*Article*

Hammill Institute on Disabilities

Journal of Learning Disabilities
2019, Vol. 52(2) 168–180

DOI: 10.1177/0022219418789376
journaloflearningdisabilities.sagepub.com
SAGE

# Examining the Impact of Group Size on the Treatment Intensity of a Tier 2 Mathematics Intervention Within a Systematic Framework of Replication

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

## Abstract

Group size and treatment intensity are understudied topics in mathematics intervention research. This study examined whether the treatment intensity and overall intervention effects of an empirically validated Tier 2 mathematics intervention varied between intervention groups with 2:1 and 5:1 student-teacher ratios. Student practice opportunities and the quality of explicit instruction served as treatment intensity metrics. A total of 465 kindergarten students with mathematics difficulties from 136 intervention groups participated. Results suggested comparable performances between the 2:1 and 5:1 intervention groups on six outcome measures. Observation data indicated that student practice differed by group size. Students in the 5:1 groups received more opportunities to practice with their peers, while students in the 2:1 groups participated in more frequent and higher quality individualized practice opportunities. Implications in terms of delivering Tier 2 interventions in small-group formats and engaging at-risk learners in meaningful practice opportunities are discussed.

## Keywords

treatment intensity, group size, student practice opportunities, explicit mathematics instruction, mathematics difficulties

Within multitiered approaches to mathematics instruction and Response to Intervention (RtI) frameworks (Fuchs & Vaughn, 2012), *treatment intensity* is generally conceptualized as an alterable variable that can be purposefully manipulated to maximize student learning and obtain an optimal level of instruction (Warren, Fey, & Yoder, 2007). For example, in a three-tiered model, if a student does not adequately respond to Tier 1 mathematics instruction, the goal of Tier 2 is to provide a more systematic, intensive experience. A similar increase in intensity is conceptualized in moving from Tier 2 to Tier 3.

Instructional time is often recognized as a variable of treatment intensity, and recently, researchers have focused on ways to intensify mathematics instruction by increasing factors of time, such as the amount of time spent in each session and the number of days taught per week. Bryant et al. (2011) systematically increased the amount of instructional time across a series of mathematics intervention studies. In their most recent study, Bryant and colleagues (2011) found that increased intervention time was a decisive factor in improving the mathematics achievement of students with mathematics difficulties (MD).

More recently, Codding et al. (2016) investigated variations of treatment dosage of a small-group intervention focused on whole number operations. A total of 101 second-, third-, and fourth-grade students with MD were randomly assigned to one of the three dosage conditions (sessions once, twice, or four times per week) or a control condition. Findings from a proximal outcome measure suggested that students taught in the higher dosage small

[1]University of Texas at Austin, Austin, TX, USA
[2]University of Oregon, Eugene, OR, USA
[3]Oregon Research Institute, Eugene, OR, USA
[4]Southern Methodist University, Dallas, TX, USA

**Corresponding Author:**
Christian T. Doabler, PhD, University of Texas at Austin, 1 University Station, D5300 SZB 408B, Austin, TX 8712, USA.
Email: cdoabler@austin.utexas.edu

groups (i.e., four sessions per week) outperformed their peers in the control and other two dosage conditions.

Existing frameworks of treatment intensity also consider group size or the instructional format in which interventions are delivered as an effective way to increase treatment intensity (e.g., Codding & Lane, 2015). Group size is an instructional variable that is backed by substantial empirical evidence, particularly in the area of early literacy (Elbaum, Vaughn, Moody, Hughes, & Moody, 2000). Although the use of small-group instruction in mathematics does not have the same level of empirical support as in reading, a common recommendation among experts in the area of mathematics intervention research is to decrease group size to increase the intensity of the learning experience for students with MD (e.g., Gersten et al., 2009). Yet few studies have examined the effect of group size on the treatment intensity of early mathematics interventions.

Clarke et al. (2017) conducted a recent randomized controlled trial (RCT) in which they experimentally investigated the impact of group size on the treatment intensity and student mathematics outcomes in the context of ROOTS, an evidence-based Tier 2 kindergarten mathematics intervention (Clarke et al., 2016; Doabler et al., 2016). Participating in the study were approximately 600 students from 60 kindergarten classrooms in Oregon. These kindergarten students represented the first two cohorts of the larger, federally funded ROOTS Efficacy Project. Aligned with other Tier 2 mathematics interventions (Bryant et al., 2011), the ROOTS program is delivered in small-group formats and is designed to promote number sense development among students with MD. To build students' conceptual understanding of and procedural fluency with whole numbers and operations, the 50-lesson ROOTS program centers on a systematic and explicit instructional design framework (Archer & Hughes, 2011; Gersten et al., 2009). In this way, the intervention engages students in purposefully planned and explicitly delivered mathematics tasks and activities.

To examine the effect of group size on the intervention impact and treatment intensity of ROOTS, Clarke et al. (2017) focused on two treatment conditions. In one condition, students were randomly assigned to receive the ROOTS intervention in groups with 2:1 student-teacher ratios (2:1 ROOTS groups), whereas the other condition provided ROOTS in groups with 5:1 student-teacher ratios (5:1 ROOTS groups). Random assignment resulted in 60 and 59 ROOTS 2:1 and 5:1 intervention groups, respectively. Students in both treatment conditions continued to receive core mathematics instruction.

Since the ROOTS intervention was purposefully designed to deeply engage students in foundational whole number concepts and skills, Clarke et al. (2017) used the frequency of student practice opportunities as a metric of treatment intensity. The research suggests that frequent practice opportunities are critical for fostering mathematics proficiency among students with and without MD (Clements, Agodini, & Harris, 2013; Doabler et al., 2015; Gersten et al., 2009). Similar to other explicit mathematics interventions (Bryant et al., 2011; Sood & Jitendra, 2013), the ROOTS intervention offers students guided practice opportunities to promote a high success rate with the targeted mathematical content. In ROOTS, such practice opportunities consist of individual students or the group at large working with concrete representations of mathematical ideas and engaging in mathematics verbalizations. Group response opportunities allow all students to practice in unison, whereas individualized practice permits an opportunity for one student to convey or demonstrate her mathematical thinking, understanding, and reasoning.

