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 Tier 2 Mathematics Intervention: A Conceptual Replication Study

Exceptional Children
2016, Vol. 83(1) 92–110
© 2016 The Author(s)
DOI: 10.1177/0014402916660084
ec.sagepub.com

Christian T. Doabler[1], Ben Clarke[2], Derek B. Kosty[3], Evageline Kurtz-Nelson[2], Hank Fien[2], Keith Smolkowski[3], and Scott K. Baker[4]

Abstract

The purpose of this closely aligned conceptual replication study was to investigate the efficacy of a Tier 2 kindergarten mathematics intervention. The replication study differed from the initial randomized controlled trial on three important elements: geographical region, timing of the intervention, and instructional context of the counterfactual. Similar to the original investigation, however, the current study tested the same intervention, used the same outcome measures and statistical analyses, and involved the same population of learners. A total of 319 kindergarten students with mathematics difficulties from 36 kindergarten classrooms participated in the study. Students who were randomly assigned to the treatment condition received the intervention in small-group formats, with 2 or 5 students per group. Control students participated in a no-treatment control condition. Significant effects on proximal and distal measures of mathematics achievement were found. Effect sizes obtained for all measures fell within or exceeded the upper bound of the effects reported in the initial study. Implications for systematically situating replication studies in larger frameworks of intervention research and reporting rates of treatment response across replication studies are discussed.

Replication is a fundamental principle of scientific research (Flay et al., 2005; Gottfredson et al., 2015; Schmidt, 2009; Valentine et al., 2011). Replication studies allow the research community to rule out chance as a plausible explanation of previous findings and build a convergence of empirical evidence in favor of an intervention or instructional practice (Coyne, Cook, & Therrien, 2016; Gottfredson et al., 2015). These studies also serve as a way to determine if findings obtained in a previous study hold up to variations in settings and contextual factors (Coyne et al., 2013). Establishing evidence of replication, therefore, adds credibility and generalizability to the hypotheses and claims of the original research (Schmidt, 2009).

Operationalizing Replication in Educational Research

Within most scientific fields, replication studies are categorized in two ways: *direct replication* and *conceptual replication* (Coyne et al., 2016; Makel, Plucker, & Hegarty, 2012; Schmidt, 2009). Direct replications are conducted using the same methods and under the

[1]University of Texas at Austin
[2]University of Oregon
[3]Oregon Research Institute
[4]Southern Methodist University

Corresponding Author:
Christian T. Doabler, University of Texas at Austin, 1 University Station, D5300, SZB 408B, Austin, TX 78712, USA.
E-mail: cdoabler@austin.utexas.edu

same conditions as the original research. In the field of education research, direct replications are difficult, if not impossible, to conduct given the complex and dynamic environments of schools. A more feasible option for educational researchers is to conduct *closely aligned* conceptual replications (Coyne et al., 2016). Closely aligned conceptual replications are conducted to verify whether findings generalize across settings, conditions, and participants. Such replications typically vary from the original study on one or two elements (Schmidt, 2009). If such variations are minimal, conceptual replications can demonstrate similar capacity as direct replications. For example, a closely aligned replication can uncover whether treatment effects demonstrated in the original study replicate in a different geographical region (Coyne et al., 2013).

Replication Research and the Changing Landscape of the Counterfactual

Another aspect of replication studies concerns the control condition, or the counterfactual. The counterfactual represents what might have occurred had the treated sample not received the treatment (Lemons, Fuchs, Gilbert, & Fuchs, 2014; Shadish, Cooke, & Campbell, 2002). Replication studies of beginning reading interventions have discovered that dimensions of the counterfactual change across time (Lemons et al., 2014) and vary by school and across different geographical regions (Coyne et al., 2013). Results from these studies suggest that an intervention implemented in a similar fashion to the same population may produce vastly different results based on the nature and strength of the counterfactual. Variability of the counterfactual therefore may change interpretation of observed treatment effects across a program of research (Lemons et al., 2014). For example, a study that establishes a mathematics intervention as evidence based relative to a counterfactual could lose its treatment effect if the counterfactual shifts or strengthens in a subsequent replication study.

Instructional dimensions of the counterfactual are commonly billed as "business-as-usual" (BAU) instruction. In many mathematics intervention studies, BAU instruction represents the core mathematics instruction provided in general education settings. The aim of core mathematics instruction is to address the range of mathematics standards that students are expected to know at the end of each grade level (e.g., Common Core State Standards; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). When core mathematics instruction represents the counterfactual in mathematics intervention studies, its nature and strength is also expected to vary within and across classrooms. In part this is due to the diverse range of commercially available core mathematics programs and instructional approaches used to teach mathematics in U.S. classrooms (Morgan, Farkas, & Maczuga, 2015).

Replication studies allow the research community to rule out chance as a plausible explanation of previous findings and build a convergence of empirical evidence in favor of an intervention or instructional practice.

In summary, the control condition or counterfactual is an essential component of intervention research as it allows researchers to rule out alternative explanations of a treatment effect. However, the nature and strength of the counterfactual can vary by location and improve across time. As evidenced by prior research, such variations of the counterfactual may attenuate observed treatment effects (Coyne et al., 2013; Lemons et al., 2014).

Purpose of the Study

The primary purpose of this study was to conduct a closely aligned conceptual replication that tested the efficacy of the ROOTS mathematics intervention in schools that offered different instructional and contextual dimensions than those included in a previous investigation of the intervention (Clarke et al., 2016). Because ROOTS had demonstrated preliminary treatment effects (Clarke et al., 2016), the current study sought to examine whether the intervention's efficacy would hold up to planned variations in geographical location, intervention timing, and instructional context of the counterfactual while holding constant other theoretically important variables (e.g., intervention dosage). A secondary purpose was to situate the current study within an a priori framework of systematic replication (Coyne et al., 2016).

