*, 570–576. doi:10.1080/19345747.2016.1222144
- Pro-Ed. (2007). *Test of Early Mathematics Ability—Third Edition (TEMA-3)*. Austin, TX: ProEd.
- 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 Inc. (2009). *SAS/STAT 9.2 user's guide*. Cary, NC: Author.
- 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. (2013). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of Learning Disabilities, 46*, 328–346. doi:10.1177/0022219411422380
- Thurlow, M. L., Ysseldyke, J. E., Wotruba, J. W., & Algozzine, B. (1993). Instruction in special education classrooms under varying student-teacher ratios. *The Elementary School Journal, 93*, 305–320. doi:10.2307/1001897
- Van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). Hoboken, NJ: Wiley.
- Vaughn, S., Cirino, P. T., Wanzek, J., Wexler, J., Fletcher, J. M., Denton, C. D., . . . Francis, D. J. (2010). Response to intervention for middle school students with reading difficulties: Effects of a primary and secondary intervention. *School Psychology Review, 39*, 3–21.
- Vaughn, S., Denton, C. A., & Fletcher, J. M. (2010). Why intensive interventions are necessary for students with severe reading difficulties. *Psychology in the Schools, 47*, 432–444. doi:10.1002/pits.20481
- Vaughn, S., & Swanson, E. A. (2015). Special education research advances knowledge in education. *Exceptional Children, 82*, 11–24. doi:10.1177/0014402915598781
- Vaughn, S., Thompson, S. L., Kouzekanani, K., & Dickson, S. (2003). The effects of three grouping formats on the reading performance of monolingual and English language learners with reading problems. *Journal of Educational Psychology, 24*, 301–315.
- Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*, 541–561.
- 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. (2014). *Procedures and standards handbook. Version 3.0*. Washington, DC: Institute of Education Sciences. Retrieved from <http://ies.ed.gov/ncee/www/DocumentSum.aspx?sid=19>
- Woodward, J. (2016). Commentary on intensive interventions: What are the limits of highly structured curriculum for at-risk students? *Journal of Research on Educational Effectiveness, 9*, 678–683. doi:10.1080/19345747.2016.1212630

Authors

BEN CLARKE, University of Oregon; research interests: early numeracy, curriculum-based measurement, instructional design.

CHRISTIAN T. DOABLER, University of Texas at Austin; research interests: curriculum design, mathematics interventions, learning disabilities.

DEREK KOSTY, Oregon Research Institute; research interests: design and analysis of complex efficacy and effectiveness trials.

EVANGELINE KURTZ NELSON, University of Oregon; research interests: Tier 2 mathematics interventions, supporting learners with intellectual and developmental disabilities.

KEITH SMOLKOWSKI, Oregon Research Institute; research interests: design and analysis of complex efficacy and effectiveness trials.

HANK FIEN, University of Oregon; research interests: early reading and mathematics interventions, formative assessments.

JESSICA TURTURA, University of Oregon; research interests: early math and reading intervention strategies and materials, factors predicting nonresponse to intervention, schoolwide multitiered systems of support.