While the ROOTS intervention was designed to provide intensive learning experiences regardless of group size, Clarke et al. (2017) hypothesized that the 2:1 ROOTS groups would demonstrate stronger treatment effects and receive more opportunities to practice than the 5:1 ROOTS groups based on the lower student-to-teacher ratio. Clarke et al. (2017) reported nonsignificant differences in student mathematics outcomes when comparing the 2:1 and 5:1 ROOTS groups. Findings suggested that the impact of ROOTS was essentially the same regardless of whether students participated in the 2:1 or 5:1 groups. Results for their second hypothesis, however, indicated that the frequency of learning or practice opportunities students received differed by group size. Whereas students in the 2:1 ROOTS groups received more opportunities to practice on their own, students in 5:1 ROOTS groups participated in more group-level practice.

In sum, because the study conducted by Clarke et al. (2017) was part of the larger ROOTS Efficacy Project, it represented the initial investigation of group size and treatment intensity of the ROOTS intervention. Confirming and identifying the generalizability of its reported findings within a planned sequence of replication (Coyne, Cook, & Therrien, 2016), therefore, was considered crucial to the broader contributions of our program of research. As such, continued research on the treatment intensity of the ROOTS intervention involving kindergarten students from other geographical regions was deemed warranted.

## *Purpose of the Study*

The purpose of this RCT was to extend the existing literature on Tier 2 mathematics interventions by investigating the extent to which the purposeful manipulation of group size affected the overall intervention impact and treatment intensity of the ROOTS intervention. Participating in the RCT were 72 kindergarten classrooms from two school districts in the metropolitan area of Boston, Massachusetts. In conducting the current study in Boston, we sought to determine whether the results of Clarke et al. (2017) generalized across instructional settings and participants from a different geographical region of the United States. Accumulating converging evidence through a framework of systematic replication, as noted by Coyne et al. (2016), increases the trustworthiness of information about an intervention or approach, such as selecting the appropriate group size to deliver mathematics interventions for students with MD. Therefore, extending the research of Clarke et al. (2017), the current study assessed whether the overall intervention impact and treatment intensity, as measured by the frequency of student practice opportunities, of the ROOTS intervention varied across small group formats with 2:1 and 5:1 student-teacher ratios. Additionally, to expand our work on treatment intensity, we also investigated whether the quality of explicit instruction delivered during the ROOTS intervention varied by group size.

Three research questions were investigated: (1) Does the effect of the ROOTS intervention on student mathematics achievement vary by group size (2:1 ROOTS group vs. 5:1 ROOTS group)? (2) Does the frequency of student practice opportunities facilitated during ROOTS instruction vary by group size (2:1 ROOTS group vs. 5:1 ROOTS group)? (3) Does the quality of explicit instruction during the ROOTS intervention vary by group size (2:1 ROOTS group vs. 5:1 ROOTS group)?

## Method

This study employed a randomized block design (blocking on classrooms), randomly assigning students within classrooms to one of three conditions: 2:1 ROOTS group, 5:1 ROOTS group, and a no-treatment control condition. Because a separate line of research has demonstrated the efficacy of the ROOTS intervention relative to a core mathematics program (Hedges' $g$ = .31 to .37; Clark et al., 2016) and no-treatment control conditions (Hedges' $g$ = .16 to 1.08; Clarke et al., 2017; Doabler et al., 2016), the primary focus of the current study was a comparison between the 2:1 and 5:1 ROOTS groups. Thus, the current analyses did not include students in the control condition. The study was conducted across 2 years, with Year 1 and Year 2 representing the 2014–2015 and 2015–2016 school years, respectively. Each study year involved a different cohort of kindergarten students. These cohorts represented the final two cohorts of the larger, federally funded ROOTS Efficacy Project.

### Participants

*Schools.* Nine elementary schools from two Boston area school districts participated in both Year 1 and Year 2 of the present study. District A had a total enrollment of 6,118 students in Year 1 and 6,350 students in Year 2. In District A, 83% of students in Year 1 and 48% of students in Year 2 qualified for free and reduced lunch. All kindergarten students in District A attended the same school. District B had a total enrollment of 6,834 students in Year 1 and 6,721 students in Year 2. In District B, 30% of students in Year 1 and 22% of students in Year 2 qualified for free and reduced lunch. Eight separate schools from District B participated in the study.

*Classrooms.* Participants were drawn from 36 classrooms in Year 1 and 36 classrooms in Year 2 ($N$ = 72). Of the 36 Year 1 classrooms, 28 participated in Year 2. Thus, a total of 44 distinct kindergarten classrooms participated across the study. Half-day kindergarten programs were offered in 11 and 7 classrooms in Year 1 and Year 2, respectively. Across both years of the study, the average classroom size was 24.9 students ($SD$ = 5.9).

Kindergarten classrooms were taught by 44 certified kindergarten teachers, of whom 39 provided the following demographic information: 100% of teachers identified as female and 92% as White, with 8% of teachers declining to provide ethnicity information. Teachers had an average of 14.0 years of teaching experience and 8.9 years of kindergarten teaching experience. Of the 44 teachers, 72% of teachers had a master's degree in education, and 51% 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, & Rolfhus, 2011) and the Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2010), which are standardized measures of early mathematics proficiency. Students were eligible for the ROOTS intervention and thus considered at risk for MD if they received an NSB score of 20 or less and an ASPENS' composite score in the *strategic* or *intensive* ranges.

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 at-risk student. Composite scores within each classroom were then rank ordered, and the 10 lowest ROOTS-eligible students were randomly assigned to one of three conditions: (a) 2:1 ROOTS groups, (b) 5:1 ROOTS groups, or (c) a no-treatment control condition. As previously noted, the current analyses included only those students randomly assigned to the two ROOTS conditions.