The ROOTS intervention is a 50-lesson Tier 2 kindergarten program designed to accelerate the learning of kindergarten students with mathematics difficulties (MD). In this article, we use the term MD rather than *mathematics disability* because it encompasses a broader range of students. Students with MD perform at or below the 35th percentile on standardized mathematics measures. Central to the ROOTS intervention is the integration of foundational concepts of whole number and validated principles of systematic and explicit mathematics instruction (Archer & Hughes, 2010; Doabler & Fien, 2013; Gersten et al., 2009). Our initial efficacy trial of ROOTS, which occurred in the 2012–2013 school year, utilized a randomized block design (Clarke et al., 2016). A total of 290 students from 37 kindergarten classrooms in Oregon were randomly assigned within classrooms to one of three conditions: (a) a ROOTS-Small intervention group with a 2:1 student-teacher ratio (*n* = 58), (b) a ROOTS-Large intervention group with a 5:1 student-teacher ratio (*n* = 145), or (c) a no-treatment (BAU) control condition (*n* = 87). Despite group size differences, students in both ROOTS groups received the same intervention dosage (i.e., 50 lessons, 5 days per week, 20 min per day). Control students received only the core mathematics instruction provided in their kindergarten classrooms. Treatment students, however, received the ROOTS intervention in addition to their core mathematics instruction.

To test the efficacy of ROOTS in Oregon, Clarke et al. (2016) aggregated treatment students in ROOTS-Small and ROOTS-Large intervention groups and compared their mathematics achievement gains across the intervention time period to students in the control condition. Findings from the initial efficacy trial indicated statistically significant differences in favor of ROOTS students on four of the six outcome measures. Reported effect sizes (Hedges' *g*) for the Test of Early Mathematics Ability–Edition 3 (TEMA-3; Ginsburg & Baroody, 2003), Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke, Rolfhus, Dimino, & Gersten, 2012); ROOTS Assessment of Early Numeracy Skills (RAENS; Doabler, Clarke, & Fien, 2012), and oral counting (Clarke & Shinn, 2004) were 0.32, 95% confidence interval (CI) [0.13, 0.50]; 0.58, 95% CI [0.36, 0.79]; 0.75, 95% CI [0.53, 0.96]; and 0.28, 95% CI [0.02, 0.54], respectively. Statistically significant differences between conditions were not found for the Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2008), the Stanford Early School Achievement Test (SESAT; Harcourt Brace Educational Measurement 2003), and Stanford Achievement Test–10th Edition (SAT-10; Harcourt Brace Educational Measurement, 2002).

Variability of the counterfactual therefore may change interpretation of observed treatment effects across a program of research.

We consider the current study as a closely aligned replication based on the degree of overlap with the Oregon efficacy trial. For example, both studies employed a randomized block design and applied the same a priori criteria for determining students' eligibility for the intervention. In both studies, the 10 lowest-performing kindergarten students from each classroom who met the eligibility criteria were randomly assigned within classrooms to the same three experimental conditions. In addition, similar to the Oregon study, the replication had district-employed personnel deliver the ROOTS intervention in small-group formats and at the same dose frequency (i.e., 50 lessons, 20 min per day, 5 days per week for approximately 10 weeks). Analyses used in both studies aggregated students in the ROOTS groups to compare gains in mathematics outcomes relative to students in the control condition.

Notwithstanding strong overlap with the research design, analytic procedures, and intervention dosage levels, the current study varied from the Oregon efficacy trial on several important contextual and instructional dimensions. One difference between the two studies was geographical location. Whereas Clarke et al. (2016) investigated ROOTS in suburban and rural schools from across Oregon, the current study was conducted in urban and suburban schools from the metropolitan area of Boston, Massachusetts. It is important to note that in the current study, we purposefully varied the geographical region to determine whether the effects of ROOTS held up in classrooms with different instructional contexts and students with different sociodemographic characteristics. For example, schools that participated in the replication had a more diverse demography and served larger percentages of students from economically disadvantaged backgrounds than the Oregon schools. Onset of the intervention in the Boston study also occurred at a different time point in the school year than the original efficacy trial. Students in the Oregon study received ROOTS starting in mid-January; however, the replication began nearly 2 months earlier (mid-November). We altered the onset of the intervention to see if an earlier start time in the school year was more effective for at-risk kindergarteners and would better align with students' acquisition of early number sense.

In addition, the core mathematics programs used in the replication study's kindergarten classrooms (i.e., control condition) differed from those delivered in the Oregon study. The replication's programs were also noted as having a more robust evidence base for improving student mathematics achievement. Recognizing these contextual and instructional differences, the new research sites were expected to offer a unique counterfactual. Thus, positive findings from the replication would increase the credibility and generalizability of the intervention's beneficial impact on student mathematics outcomes.

Method

This replication of the efficacy evaluation of ROOTS was examined in a partially nested randomized controlled trial (RCT) described by Bauer, Sterba, and Hallfors (2008) and Baldwin, Bauer, Stice, and Rohde (2011). The 10 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. The study is partially nested because the intervention students were nested within interventionists' small groups whereas control students were not. 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. This design requires a specific analysis approach to account for clustering, which we describe later.

The study was designed to allow for four replications of the efficacy of ROOTS. The potentially important difference between the two treatment groups (i.e., 2:1 ROOTS-Small vs. 5:1 ROOTS-Large) was a key aim of the full, 4-year efficacy trial. Because the differences between students in the two intervention groups was expected to be smaller than those between intervention and control, this replication study, by itself, is underpowered to test questions about group size. Students in the ROOTS groups were combined to compare their gains on important mathematics outcomes relative to students in the no-treatment control condition. Future research will investigate the effect of grouping within the ROOTS intervention.

Participants

The principal investigators and key personnel of the ROOTS project conducted all recruitment efforts. These efforts entailed contacting district leaders of public school districts in the metropolitan area of Boston, Massachusetts. District leaders were provided information on the study's research aims and activities. Interested district leaders then identified potential schools for participation, namely, those that contained large percentages of students in need of intensive instructional support in mathematics. Schools targeted for recruitment were those that received Title I funding. Principals and the kindergarten teachers of schools were then contacted. All kindergarten teachers in each participating school were eligible to participate in the study.

Schools. This study took place in 36 kindergarten classrooms across two school districts in the metropolitan area of Boston, Massachusetts. Districts A and B had total enrollments of 6,118 and 6,843 students, respectively. Nine schools participated from these two districts. In District A, all kindergarten students attended the same school, whereas eight separate schools from District B participated in the study.

Classrooms. Of the 36 classrooms, 32 provided a full-day kindergarten program, and four classrooms provided a half-day kindergarten program. All classrooms provided 5 days per week of core mathematics instruction in English and had an average of 23.7 students (*SD* = 6.1). Participating classrooms were taught by 36 certified kindergarten teachers. One teacher went on maternity leave during the course of the intervention, and the remaining 35 teachers participated for the duration of the study. Of the 36 teachers, 32 teachers provided the following demographic information. All teachers were female, and the majority of teachers were White (91%). Teachers had an average of 15 years (*SD* = 8.68) of teaching experience and 10 years (*SD* = 7.29) of experience teaching at the kindergarten level. Of the 32 teachers, 66% held a graduate degree in education, and 56% had previously completed at least one graduate course in algebra.

Interventionists. Similar to the Oregon efficacy trial, ROOTS groups were taught by instructional assistants who were employed by the two participating Boston districts. All interventionists were female, and 62% had a bachelor's degree or higher. Among the interventionists, 91% had prior experience with providing small-group instruction, 53% had taken at least one college level course in algebra, 56% were White, and 15% reported their ethnicity as Hispanic.

Criteria for participation. The current study applied the same process for determining students' eligibility for the intervention as the Oregon efficacy trial. To determine eligibility for ROOTS, all participating students from the 36 classrooms, who had parental consent, were screened on the ASPENS (Clarke, Rolfhus, et al., 2012) and NSB (Jordan et al., 2008). Students with both an NSB score of 20 or less and an ASPENS composite score in the "strategic" or "intensive" ranges were considered at risk for MD and eligible for the intervention. The ASPENS and NSB scores of ROOTS eligible students were separately converted into standard scores and then combined to form an overall composite standard score. In each classroom, ROOTS-eligible students' composite standard scores were rank ordered. The 10 ROOTS-eligible students with the lowest composite standard scores were randomly assigned to one of three conditions: (a) a ROOTS-Small (2:1) group, (b) a ROOTS-Large (5:1) group, or (c) a no-treatment BAU control condition. Thus, in each classroom, two students were in the ROOTS-Small group, five students were in the ROOTS-Large group, and three were in the control condition.

Of the 36 classrooms, 26 had at least 10 ROOTS-eligible students per classroom. After random assignment, these 26 classrooms provided 52 ROOTS groups (*n* = 26 ROOTS-Small, *n* = 26 ROOTS-Large). For the 10 classrooms that had fewer than 10 ROOTS-eligible students, we applied a cross-class grouping procedure. For example, in one school, we combined two half-day session classrooms and then randomly assigned the 10 ROOTS-eligible students to one of three conditions: ROOTS-Small, ROOTS-Large, or control. A total of seven ROOTS-Small groups, seven ROOTS-Large groups, and six control groups were formed using the cross-class grouping procedure. In four of these combined classrooms, the ROOTS-Large group had four students instead of five, as only nine eligible students were available after combining classrooms. In one of the combined classrooms, a control group could not be formed, as only six students in the classroom were eligible for ROOTS. In all, the 36 classrooms provided a total of 66 ROOTS groups (n = 33 ROOTS-Small, n = 33 ROOTS-Large).

Students. A total of 878 kindergarten students were screened for ROOTS eligibility in the late fall of 2013. Of these students, 319 met the ROOTS eligibility criteria and were randomly assigned to the ROOTS-Small condition (n = 67), the ROOTS-Large condition (n = 162), or the control condition (n = 90). Similar to the initial ROOTS efficacy trial, the current study combined students in the two ROOTS conditions (small and large) in order to assess the effects of ROOTS intervention as compared to the control condition. Demographic information for all ROOTS-eligible students is presented in Table 1.

Procedures

ROOTS intervention. ROOTS is a Tier 2 kindergarten mathematics intervention program that consists of 50 lessons delivered in 20-min sessions. Similar to the initial study, ROOTS was delivered in small-group formats (two or five students per group), 5 days per week for approximately 10 weeks. However, unlike the initial study, the onset of the ROOTS intervention in the Boston study began in November and ended in March of the 2013–2014 school year, which was an earlier start by nearly 2 months. It is important to note that the intervention occurred at times that did not conflict with students' core mathematics instruction.

Table 1. Descriptive Statistics for Student Characteristics by Condition.

Student characteristic	ROOTS	Control
Age at pretest, *M* (*SD*)	5.2 (0.4)	5.2 (0.4)
Male	47%	58%
Race		
American Indian/Alaskan Native	0%	1%
Asian	1%	2%
Black	6%	7%
Native Hawaiian/Pacific Islander	0%	0%
White	91%	84%
More than one race	2%	6%
Hispanic	49%	51%
Limited English proficiency	23%	28%
SPED eligible	8%	14%

Note. The sample included 229 students in the ROOTS condition and 90 students in the control condition. SPED = special education.

The ROOTS intervention is designed to promote procedural fluency and conceptual understanding in whole-number concepts and skills. This focus is consistent with the Common Core State Standards for mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and calls from expert panels (Gersten et al., 2009). Specifically, ROOTS instruction prioritizes concepts from the Counting and Cardinality and Operations and Algebraic Thinking domains of the CCSS-M. This deep focus on whole-number concepts is designed to help struggling students develop a robust number sense. ROOTS facilitates mathematics proficiency by prioritizing principles of explicit and systematic instruction. For example, lessons judiciously include essential features of explicit mathematics instruction, such as teacher modeling, deliberate practice, visual representations of mathematics, and academic feedback. ROOTS also incorporates opportunities for students to verbalize their mathematical thinking and discuss problem-solving methods. More information on ROOTS is described in Clarke et al. (2016).

Professional development. Similar to the Oregon efficacy trial, all participating interventionists received two 5-hr professional development workshops delivered by project staff. The first workshop focused on the instructional objectives and content of Lessons 1 to 25, empirically validated instructional practices in mathematics, and small-group management techniques. The second workshop focused on Lessons 26 to 50. Both workshops incorporated opportunities for interventionists to practice and receive feedback on lesson delivery from coaches and project staff. During intervention implementation, all interventionists received instructional support from two ROOTS coaches. The two coaches were former educators with specialized knowledge and training in effective early mathematics instruction and small-group instructional practices. Coaching visits included direct observations of lesson delivery followed by feedback on the quality of instruction as well as fidelity of intervention implementation.

To ensure a high level of implementation fidelity, the number of coaching visits varied based on interventionists' implementation needs. Although interventionists received an average of 2.3 (*SD* = 0.98) coaching visits over the course of the study, 22 interventionists demonstrated implementation difficulties and thus received three or four coaching visits. The remaining interventionists had no difficulties with the intervention and thus received one or two visits.