Out of the 36 participating classrooms in each year of the study, 26 had at least 10 students who met the eligibility criteria. However, 10 classrooms in Year 1 and 10 classrooms in Year 2 had fewer than 10 ROOTS-eligible

**Table 1.** Descriptive Statistics for Student Variables by Assessment Time and Condition.

| | Fall of Kindergarten | | | | | | Spring of Kindergarten | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2:1 ROOTS | | | 5:1 ROOTS | | | 2:1 ROOTS | | | 5:1 ROOTS | | |
| Measure | *M* (*SD*) | % | *n* | *M* (*SD*) | % | *n* | *M* (*SD*) | % | *n* | *M* (*SD*) | % | *n* |
| Demographics | | | | | | | | | | | | |
| Age at pretest | 5.3 (0.5) | | | 5.3 (0.4) | | | | | | | | |
| Male | | 45 | | | 47 | | | | | | | |
| Race | | | | | | | | | | | | |
| Asian | | 1 | | | 2 | | | | | | | |
| Black | | 7 | | | 6 | | | | | | | |
| White | | 59 | | | 59 | | | | | | | |
| More than one race | | 1 | | | 1 | | | | | | | |
| Unknown | | 33 | | | 31 | | | | | | | |
| Hispanic | | 50 | | | 44 | | | | | | | |
| LEP | | 18 | | | 22 | | | | | | | |
| SPED eligible | | 8 | | | 5 | | | | | | | |
| Outcomes | | | | | | | | | | | | |
| NSB | 12.3 (3.9) | | 138 | 12.1 (3.8) | | 326 | 20.2 (4.8) | | 131 | 20.0 (4.9) | | 294 |
| ASPENS | 21.6 (16.8) | | 137 | 21.1 (16.5) | | 323 | 94.8 (33.1) | | 131 | 92.9 (34.1) | | 294 |
| Oral Counting | 19.5 (12.5) | | 138 | 19.8 (13.1) | | 326 | 47.2 (21.0) | | 131 | 47.3 (21.8) | | 293 |
| TEMA-3 | 16.8 (7.1) | | 137 | 16.9 (6.9) | | 323 | 27.6 (6.6) | | 134 | 27.1 (7.6) | | 300 |
| RAENS | 11.7 (5.7) | | 137 | 11.4 (5.7) | | 324 | 24.7 (5.6) | | 134 | 24.6 (5.5) | | 301 |
| SESAT Total | | | | | | | 473.1 (34.4) | | 130 | 465.0 (36.7) | | 295 |

*Note*. The complete sample included 138 students in the 2:1 ROOTS group condition and 327 students in the 5:1 ROOTS group condition. The sample sizes (*n*) represent students with a particular measure at each assessment period. ASPENS = Assessing Student Proficiency in Early Number Sense; LEP = Limited English proficiency; NSB = Number Sense Brief; RAENS = ROOTS Assessment of Early Numeracy Skills; SESAT = Stanford Early School Achievement Test; TEMA-3 = Test of Early Mathematics Ability (3rd ed.).

students. In these instances, we combined at-risk students from these classrooms to meet the random assignment's 10-student requirement. For example, in Year 1, at-risk students from two classrooms were combined to form a "virtual" ROOTS classroom, which provided a 2:1 ROOTS group and a 5:1 ROOTS group. After these cross-class grouping procedures were applied, a total of 136 ROOTS groups were formed ($n$ = 69 2:1 ROOTS groups, $n$ = 67 5:1 ROOTS group).

*Students.* A total of 1,580 kindergarten students were screened for ROOTS eligibility. Of these students, 659 met eligibility criteria and were randomly assigned to the two-student group condition ($n$ = 138), the five-student group condition ($n$ = 327), or the no-treatment control condition ($n$ = 194). See Table 1 for demographic information on the ROOTS students.

*Interventionists.* ROOTS intervention groups were taught by district-employed instructional assistants (75%) and by interventionists hired specifically for this study (25%). All interventionists were female, with 61% identifying as White, 20% as Hispanic, 2% as African American, 2% as Asian American/Pacific Islander, 2% as Native American, and 8% declining to provide ethnicity information. Most interventionists had prior experience with small-group instruction (91%), a bachelor's degree or higher (63%), and an average of 13 years of teaching experience. Half of the interventionists (50%) had taken a college-level algebra course and 24% had a current teaching license.

## Procedures

*ROOTS intervention.* ROOTS is a Tier 2 kindergarten mathematics intervention program that consists of 50 lessons (20 minutes each) delivered in small-group formats. The primary aim of ROOTS is to support kindergarten students with MD in developing a robust understanding of whole number concepts and skills. Specifically, the ROOTS intervention prioritizes concepts from the Counting and Cardinality, Operations and Algebraic Thinking, and Number and Operations in Base Ten domains of the Common Core State Standards for Mathematics (2010). Lessons 1–25 of the intervention primarily focus on the relationship between numbers and quantities and then begin to incorporate instruction on early concepts of addition and subtraction (e.g., solving word problems). The second half of the intervention (Lessons 26–50) targets building students' understanding of place value with numbers 11–19.

Each ROOTS lesson contains four mathematics activities with scripted guidelines for interventionists to (a) provide overt demonstrations and explanations of new mathematical content; (b) incorporate visual representations of mathematics into lesson activities; (c) facilitate frequent student practice opportunities, including student mathematics verbalizations; and (d) offer specific academic feedback to students based on their performance when completing tasks and activities. ROOTS lessons provide systematic practice and review within and across lessons to promote mastery, maintenance, and transfer. Of particular relevance to the current study's investigation of treatment intensity is the frequency and quality of student practice opportunities prioritized by the ROOTS intervention. Such practice opportunities include students verbalizing their mathematical thinking and working with visual representations of mathematical ideas (e.g., Base 10 blocks).