Control condition. Core (Tier 1) math instruction delivered in the kindergarten classroom served as the control condition or counterfactual. All ROOTS (small and large) and control students received daily core mathematics instruction. The control condition was documented through teacher surveys and direct observations of core mathematics instruction. Observation and survey data indicated that classrooms in District A primarily used the Scott Foresman mathematics curriculum during core math instruction, whereas classrooms in District B primarily used the enVisionMath curriculum. Based on the level of available evidence, the What Works Clearinghouse (WWC; n.d.) has rated both of these core mathematics programs as effective for improving student mathematics achievement. Teachers also reported that they supplemented these curricula with their own materials. Observations indicated that ROOTS intervention materials were not used during core mathematics instruction.

Kindergarten classroom teachers reported that they provided an average of 46 min per day of core mathematics instruction (*SD* = 13.64). Survey data also indicated that students received whole-number instruction during calendar time. All teachers reported that core instruction included activities on counting and cardinality as well as numbers and operations in base 10. The majority of teachers (79%) indicated that knowing number names and the count sequence was the main instructional priority when teaching whole-number concepts and skills, and 17.2% reported that their core instruction emphasized counting to tell the number of objects. All teachers provided teacher-led instruction, and the majority of teachers also reported that core mathematics instruction included peer or group work, independent student work, and learning centers. The majority of the teachers reported that small-group formats (93%) and individualized instruction (66%) were used during core mathematics instruction. Finally, teachers reported that they regularly incorporated principles of explicit instruction, such as demonstrations of mathematics concepts, visual representations, and opportunities for mathematics verbalizations.

In each participating classroom, trained research staff conducted direct observations of core mathematics instruction. Observation data indicated that all teachers provided teacher-led mathematics instruction, with some teachers providing a variety of other learning opportunities, such as independent practice, peer and small-group activities, and technology-based activities. The primary mode of instructional delivery was teacher-led instruction (92%), and nearly 75% of core instruction focused on operations and algebraic thinking (e.g., addition and subtraction), whereas 17.9% of instructional periods primarily focused on counting. The majority of observations showed clear evidence of the following principles of explicit and systematic instruction (Gersten et al., 2009): academic feedback, visual representations, teacher demonstrations, guided practice opportunities, and opportunities for students to verbalize their mathematical thinking. However, observations documented that teachers were less likely to provide independent and written mathematics practice and scaffolded instruction for struggling students.

Fidelity of Implementation

In order to determine the extent to which the ROOTS intervention was delivered as intended, fidelity of ROOTS implementation was directly observed by trained research staff. Each ROOTS group was observed three times over the course of the intervention. A total of 190 observations of implementation fidelity were conducted, of which 31 included two observers. It is important to note that these observations were separate from the fidelity observations conducted by the ROOTS coaches. For these fidelity observations, observers used a 4-point scale (4 = *all*, 3 = *most*, 2 = *some*, 1 = *none*) to rate the extent to which interventionists met the lesson's instructional objectives, followed the provided teacher scripting, and used the prescribed math models for that lesson. Observers also recorded the number of prescribed activities delivered during the lesson. Interventionists were observed to deliver the majority of the prescribed activities (*M* = 4.26 out of 5 activities, *SD* = 0.34). Observers also noted that interventionists met lesson objectives (*M* = 3.46, *SD* = 0.46), followed teacher scripting (*M* = 3.37, *SD* = 0.48), and used prescribed math models (*M* = 3.58, *SD* = 0.42).

To describe interobserver reliability for the fidelity measures, intraclass correlation coefficients (ICCs) were calculated. The ICC for the aggregate fidelity score was substantial (0.87). Further, ICCs for individual fidelity ratings indicated substantial to perfect agreement: (a) 1.0 for number of activities delivered, (b) 0.68 for met math objectives, (c) 0.79 for followed teacher scripting, and (d) 0.87 for used prescribed math models (Landis & Koch, 1977).

ROOTS coaches also used direct observations to rate fidelity of implementation. Each ROOTS group was observed and rated between two and four times. During these observations, coaches rated the quality of student–teacher interactions and fidelity of implementation on separate 5-point scales (1 = *low*, 3 = *medium*, 5 = *high*) as well as the duration of each lesson in minutes. On average, ROOTS lessons were delivered in 22.36 min (*SD* = 3.87). Coaches rated overall implementation fidelity in the medium to medium-high range, with variability across interventionists (*M* = 3.72, *SD* = 0.92). Quality of student–teacher interactions was also rated in the medium-high range (*M* = 4.02, *SD* = 0.86).

Intervention dosage was also recorded as an indicator of implementation fidelity. Of the 66 ROOTS groups, 62 groups provided a full or nearly full dose of the intervention (98% to 100% of the lessons). Four of the groups did not provide information on lesson completion.

Measures

All treatment and control students were administered five measures of whole-number understanding at pretest (T1) and posttest (T2). These measures consisted of one proximal measure that measured skills taught in ROOTS, a set of curriculum-based measures that assessed discrete skills related to early number sense, and two distal measures of whole-number understanding. An additional distal outcome measure was administered approximately 6 months into students' first-grade year (T3). Trained research staff administered all student measures. Interscorer reliability criteria were met for all assessments (i.e., >.95).

RAENS (Doabler et al., 2012) is a researcher-developed instrument that was administered at T1 and T2 time periods. RAENS is individually administered and consists of 32 items. Items assess aspects of counting and cardinality, number operations, and the base-10 system. In an untimed setting, students are asked to count and compare groups of objects; write, order, and compare numbers; label visual models (e.g., 10-frames); and write and solve single-digit addition expressions and equations. RAENS's predictive validity ranges from .68 to .83 for the TEMA-3 and the NSB. Interrater scoring agreement is reported at 100% (Clarke et al., 2016).

Oral counting–early numeracy curriculum-based measurement (Clarke & Shinn, 2004), a curriculum-based measure, which was administered at T1 and T2, has students orally count in English for 1 min, and the discontinue rule applies after the first counting error. The highest correct number counted represents a student's score. Fall-of-kindergarten oral counting skills have predictive validity with SAT-10 performance in the spring of kindergarten as well as high interscorer and test-retest reliability. Test-retest reliability and alternate-form reliability are reported at above .80, concurrent validity is reported as ranging from .49 to .70, and predictive validity with the Woodcock-Johnson Applied Problems subtest (Woodcock & Johnson, 1989) and the Number Knowledge Test (Okamoto & Case, 1996) is reported as ranging from .46 to .72 (Clarke et al., 2011).

ASPENS (Clarke et al., 2012) is a set of three curriculum-based measures validated for screening and progress monitoring in kindergarten mathematics (Clarke et al., 2012). Each 1-min 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 et al., 2012). ASPENS was administered at T1 and T2.

NSB (Jordan et al., 2008) is an individually administered measure with 33 items that assess counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations. NSB, which was administered at T1 and T2, has a coefficient alpha of .84 (Jordan et al., 2008).

TEMA-3 (Ginsburg & Baroody, 2003) is a standardized, norm-referenced, individually administered measure of beginning mathematical ability. The TEMA-3 assesses whole-number understanding for children ranging in age from 3 to 8 years 11 months. Alternate-form and test-retest reliabilities of the TEMA-3 are .97 and .93, respectively. The TEMA-3, which was administered at T1 and T2, has concurrent validity with other mathematics measures ranging from .54 to .91.

The SAT-10 (Harcourt Brace Educational Measurement, 2002) and the SESAT (Harcourt Brace Educational Measurement, 2003) are both group-administered, standardized, norm-referenced measures. Both measures have two mathematics subtests: Problem Solving and Procedures. The internal consistency for the SESAT is .88. Internal consistency reliabilities for the SAT-10 mathematics subtests range from .80 to .90. All treatment and control students were administered the SESAT at posttest (T2) and the SAT-10 midway through their first-grade year (T3).

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 et al., 2011; Bauer et al., 2008). The study design called for the randomization of individual students to receive ROOTS, nested within small 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, whereas the unclustered controls did not. Further, because the residual variances may have differed between conditions, we tested the assumption of homoscedasticity of residuals.

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 (T1) to spring (T2) of kindergarten. The basic model includes time, T, coded 0 at T1 and 1 at T2; condition, C, coded 0 for control and 1 for ROOTS; and the interaction between the two:

$$\begin{aligned} Y_{ij} &= \pi_{0j} + \pi_{1j}C_j + \pi_{2j}T_{ij} \\ &\quad + \pi_{3j}T_{ij}C_j + e_{ij} \qquad e_{ij} \sim N\left(0, \sigma^2\right) \end{aligned} \tag{1}$$

$$\pi_{0j} = \beta_{00} \tag{2}$$

$$\pi_{1j} = \beta_{10} \tag{3}$$

$$\pi_{2j} = \beta_{20} \tag{4}$$

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

For Level 1, the first equation, Y_{ij} represents a score for individual i within cluster j, and the model includes condition, C_j; time, T_{ij}; and their interaction as predictors. Although the e_{ij} are distributed $N(0, \sigma^2)$, the Time × Condition analysis decomposes the individual-level variance, σ^2, into a variance for the assessments, σ_s^2, and covariance between the T1 and T2 assessments, r_s^2, where σ_s^2, and r_s^2 sum to σ^2 (Murray, 1998). For simplicity, 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:

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

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:

$$\mathrm{V}\left(e_{ij} \mid C_j = 0\right) = \sigma_0^2 \tag{7}$$

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

As described previously, both σ_0^2 and σ_1^2 decompose into a variance and covariance in the analyses.

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 & Dickenson, 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 earlier to account for clustering within classrooms or schools. Aligned with the observations of Raudenbush and Sadoff (2008), the overall pattern of intervention results remained similar in all models, whether or not we included classroom or school levels in the model. Condition effects did not vary by classroom or school, either, so we omitted these results.

Model estimation. We fit models to our data with SAS PROC MIXED Version 9.2 (SAS Institute, 2009) using restricted maximum likelihood, generally recommended for multilevel models (Hox, 2002). Maximum likelihood estimation for the Time × Condition analysis uses of all available data to provide potentially unbiased results even in the face of substantial attrition, provided the missing data were missing at random (Graham, 2009). 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 or left the school.

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 2.3. However, multilevel regression methods have been found quite robust to violations of normality (e.g., Hannan & Murray, 1996).

Effect sizes. To ease interpretation, we computed an effect size, Hedges' *g* (Hedges, 1981), for each fixed effect. Hedges' *g*, recommended by the WWC (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 condition and assessment time. Overall, 35% of the sample scored below the 10th percentile on the TEMA-3 at pretest (38% of control, 34% of ROOTS). At posttest, 22% of the sample remained below the 10th percentile (32% of control, 17% of ROOTS).

Attrition

Student attrition was defined as students with data at T1 but missing data at T2. Attrition rates varied between 6%, for the TEMA-3 and RAENS, and 8%, for NSB, ASPENS, and oral counting. Only 5% of students were missing all posttest data. The proportion of students missing all posttest data did not differ between conditions, $\chi^2(1) = 3.61$, $p = .0575$. 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 any of the outcome measures ($p \geq .2333$) and no pretest differences between conditions for students with posttest data (see Table 3). The primary impact analyses incorporated all available data, further reducing the likelihood of bias (Graham, 2009).

Efficacy Effects for ROOTS

Table 3 presents the results of the statistical models. The table presents the results of the homoscedastic model for all outcomes, as it was deemed equivalent to the more complicated heteroscedastic model. The bottom two rows of the table show the likelihood ratio test results that compared homoscedastic residuals to heteroscedastic residuals. Although the variance structures differed between these models, the condition effect estimates and statistical significance values

Table 2. Descriptive Statistics for Mathematics Measures by Condition and Assessment Time.