The opening activity in each ROOTS lesson consists of a brief warm-up (3 minutes), the "Nifty Fifty" activity, focused on number identification skills and use of efficient counting strategies with a 1–50 number chart. Each Nifty Fifty activity corresponds to the number of lessons completed in the intervention program. In Lesson 17, for example, interventionists use the Nifty Fifty activity to help children count and identify numbers up to 17. The warm-up activities also support students' knowledge of rational counting (i.e., one-to-one correspondence) and identifying whether one group of objects is greater than, less than, or equal to another group of objects. Next, interventionists deliver a 5-minute activity that overtly introduces a new mathematical concept or skill that is central to the lesson's overall objective. For this activity, interventionists use concrete objects (e.g., counting blocks or number lines) to explicitly demonstrate and explain the targeted concept or skill. The third activity (7 minutes) involves either guided practice of content introduced in the second activity or a review of previously learned material. The final activity is a brief worksheet activity (5 minutes) that interventionists use to review the lesson's content. Worksheets contain a "note home" (in both English and Spanish) to provide students with additional practice opportunities outside of school.

In the current study, ROOTS was delivered in small-group formats (2:1 or 5:1 student-teacher ratios). Interventionists delivered the 20-minute lessons, 5 days per week for approximately 10 weeks. ROOTS began in late fall and ended in the spring and was delivered in addition to students' core mathematics instruction.

*Professional development.* All participating interventionists received two 5-hour professional development workshops. The first workshop focused on Lessons 1–25, while the second workshop targeted Lessons 26–50. Both workshops, which were delivered by project staff, also gave interventionists exposure to empirically validated practices of mathematics instruction and small-group management techniques. During the workshops, interventionists received opportunities to practice and receive feedback on lesson delivery. To bolster implementation, all interventionists received in-class coaching support during the intervention. Coaching visits offered feedback on the fidelity and quality of intervention implementation. Each intervention group received two coaching visits over the course of the study.

*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. 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. Overall, observations indicated that instruction in the 2:1 and 5:1 ROOTS groups was delivered with similar levels of implementation fidelity. As shown in Table 3, no significant differences in fidelity of implementation were observed between the 2:1 and 5:1 ROOTS groups ($p$s > .18).

*Core mathematics instruction.* Throughout the study, ROOTS students continued to receive core mathematics instruction delivered in their kindergarten classroom. Survey data reflected that teachers in District A primarily used the Scott Foresman mathematics curriculum during core mathematics instruction, while teachers in District B primarily used the enVisionMath curriculum. Teachers also supplemented core instruction with their own materials. Teachers reported that they provided an average of 32.8 minutes of mathematics instruction per day ($SD$ = 22.8). Teachers also noted that a main instructional focus when teaching whole number concepts was reading number names and knowing the count sequence.

## Measures

All ROOTS students were administered five measures of whole number understanding at pretest and posttest. One distal measure of mathematics achievement was administered at posttest only. Trained research staff administered all student measures. Interscorer reliability criteria were met for all assessments (i.e., >.95).

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

Oral Counting–Early Numeracy Curriculum-Based Measurement (Clarke & Shinn, 2004), a curriculum-based measure, has students orally count in English for 1 minute, and the discontinue rule applies after the first counting error. The highest correct number counted represents a student's score. Test-retest reliability and alternate-form reliability are reported at above .80, concurrent validity is reported as ranging from .49 to .70, and predictive validity with standardized measures of mathematics ranged from .46 to .72.

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

NSB (Jordan et al., 2010) is an individually administered measure with 33 items that assess counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations. NSB has a coefficient alpha of .84.

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

Stanford Early School Achievement Test (SESAT; Harcourt Brace Educational Measurement, 2003) is a group-administered, standardized, norm-referenced measure, with two mathematics subtests: Problem Solving and Procedures. The internal consistency for the SESAT is .88. ROOTS students were administered the SESAT at posttest only.

### *Observations of ROOTS Instruction*

Each ROOTS group was observed approximately three times ($M = 2.9$, $SD = 0.8$) over the course of the intervention, with approximately 3 weeks separating each observation occasion. A total of 391 observations were conducted, of which 124 included two observers. Trained observers, who were blind to our research hypotheses, conducted all observations using two observation measures.

Observers used Classroom Observations of Student-Teacher Interactions—Mathematics (COSTI-M; Doabler et al., 2015) to document four types of student practice opportunities in the 2:1 and 5:1 ROOTS groups. These practice opportunities, which served as metrics of treatment intensity, included (a) individual practice, (b) group practice, (c) guided practice, and (d) independent practice. While the COSTI-M also documents teachers' use of explicit demonstrations and provision of academic feedback, the current study focused specifically on student practice opportunities because prior research with the COSTI-M suggests their association with student mathematics achievement (Doabler et al., 2015). Individual practice represented a practice opportunity provided to one student, while group practice represented a practice opportunity provided to two or more students. Individual and group practice included student mathematics verbalizations and opportunities to manipulate concrete representations of mathematical ideas (e.g., Base 10 blocks). Observers also coded whether individual and group practice entailed concurrent teacher support (guided or independent). Guided practice was operationally defined as an opportunity for one or more students to verbalize or physically demonstrate their mathematical understanding with concurrent instructional support from the teacher. Independent practice represented an opportunity for one or more students to verbalize or physically demonstrate their mathematical understanding without teacher support. Rates of these practice opportunities were calculated by dividing their observed frequency by the number of instructional minutes. Analyses also included an "all practice" variable, which comprised guided, independent, group, and individual practice opportunities. Mean rates of practice opportunities across observation occasions were calculated and used as treatment intensity predictors in subsequent analyses.

Quality of Explicit Mathematics Instruction (QEMI; Doabler & Clarke, 2012) is a broad measure of instruction quality. The QEMI comprises seven items that target the quality of explicit mathematics instruction, including group and individual practice opportunities, student participation, teacher modeling, academic feedback, efficiency of instructional delivery, and instructional scaffolding. Internal consistency of the measure was high—.93 (coefficient alpha). To rate the quality of each item, observers used a 4-point rating scale, with scores of 1–2 representing the lower quality range and 3–4 representing the upper quality range. Observers completed the QEMI at the conclusion of each observation occasion. Total QEMI scores were computed as the mean across all items. The mean across the three observations for the 2:1 and 5:1 ROOTS groups was used as a treatment intensity predictor in subsequent analyses.

*Estimates of interobserver reliability and stability.* Interobserver reliability for the COSTI-M variables, which were represented by intraclass correlation coefficients (ICCs), were as follows: .89 for all practice, .92 for individual practice, .91 for group practice, .41 for guided practice, and .79 for independent practice. ICCs for the QEMI's total score (i.e., overall quality of explicit mathematics instruction) and the ROOTS fidelity of implementation tool were .93 and .88, respectively. Guidelines proposed by Landis and Koch (1977) suggest that these ICCs indicate moderate to nearly perfect interobserver agreement.

To provide an estimate of stability, ICCs were calculated across the three observations within each ROOTS group. Stability ICCs for COSTI-M variables were as follows: .28 for all practice, .23 for individual practice, .20 for group practice, .27 for guided practice, and .23 for independent practice. These ICCs represent low stability, indicating that rates of practice opportunities differed across observations. The stability ICC for the QEMI was .32.

### Statistical Analysis

For our first research question, we examined the effects of the 2:1 versus 5:1 condition on student outcomes using a nested mixed-model (multilevel) Time × Condition analysis (Murray, 1998) to account for the intraclass correlation associated with students nested within groups. The analysis tested for differences between conditions on gains in outcomes from the fall to spring of kindergarten. The statistical model included time, coded 0 at pretest and 1 at posttest; condition, coded 0 for 5:1 ROOTS groups and 1 for 2:1 ROOTS groups; and the interaction between the two. A mixed analysis of covariance model was used for the SESAT measured only at posttest. Our second research question examined whether 2:1 and 5:1 ROOTS groups experienced different rates of student practice opportunities or QEMI ratings using independent-samples *t* tests.

Because students were randomly assigned within classrooms, we tested an additional set of mixed models that extended those discussed above to account for clustering within classrooms. Results were similar in both sets of models, and condition effects did not vary by classroom, 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 (REML), generally recommended for multilevel models (Hox, 2002). Maximum likelihood estimation for the Time × Condition analysis uses all available data to provide potentially unbiased results even in the face of substantial attrition, provided the missing data were missing at random (Graham, 2009). 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 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 did not markedly deviate from normality; skewness and kurtosis fell within ±2.0 for all measures except for oral counting, where kurtosis was 2.9. Nonetheless, multilevel regression methods have also been found quite robust to violations of normality (e.g., Hannan & Murray, 1996).

*Effect sizes.* To further interpretation, we computed Hedges' *g* (Hedges, 1981) for each fixed effect as recommended by the What Works Clearinghouse (2017). Assuming ICCs from .1 to .2, approximately 65 groups per condition, an average of 3.5 students per group, and pre-post correlations of .50 and .71, the minimally detectable effect sizes (*g*) ranged from 0.23 to 0.32.

## Results

Table 1 presents means, standard deviations, and sample sizes for the six dependent variables by assessment time and condition. Below we present results from tests of bias due to attrition, effects of the 2:1 versus 5:1 conditions on student outcomes, and differential rates of student practice opportunities and QEMI ratings between 2:1 and 5:1 conditions.

### Attrition

Student attrition was defined as students with data at pretest but missing data at posttest. Attrition rates were between 7% and 9% for all outcomes measured at posttest. Only 6% of students were missing all posttest data. The proportion of students missing all posttest data did not differ between 2:1 and 5:1 conditions, $\chi^2_{(1)} = 3.03$, $p = .082$. Although differential rates of attrition are undesirable, differential scores on math tests present a far greater threat to internal 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. We found no statistically significant interactions or evidence that mathematics scores were differentially affected by attrition across conditions ($p$s > .31).

### Impact on Student Outcomes

Table 2 presents the results of the statistical models comparing gains between 2:1 and 5:1 ROOTS groups. The models

**Table 2.** Results From a Nested Time × Condition Analysis on Fall-to-Spring Gains in Math Comparing 2:1 and 5:1 ROOTS Groups.

| Variables | | NSB | ASPENS | Oral Counting | TEMA-3 | RAENS |
|---|---|---|---|---|---|---|
| Fixed effects | Intercept | 12.17*** | 21.20*** | 19.75*** | 16.95*** | 11.48*** |
| | | (0.35) | (1.94) | (1.20) | (0.58) | (0.44) |
| | Time | 7.85*** | 71.83*** | 27.68*** | 10.03*** | 13.08*** |
| | | (0.28) | (1.92) | (1.44) | (0.36) | (0.36) |
| | ROOTS group | 0.19 | 0.20 | −0.21 | −0.06 | 0.25 |
| | | (0.55) | (3.13) | (2.00) | (0.90) | (0.70) |
| | Time × ROOTS group | −0.04 | 1.49 | 0.03 | 0.70 | −0.03 |
| | | (0.48) | (3.25) | (2.39) | (0.61) | (0.60) |
| Variances | ROOTS group intercept | 4.76*** | 99.48** | 13.04 | 13.05*** | 6.75*** |
| | | (1.04) | (30.41) | (11.98) | (2.91) | (1.66) |
| | ROOTS group gains | 0.71 | 36.90 | 28.15* | 1.60* | 1.46* |
| | | (0.46) | (20.29) | (12.18) | (0.78) | (0.72) |
| | Student | 4.53*** | 167.50*** | 83.33*** | 23.82*** | 10.49*** |
| | | (0.80) | (32.04) | (15.66) | (2.42) | (1.44) |
| | Residual | 8.83*** | 384.28*** | 184.70*** | 12.55*** | 12.79*** |
| | | (0.71) | (30.36) | (15.13) | (1.03) | (1.03) |
| Hedges' *g* | Time × ROOTS group | −0.01 | 0.04 | <0.01 | 0.10 | −0.01 |
| *p* values | Time × ROOTS group | .9334 | .6480 | .9905 | .2515 | .9629 |
| *df* | Time × ROOTS group | 173 | 175 | 148 | 154 | 167 |

*Note*. 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 2:1 and 5:1 ROOTS conditions. ASPENS = Assessing Student Proficiency in Early Number Sense; NSB = Number Sense Brief; RAENS = ROOTS Assessment of Early Numeracy Skills; TEMA-3 = Test of Early Mathematics Ability (3rd ed.).
*$p < .05$. **$p < .01$. ***$p < .001$.