	T_1		T_2		T_3	
Measure	ROOTS	Control	ROOTS	Control	ROOTS	Control
NSB						
M (*SD*)	12.66 (3.71)	11.89 (3.35)	19.71 (4.90)	17.01 (4.81)		
n	229	90	208	86		
ASPENS						
M (*SD*)	21.67 (17.06)	17.80 (14.85)	89.10 (33.60)	63.97 (35.00)		
n	225	90	208	86		
Oral counting						
M (*SD*)	19.80 (12.93)	18.20 (11.49)	45.83 (21.56)	41.31 (21.41)		
n	229	90	207	85		
TEMA-3						
M (*SD*)	17.08 (7.03)	16.23 (6.67)	26.48 (7.62)	23.27 (8.06)		
n	227	88	213	88		
RAENS						
M (*SD*)	11.83 (5.74)	11.39 (5.74)	24.31 (6.01)	17.34 (5.93)		
n	228	89	213	88		
SESAT Total						
M (*SD*)			463.75 (38.51)	451.90 (34.24)		
n			200	82		
SAT-10 Total						
M (*SD*)					497.09 (29.07)	495.45 (27.55)
n					186	77

Note. NSB = Number Sense Brief (Jordan et al., 2008); ASPENS = Assessing Student Proficiency in Early Number Sense (Clarke et al., 2012); TEMA-3 = Test of Early Mathematics Ability (3rd ed.; Ginsburg & Baroody, 2003); RAENS = ROOTS Assessment of Early Numeracy Skills (Doabler et al., 2012); SESAT = Stanford Early School Achievement Test (Harcourt Brace Educational Measurement, 2003); SAT-10 = Stanford Achievement Test Series (10th ed.; Harcourt Brace Educational Measurement, 2002). The sample sizes represent students with a particular measure at each assessment period.

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 3 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 ($p \geq .1573$, Hedges' $g \leq 0.16$ for all measures), which suggested that ROOTS and control students were similar in the fall of kindergarten. We found statistically significant differences by condition in gains from fall to spring for five dependent variables. Students in the ROOTS condition made greater gains than control students on the NSB ($t = 3.15$, $df = 94$, $p = .0022$), ASPENS ($t = 5.60$, $df = 118$, $p < .0001$), TEMA-3 standard scores ($t = 3.45$, $df = 99$, $p = .0008$), and RAENS ($t = 9.20$, $df = 126$, $p < .0001$). Students in the ROOTS condition also had higher covariate-adjusted SESAT scores using ASPENS and TEMA-3 as pretest covariates ($t = 2.45$, $df = 136$, $p = .0158$). We did not detect statistically significant differences between conditions in gains on oral counting or differences between conditions on covariate adjusted SAT-10 scores with ASPENS and TEMA-3 as pretest covariates. The Time × Condition model estimated differences in gains between conditions of 1.94 for the NSB ($g = 0.40$), 21.78 for the ASPENS ($g = 0.64$), 2.43 for the TEMA-3 standard score

Table 3. Time × Condition Analysis With Fall-to-Spring Gains in Mathematics.

Variable	NSB	ASPENS	Oral counting	TEMA-3	RAENS
Fixed effects					
Intercept	11.89**** (0.44)	17.80**** (2.75)	18.20**** (1.83)	16.23**** (0.77)	11.39**** (0.61)
Time	5.17**** (0.48)	46.33**** (3.10)	23.13**** (2.18)	7.05**** (0.56)	5.89**** (0.56)
Condition	0.77 (0.55)	3.62 (3.38)	1.62 (2.31)	0.81 (0.92)	0.48 (0.74)
Time × Condition	1.94** (0.62)	21.78**** (3.89)	3.43 (2.83)	2.43*** (0.70)	6.50**** (0.71)
Variances					
Gains between ROOTS groups	2.55* (1.03)	62.13 (37.02)	53.97*** (15.70)	2.25 (1.50)	2.34* (1.19)
Pre/post covariance	7.41**** (1.32)	263.06**** (47.97)	95.48**** (17.50)	38.62**** (4.06)	19.59**** (2.45)
Residual	7.28**** (0.70)	357.53**** (34.49)	152.79**** (16.10)	11.51**** (1.23)	11.49**** (1.11)
Time × Condition					
Hedges' *g* [95% CI]	0.398 [0.150, 0.646]	0.641 [0.416, 0.865]	0.159 [–0.099, 0.417]	0.314 [0.136, 0.492]	1.085 [0.854, 1.317]
p values	.0022	<.0001	.2276	.0008	<.0001
df	94	118	192	99	126
Likelihood ratio χ^2	2.57	0.20	0.32	0.36	3.11
p value	.2766	.9039	.8510	.8354	.2108

Note. NSB = Number Sense Brief (Jordan et al., 2008); ASPENS = Assessing Student Proficiency in Early Number Sense (Clarke et al., 2012); TEMA-3 = Test of Early Mathematics Ability (3rd ed.; Ginsburg & Baroody, 2003); RAENS = ROOTS Assessment of Early Numeracy Skills (Doabler et al., 2012); CI = confidence interval. Table entries show parameter estimates with standard errors in parentheses except for Hedges' *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.

*p < .05. **p < .01. ***p < .001. ****p < .0001.

($g = 0.31$), and 6.50 for the RAENS ($g = 1.08$). The analyses-of-covariance model estimated differences in covariate-adjusted SESAT scores between conditions of 8.95 ($g = 0.24$).

Discussion

This study conducted a closely aligned conceptual replication to further investigate the efficacy of the ROOTS intervention program 1 year after the initial study. Participating kindergarten classrooms in the current study were from a different geographical region and offered a different instructional context for the counterfactual than the original study. Overall, statistically significant effects on the TEMA-3, ASPENS, and RAENS were replicated in the Boston study. The effect sizes obtained for these measures ranged from 0.31 to 1.08. In the Oregon study, Clarke and colleagues (2016) failed to find statistically significant differences on the NSB or the SESAT. Notably, however, results from the Boston study indicated a statistically significant impact on these two distal outcome measures. Effect sizes for the NSB and the SESAT, respectively, were .40 and .24. With exception of the RAENS, all effect sizes fell within the confidence intervals of the Oregon study. The effect size for the Boston RAENS exceeded the upper bound of the effect obtained in the Oregon study.

These findings are noteworthy because the replication's control condition, relative to the Oregon study, included programs that had a stronger evidentiary basis. In fact, the two primary core programs (enVisionMATH and Scott Foresman) were reviewed by the WWC (n.d.) and rated as effective for improving student mathematics achievement. These differences in instructional dimensions (i.e., mathematics programs) were deemed as "raising the bar" of the counterfactual for the current study. The results are also important because the Boston study included a more ethnically diverse student sample. Demographic data indicated a significant increase (18%) of students whose reported ethnicity was Hispanic. By diversifying the sample, effects of ROOTS may be generalized to a broader range of students.

One plausible explanation for the statistically significant differences between the ROOTS and control conditions in the replication study is the timing of the intervention. Onset of ROOTS in the Boston study occurred at a different time point in the school year than the Oregon study. Whereas students in the initial study received ROOTS starting in mid-January, the replication began nearly 2 months earlier (mid-November). Based on the foundational concepts and skills addressed in ROOTS, it is plausible that the intervention has more beneficial effect on mathematics outcomes when delivered earlier in students' kindergarten year.

To obtain a broader perspective of replication, researchers may need to consider whether there is a "decline effect" with students' response to treatment.

In addition to our formal investigation of treatment effects, we were also interested in exploring whether rates of treatment responsiveness among ROOTS students were similar across the Oregon and Boston studies. Although this was not formally tested, we found similar percentages of ROOTS students from the Oregon and Boston studies tested below the 10th percentile on the TEMA-3 at both pretest (34% and 29%) and posttest (17% and 16%). Our interest in these percentages is based on the idea of invoking the field to bring treatment responsiveness into the replication conversation. To obtain a broader perspective of replication, researchers may need to consider whether there is a "decline effect" (Cook, 2014) with students' response to treatment. Rates of treatment response equal to or better than those obtained in earlier investigations may serve as an additional indicator of replication. A decline of treatment effects for at-risk subgroups (e.g., English learners), however, may suggest a lack of replication. Moreover, it may indicate a shift or strengthening of the counterfactual (Lemons et al., 2014). Our informal explorations suggest similar percentages of treatment responsiveness in the two ROOTS studies.

Implications for Practice and Research

Findings from the current study align with the growing knowledge base of effective mathematics instruction that suggests high concentrations of whole-number instruction delivered via systematic and explicit mathematics interventions are crucial for students who struggle to develop mathematics proficiency (Bryant et al., 2011; Gersten et al., 2009; Sood & Jitendra, 2013). A profound understanding of whole numbers in the early grades has strong implications for a child's success in later mathematics. Therefore, results obtained by our study and those found by previous mathematics intervention research showcase the need to privilege the development of number sense among young students who face difficulties in mathematics.

Rates of treatment response equal to or better than those obtained in earlier investigations may serve as an additional indicator of replication.

We also believe the current study has implications for future replication research. At a broad level, this study supports the crucial role replication has in educational research. Replication studies, when conceptualized and implemented well, are valuable for ruling out chance findings obtained in previous research. We regard findings from the current study as replicated evidence of the ROOTS intervention's impact on student outcomes. As recommended by Coyne et al. (2016), a logical next step for our research on ROOTS would be to conduct a series of increasingly distal replication studies. Such replication studies will enable greater generalization regarding ROOTS' efficacy and also allow us to systematically explore various components of the intervention theorized to influence student outcomes.

For example, roughly one fifth of ROOTS students remained in what could be described as a severe risk category (i.e., below the 10th percentile on the TEMA-3 at posttest). Persistent findings related to low response or nonresponse to generally effective interventions continue to challenge the field of academic intervention research (Fuchs, Fuchs, & Compton, 2012). The findings of the current study fit this pattern. Therefore, having begun to establish overall intervention efficacy, a next series of studies might investigate whether complementary features to the ROOTS intervention, such as components that target students' cognitive processes (e.g., attention), bolster treatment impact and improve student response.

This study also adds support for situating replication studies within larger frameworks of research. For example, given that the duration of Goal 3 efficacy projects funded by the Institute of Education Sciences (IES) is typically 4 years, it is possible that researchers can conduct a series of a priori replication studies within these large-scale projects. When conceptualizing the larger, IES-funded ROOTS efficacy project (Clarke, Doabler, Fien, Baker, & Smolkowski, 2012), we decided to ground it in a systematic framework of replication (Coyne et al., 2016). Our aim was to accumulate converging evidence in support of the intervention across four separate cohorts of kindergarten students from two different geographical regions. We acknowledge that when planning for replication studies in large-scale efficacy projects, factors such as statistical power and available resources must be taken into full consideration. However, if a series of replications is planned well, rigorous and trustworthy evidence in support of an intervention can be obtained within a relatively short time frame (Coyne et al., 2016).

Finally, the recommendations for systematic replication frameworks proposed by Coyne et al. (2016) yield a cogent argument that the field of special education needs to delineate standards for replication research. Fields such as prevention science have endorsed replication standards for efficacy, effectiveness, and scale-up research (Flay et al., 2005; Gottfredson et al., 2015). Of foremost importance, standards in the field of special education would establish the basis for the design and implementation of replication research. Moreover, such standards would increase consistency in the way researchers conceptualize, operationalize, and disseminate replication studies. Currently, replication research in special education markedly differs on how it is classified for dissemination (Cook, Collins, Cook, & Cook, 2016; Lemons et al., 2016).

Limitations

Results from the current study are specific to the targeted sample. Therefore, additional replications should be conducted in different geographical areas and with more diverse samples to increase confidence in the generalizability of findings. Additionally, the partially nested design used for this study has weakened internal validity compared to fully clustered or unclustered RCTs. Although the randomly assigned students represent potential outcomes for each other in all background characteristics (Rubin, 1974, 2005), the two groups may differ in postassignment clustering effects (Bauer et al., 2008). We have neither a theoretical rationale nor empirical evidence that simply clustering kindergarten students in small groups would lead to improved math outcomes, but we cannot rule out such an effect in the present study because the difference in clustering across conditions has been confounded with the intervention delivery. On the other hand, the external validity may be stronger. In our schools, students at risk for math difficulties typically receive no supplemental instruction in kindergarten, so the ungrouped control students likely represent the most appropriate contrast condition. The control group experience for this study was chosen to represent most classrooms in the states where the project took place. That said, a limitation in the present study is that ROOTS students received substantially more mathematics instruction than their control peers. This additional time may explain some of the improvement. Additionally, because we simultaneously varied three important elements (i.e., location, timing of intervention, and instructional context of the counterfactual), it is difficult to determine which element may account for the differences in findings between the current study and the original investigation.

The possibility of unintentional bias or potential repetition of mistakes in study logic or procedures because of author overlap is an additional limitation (Makel & Plucker, 2014). Author overlap occurs when the same research team is responsible for and carries out replications of their original research (Coyne et al., 2016). The same research team conducted the original and current investigations of ROOTS. Concerns with author overlap in the current study are largely mitigated given that an independent evaluator conducted the statistical analysis and an external entity from the Boston metropolitan area collected all student achievement and direct observation data. It is important to note that our research team oversaw these data collection efforts. Additionally, both ROOTS RCTs used school-based personnel to implement the intervention. We believe this approach more closely resembles the conditions studied in large-scale evaluation projects. Efficacy trials typically have research team members implement the intervention in order to provide a more controlled "ideal" setting and maximize treatment effects. Thus, our estimates of intervention impact may be closer in approximation to routine classroom conditions.