in Table 4 tested fixed effects for differences between conditions at pretest (2:1 ROOTS group effect), gains across time, and the interaction between the two. We found no statistically significant differences at pretest ($p$s > .72), which suggested that students were similar in the fall of kindergarten. We also found no statistically significant differences by condition in gains from fall to spring ($p$s > .25). The Time × Condition model estimated differences in gains between conditions of −0.04 for the NSB (Hedges' $g = -0.01$), 1.49 for the ASPENS ($g = 0.04$), 0.03 for oral counting ($g < 0.01$), 0.70 for the TEMA-3 standard score ($g = 0.10$), and −0.03 for the RAENS ($g = -0.01$). The analysis of covariance model estimated differences between 2:1 and 5:1 ROOTS groups of 1.01 for the SESAT ($g = 0.03$, $p = .725$).[1]

### Impact on Student Practice and Quality of Explicit Math Instruction

Table 3 presents descriptive statistics for the observed rates per minute of student practice opportunities and QEMI ratings as well as results of independent-samples $t$ tests comparing these outcomes by condition. Compared to the 5:1 ROOTS groups, 2:1 groups experienced higher rates of individual practice ($t = 2.95$, $p = .004$, $g = 0.51$) and lower rates of group practice ($t = -2.12$, $p = .036$, $g = -0.36$). We found no effects of condition on the rate of guided practice ($t = 1.15$, $p = .254$, $g = 0.20$), independent practice ($t = 0.80$, $p = .424$, $g = 0.14$), or all practice combined ($t = 1.45$, $p = .150$, $g = 0.24$). With respect to QEMI ratings, 2:1 ROOTS groups had higher quality individual practice opportunities compared to 5:1 groups ($t = 2.19$, $p = .031$, $g = 0.37$). We observed no effects of condition on other QEMI items or the QEMI total score ($p$s > .13).

## Discussion

The purpose of this study was to examine whether an experimental manipulation of group size affected the overall intervention impact and treatment intensity of a Tier 2 mathematics intervention. The frequency of student practice opportunities and the quality of explicit mathematics instruction served as metrics of the intervention's treatment intensity. The study investigated three research questions.

Results from our first research question suggested no statistically significant differences in student mathematics outcomes when comparing the 2:1 and 5:1 ROOTS groups. Essentially, students in the 2:1 and 5:1 ROOTS groups demonstrated comparable performances on the six mathematics outcome measures. For our second research question, we found that both the 2:1 and 5:1 ROOTS groups facilitated similarly high rates of guided and independent practice opportunities. Our findings did show that the ROOTS groups differed on how frequently they facilitated group and individual practice opportunities. Specifically,

**Table 3.** Results of Independent-Samples *t* Tests Comparing Rates of Student Practice, Quality of Explicit Instruction, and Fidelity of Implementation by Size of ROOTS Group.

| Variables | 2:1 ROOTS Groups, *M* (*SD*) | 5:1 ROOTS Groups, *M* (*SD*) | *t* | *p* | Hedges' *g* |
|---|---|---|---|---|---|
| Rates of student practice opportunities | | | | | |
| Guided practice | 0.9 (0.5) | 0.8 (0.4) | 1.15 | .254 | 0.20 |
| Independent practice | 3.0 (0.7) | 2.9 (0.7) | 0.80 | .424 | 0.14 |
| Individual practice | 2.2 (0.9) | 1.8 (0.7) | 2.95 | .004 | 0.51 |
| Group practice | 1.7 (0.7) | 1.9 (0.5) | −2.12 | .036 | −0.36 |
| All practice | 3.8 (0.8) | 3.7 (0.7) | 1.45 | .150 | 0.24 |
| Quality of Explicit Math Instruction (QEMI) | | | | | |
| Efficient delivery of instruction | 3.2 (0.6) | 3.1 (0.6) | 0.92 | .361 | 0.15 |
| Student participation | 3.2 (0.5) | 3.1 (0.5) | 1.23 | .221 | 0.21 |
| Effective teacher modeling | 3.2 (0.5) | 3.2 (0.5) | 0.45 | .625 | 0.08 |
| Group practice opportunities | 3.1 (0.5) | 3.2 (0.5) | −0.39 | .698 | −0.06 |
| Checks of understanding | 3.3 (0.5) | 3.2 (0.5) | 1.05 | .297 | 0.16 |
| Individual practice opportunities | 3.3 (0.5) | 3.1 (0.5) | 2.19 | .031 | 0.37 |
| Instructional scaffolding | 3.2 (0.5) | 3.1 (0.6) | 1.53 | .128 | 0.26 |
| Total QEMI score | 3.2 (0.4) | 3.1 (0.5) | 1.14 | .255 | 0.20 |
| Fidelity of implementation | | | | | |
| 1. Number of activities taught out of 5 | 4.2 (0.4) | 4.2 (0.4) | −0.04 | .972 | 0.00 |
| 2. Met math objectives | 3.6 (0.5) | 3.5 (0.5) | 0.97 | .336 | 0.15 |
| 3. Followed teacher scripting | 3.5 (0.5) | 3.4 (0.5) | 1.33 | .185 | 0.24 |
| 4. Used prescribed math models | 3.7 (0.4) | 3.6 (0.5) | 1.34 | .183 | 0.23 |
| Total fidelity | 3.6 (0.4) | 3.5 (0.5) | 1.29 | .199 | 0.21 |
| Average observation duration in minutes | 20.7 (3.7) | 22.9 (4.1) | −3.30 | .001 | −0.57 |

*Note.* Group *t* tests were based on 69 2:1 ROOTS groups and 67 5:1 ROOTS groups (134 degrees of freedom). Quality of explicit math instruction was rated from 1 = not present to 4 = highly present. Total instructional quality was calculated as the mean across items. Fidelity of implementation Items 2 through 4 were rated from 1 = *none* to 4 = *all*. Total fidelity was calculated as the mean across Items 2 through 4.

observation data revealed that the highest rates of individual practice opportunities were documented in the 2:1 ROOTS groups and that the 5:1 ROOTS groups engaged students in more group-level practice. These data suggest that students in both the 2:1 and 5:1 groups received intensive learning experiences.

Collectively, findings from the current study's first two research questions replicated those reported in Clarke et al. (2017). As juxtaposed in Table 4, comparable effect sizes were reported in both studies. Moreover, all effect sizes from the current study fell within 95% confidence intervals of the study conducted by Clarke et al. (2017). Because replication is a fundamental principle of scientific research (Coyne et al., 2016; Feuer, Towne, & Shavelson, 2002; Gottfredson et al., 2015; Valentine et al., 2011), we contend that establishing the generalizability of our findings across participants and classrooms from a different geographical region is important not only for our program of research but also for the field at large. Given that the students in the 5:1 ROOTS groups from Clarke et al. (2017) and the current study performed commensurately relative to their peers in the 2:1 ROOTS groups, it helps us build a convergence of

**Table 4.** Published and Replicated Effect Sizes From the Time × Condition Analyses of Fall-to-Spring Gains in Math Comparing 2:1 and 5:1 ROOTS Groups.

| Outcomes | Clarke et al. (2017) | This Study |
|---|---|---|
| NSB | 0.00 [−0.20, 0.20] | −0.01 [−0.20, 0.19] |
| ASPENS | −0.14 [−0.32, 0.05] | 0.04 [−0.14, 0.23] |
| Oral Counting | 0.08 [−0.12, 0.29] | 0.00 [−0.22, 0.22] |
| TEMA-3 | −0.01 [−0.17, 0.15] | 0.10 [−0.07, 0.26] |
| RAENS | 0.03 [−0.17, 0.22] | −0.01 [−0.22, 0.21] |
| SESAT Total | 0.03 [−0.14, 0.21] | 0.03 [−0.13, 0.19] |

*Note*. Table entries show Hedges' *g* effect size estimates with 95% confidence intervals in brackets. ASPENS = Assessing Student Proficiency in Early Number Sense; NSB = Number Sense Brief; RAENS = ROOTS Assessment of Early Numeracy Skills; TEMA-3 = Test of Early Mathematics Ability (3rd ed.).

evidence in support of delivering the ROOTS intervention in conventionally sized small groups (i.e., five students). Perhaps as important, we believe our findings have potential implications for the allocation of resources in today's schools in terms of delivering Tier 2 mathematics instruction to students at risk for persistent difficulties in mathematics. As schools across the nation continue to face financial shortages, they are constantly searching for ways to "do more with less," particularly in terms of human capital. Our results, while preliminary, suggest that schools may be able to use fewer interventionists to intervene with more at-risk kindergarten students at one time.

To extend our work on treatment intensity, we also examined whether the quality of explicit mathematics instruction varied by group size. Such instructional quality data were not investigated in Clarke et al. (2017). Results from our third research question suggest that the individual practice opportunities facilitated in the ROOTS 2:1 groups were found to be richer and more meaningful than those observed in the larger ROOTS groups. Observers may have rated individual practice as being higher quality in the 2:1 groups because it appeared that students received more individualized attention from the interventionists. Interestingly, ROOTS groups were not found to differ on overall quality of explicit instruction. While we anticipated these findings because of the systematic nature of ROOTS, two factors may have contributed to this finding. First, it is plausible that our study's modest sample size may have limited the capacity to detect significant overall quality differences between the two intervention groups. Second, it may be that a different quality observation tool may have been able to detect group differences.

## *Implications for Research and Practice*

One implication that arises from this study is that we see value in researchers expanding the extant literature base on the topics of group size and treatment intensity in mathematics intervention research. To our knowledge, few studies have concurrently investigated these two highly important variables in the context of early mathematics interventions. Establishing the optimal size of instructional groups could provide schools with important information on how to best intensify learning opportunities for students with MD. Future research is therefore warranted in this area.

Additionally, our research may shed light on the possibility of a "threshold effect" of student practice. While it was hypothesized that the 2:1 groups would outperform the 5:1 groups, both group sizes were expected to provide students with intensive learning experiences based on the instructional design of the ROOTS intervention. Therefore, the potential yield of additional practice in the smaller groups may have diminished after a certain rate or threshold was obtained during instruction. Given the possibility of threshold effects, future research is needed to establish optimal rates of student practice opportunities for teachers to provide when teaching students with or at risk for MD. For instance, studies of extant datasets may utilize regression methods (Lee, Seo, & Shin, 2011) to identify whether differential associations exist between rates of student practice and mathematics achievement above and below a specific level (threshold) of practice opportunities (Doabler et al., in press). Information garnered from such a priori work could, in turn, serve as a guidepost for planning future research that experimentally evaluates threshold effects. For example, in a future study, researchers might consider randomly assigning students to two conditions that employ the same validated mathematics intervention but vary in terms of practice dosages. In one condition, students receive a typical dose of practice as prescribed in the intervention, whereas students in the second condition receive additional practice opportunities above the previously established threshold.

Relatedly, we see practical value in our work on student practice opportunities. Research across a variety of disciplinary fields, including music and sports (Ericsson, Roring, & Nandagopal, 2007), neuroscience (Fields, 2005), as well as cognitive and educational psychology (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) has begun to shed light on the importance of practice. In early mathematics, practice is essential for building mathematical proficiency among the full range of learners, including students with MD. As shown in the current study, students in the 5:1 and 2:1 groups received frequent opportunities to practice with the critical concepts and skills of whole number and operations. We therefore encourage teachers to facilitate frequent practice opportunities when working with students with MD.

In addition to increasing the amount of student practice, we think it is important to have teachers consider improving the quality of practice opportunities in early mathematics instruction. Quality, here, includes appropriately scaffolding practice opportunities to meet the instructional needs of struggling learners. In this regard, teachers should guide initial practice opportunities and then gradually release their instructional support as students become more independent in the acquisition of mathematics concepts and skills. For example, teachers can increase the richness of mathematical discourse by directly guiding students with MD through the process of explaining the steps and reasoning behind a mathematical solution.