An additional limitation to the findings may be the possibility of obtaining a novelty effect wherein the impact we observed was due to the uniqueness or novelty of the experiment and intervention for the research sites. Neither Boston school district had previously used a mathematics intervention in kindergarten. Thus it may be that impact resulted from the newness and excitement related to implementing a mathematics intervention, resulting in changes to other practices (e.g., renewed focus on mathematics in core instruction) that would have translated into positive student outcomes. The plausibility of the novelty effect is lessened by the similar results found across studies and could be further attenuated by additional replications in sites that have previous experience in implementing mathematics interventions at the kindergarten level.

Conclusion

Although there is consensus among the research community on the importance of replication, recent reviews of the special education research literature suggest that of the studies that demonstrate features of replication, few use identifiers such as *replicate* or *replication* (Cook et al., 2016; Lemons et al., 2016). Consequently, this may represent a missed opportunity for the field of special education to systematically couch a continuum of intervention research within a larger replication framework (Coyne et al., 2016). In this study, we sought to replicate the beneficial treatment effects of a Tier 2 mathematics intervention on the mathematics outcomes of students with MD using the same methodological and analytical procedures applied in the initial study. Overall, we found that intervention students outperformed their control peers on all student outcome measures. At a molecular level, these results suggest that the treatment effects of ROOTS replicated in classrooms from a different geographical area and with a different instructional context. More globally, we hope the integration of systematic replication into our line of research provides further support for the importance of replication in advancing educational research and practice.

References

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. (2010). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford.

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

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

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). U.S. Department of Education, Institute of Education Sciences, Special Education Research, CFDA No. 84.324A, 2012-2016, Funding No. R324A120304.

Clarke, B., Doabler, C.T., 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*. Advance online publication. doi:10.1080/19345747.2015.1116034

Clarke, B., Rolfhus, E., Dimino, J., & Gersten, R. M. (2012). *Assessing Student Proficiency of Number Sense (ASPENS)*. Longmont, CO: Cambium Learning Group, Sopris Learning.

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

Clarke, B., Smolkowski, K., Baker, S. K., Fien, H., Doabler, C. T., & Chard, D. J. (2011). The impact of a comprehensive tier 1 core kindergarten program on the achievement of students at-risk in mathematics. *Elementary School Journal*, *111*, 1–24. doi:10.1086/659033

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cook, B. G., Collins, L. W., Cook, S. C., & Cook, L. (2016). A replication by any other name . . . : A systematic review of replicative studies. *Remedial and Special Education*, *37*, 223–234. doi:10.1177/0741932516637198

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

Coyne, M. D., Little, M., Rawlinson, D. A., Simmons, D., Kwok, O. M., Kim, M., . . . Civetelli, C. (2013). Replicating the impact of a supplemental beginning reading intervention: The role of instructional context. *Journal of Research on Educational Effectiveness*, *6*, 1–23. doi:10.1080/19345747.2012.706694

Dasgupta, A., Lawson, K. A., & Wilson, J. P. (2010). Evaluating equivalence and noninferiority trials. *American Journal of*

Health-System Pharmacy, *67*, 1337–1343. doi:10.2146/ajhp090507

Doabler, C. T., Clarke, B., & Fien, H. (2012). *ROOTS Assessment of Early Numeracy Skills*. Unpublished measure, University of Oregon, Eugene.

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

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175. doi:10.1007/s11121-005-5553-y

Fuchs, L.S., Fuchs, D., & Compton, D. (2012). The early prevention of mathematics difficulty: Its power and limitations. *Journal of Learning Disabilities*, *45*, 257–269. doi:10.1177/0022219412442167

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: U.S Department of Education, National Center for Education Evaluation and Regional Assistance.

Ginsburg, H., & Baroody, A. (2003). *Test of Early Mathematics Ability–Third Edition*. Austin, TX: Pro-Ed.

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*, 893–926. doi:10.1007/s11121-015-0555-x

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 Brace Educational Measurement. (2002). *Stanford Achievement Test (SAT-10)*. San Antonio, TX: Harcourt Brace Jovanovich.

Harcourt Brace Educational Measurement (2003). *Stanford Early School Achievement Test* (10th ed.). San Antonio, TX: Harcourt Brace Jovanovich.

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.

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., 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.

Kromrey, J. D., & Dickenson, 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-treatment 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

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*, 242–252. doi:10.3102/0013189X14539189

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., Gajjar, S. A., & Sacks, L. H. (2016). An inadvertent concurrent replication: Same roadmap, different journey. *Remedial and Special Education*, *37*, 213–222. doi:10.1177/0741932516631116

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

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher*, *43*, 304–316. doi:10.3102/0013189X14545513

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research how often do they really occur?. *Perspectives on Psychological Science*, *7*, 537–542. doi:10.1177/1745691612460688

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

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

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for mathematics*. Washington, DC: Author. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. In R. Case, Y. Okamoto, S. Griffin, R.S. Siegler, & D. P. Keating (Eds.), *The role of central conceptual structures in the development of children's thought: Monographs of the Society for Research in Child Development* (pp. 27–58). Chicago, IL: Society for Research in Child Development.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., & Evans, S. W., for the CONSORT Group. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *Journal of the American Medical Association*, *295*, 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. 10.1080/19345740801982104

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.). San Francisco, CA: McGraw-Hill.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701. doi:10.1037/h0037350

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331. doi:10.1198/016214504000001880

SAS Institute. (2009). *SAS/STAT 9.2 user's guide* (2nd ed.). Cary, NC: Author. Retrieved from the http://support.sas.com/documentation/index.html

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi:10.1037/a0015108

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

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

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117. doi:10.1007/s11121-011-0217-6

van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). New York, NY: Wiley.

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/

What Works Clearinghouse. (n.d.). *What works in math*. Retrieved from http://www.ies.ed.gov

Woodcock, R., & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised*. Chicago, IL: Riverside.

Authors' Note

Funding was provided by Institute of Education Sciences 10.13039/100005246 R324A120304. Drs. Christian Doabler, Ben Clarke, Scott Baker, 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.

Manuscript received February 2016; accepted June 2016.