Another potential implication stems from the notion of peer learning (Fuchs & Fuchs, 1998). While not formally measured, students in the 5:1 ROOTS groups may have had more opportunities to learn from their peers, which, in turn, provided an added value to the overall treatment effect for these groups. For example, when using concrete materials, students in the 5:1 groups may have been able to observe more vividly what they were expected to do during a mathematical task or activity. It may have been a similar situation with student mathematics verbalizations. Students in the 5:1 groups may have benefited from hearing a wider range of mathematical thinking. Future work should explore the role peer learning opportunities have in increasing the overall impact and treatment intensity of small-group mathematics interventions.

Finally, while our work in the area of treatment intensity has focused extensively on the frequency and quality of student practice opportunities, a logical next step for future research might be to apply a treatment intensity framework, such as the one proposed by Warren et al. (2007). Utilizing the Warren et al. framework might offer a more comprehensive way to measure the treatment intensity of the ROOTS intervention. For example, in a future efficacy trial, application of their framework would enable us to compare the *cumulative intervention intensity* of ROOTS to a different Tier 2 intervention that represents the counterfactual condition.

### *Limitations*

A number of limitations must be considered when interpreting our results. First, each ROOTS group was observed only three times. While this was primarily driven by resource constraints in the larger efficacy trial, more observations would likely permit a deeper understanding of the treatment intensity of ROOTS. One suggestion to accommodate the expense of conducting real-time observations in large-scale research and gain the capacity to paint a richer picture of instructional practices would be to conduct more observations in a limited number of classrooms of the research sample. Another possible limitation was that the intervention groups were taught by "interventionists" rather than the classroom teachers. Teachers likely have a better grasp of the current level of performance of their students and thus could provide more effective mathematics instruction for their struggling students. A third limitation was that our replication study did not consider other potential variables of treatment intensity. For example, examining the duration or complexity of the targeted student practice opportunities may provide further insight into the treatment intensity of the 2:1 and 5:1 ROOTS groups. Relatedly, the current study included the quality of explicit instruction as a metric of treatment, and results suggested that quality of individual practice opportunities was higher in the 2:1 ROOTS groups. While blind to our research hypotheses, it is plausible observers may have been partial to the smaller groups, thus impacting the quality ratings. Also, the same group of researchers carried out the current study and the initial RCT. Author overlap can introduce bias in replication research (Coyne et al., 2016). We contend, however, that the likelihood of this type of bias was largely controlled for through the inclusion of an external independent evaluator. Finally, this study focused specifically on the ROOTS intervention. Therefore, future research is warranted to determine whether our findings replicate with other Tier 2 mathematics interventions.

## Conclusion

Building a converging knowledge base of effective mathematics instruction is paramount to supporting the development of mathematical proficiency for students with MD. One way to help crystalize the mathematics intervention literature is to not only establish the efficacy of mathematics interventions but also examine alterable variables, such as group size, that are hypothesized to increase their treatment intensity. Investigations that employ this type of dual focus, such as the current study, have the potential to contribute to the knowledge base of effective mathematics instruction for students with intensive learning needs in mathematics.

### Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Drs. Christian T. Doabler, Ben Clarke, Scott K. 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. Additionally, the terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research. An independent external evaluator and coauthor of this publication completed the research analysis described in the article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education

Sciences, U.S. Department of Education through Grant R324A120304 to the Center on Teaching and Learning at the University of Oregon.

## Note

1. ROOTS and control conditions were compared in the current sample, and results were similar to those previously published. Statistically significant effects of ROOTS on NSB ($g$ = 0.26, $p$ = .0034), ASPENS ($g$ = 0.49, $p$ < .0001), TEMA-3 ($g$ = 0.21, $p$ = .0030), RAENS ($g$ = 0.88, $p$ < .0001), and SESAT ($g$ = 0.28, $p$ < .0001) were found.

## References


Archer, A. L., & Hughes, C. A. (2011). *Exploring the foundations of explicit instruction*. New York, NY: The Guilford Press.

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

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

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

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

Clements, D. H., Agodini, R., & Harris, B. (2013). *Instructional practices and student math achievement: Correlations from a study of math curricula* (NCEE Evaluation Brief ED-04-CO-0112/0003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Codding, R. S., VanDerHeyden, A. M., Martin, R. J., Desai, S., Allard, N., & Perrault, L. (2016). Maniuplating treatment dose: Evaluating the frequency of small group intervention targeting whole number operations. *Learning Disabilities Research & Practice*, *31*, 208–220. doi:10.1111.drp.12120

Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

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

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. (2012). *Quality of explicit mathematics instruction*. Unpublished observation instrument, Center on Teaching and Learning, University of Oregon, Eugene.

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

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. doi:10.1177/0014402916660084

Doabler, C. T., Stoolmiller, M., Kennedy, P., Nelson, N. J., Clarke, B., Gearin, B., Fien, H., Smolkowski, K., & Baker, S. K. (in press). Do components of explicit instruction explain the differential effectiveness of a core mathematics program for students with mathematics difficulties? A mediated moderation analysis *Assessment for Effective Intervention*.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58. doi:10.1177/1529100612453266

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

Ericsson, K. A., Roring, R., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, *18*, 3–56. doi:10.1080/13598130701350593

Feuer, M., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, *31*, 4–14.

Fields, R. D. (2005). Myelination: An overlooked mechanism of synaptic plasticity? *The Neuroscientist*, *11*, 528–531.

Fuchs, L., & Fuchs, D. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research & Practice*, *13*, 126–137.

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

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

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

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.

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.

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: Lawrence Erlbaum.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, *20*, 82–88.

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

Lee, S., Seo, M., & Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association*, *106*, 220–231.

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

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

SAS Institute. (2009). *SAS/STAT® 9.2 user's guide* (2nd ed.). Cary, NC: Author.

Sood, S., & Jitendra, A. (2013). An exploratory study of number sense program to develop kindergarten students number proficiency. *Journal of Learning Disabilities*, *46*, 328–346.

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: John Wiley & Sons.

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.

What Works Clearinghouse. (2017). *Procedures and standards handbook* (Version 4.0). Washington, DC: U.S. Department of Education, Institute of Education Sciences.