An Examination of the Promise of the NumberShire Level 1 Gaming Intervention for Improving Student Mathematics Outcomes

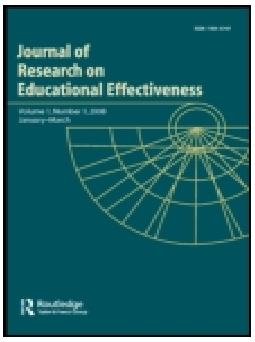
Hank Fien, Christian T. Doabler, Nancy J. Nelson, Derek B. Kosty, Ben Clarke & Scott K. Baker

Journal of Research on Educational Effectiveness, 9:4, 635-661, DOI: 10.1080/19345747.2015.1119229

Published online: 06 Jan 2016.

Funding: Development of a Game-based Integrated Learning and Assessment System to Target Whole Number Concepts (Project NumberShire), Award Number R324A120071 Source: Institute of Education Sciences. National Center for Special Education Research. And

This research was supported in part by Project NumberShire 1, Grant No. EDIES11C0026, a subcon- tract with ThoughtCycle through the U.S. Department of Education, Institute of Education Sciences, Small Business Innovation Research Program. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.



Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

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

An Examination of the Promise of the NumberShire Level 1 Gaming Intervention for Improving Student Mathematics Outcomes

Hank Fien, Christian T. Doabler, Nancy J. Nelson, Derek B. Kosty, Ben Clarke & Scott K. Baker

To cite this article: Hank Fien, Christian T. Doabler, Nancy J. Nelson, Derek B. Kosty, Ben Clarke & Scott K. Baker (2016) An Examination of the Promise of the NumberShire Level 1 Gaming Intervention for Improving Student Mathematics Outcomes, Journal of Research on Educational Effectiveness, 9:4, 635-661, DOI: 10.1080/19345747.2015.1119229

To link to this article: http://dx.doi.org/10.1080/19345747.2015.1119229

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

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. 4, 635–661
http://dx.doi.org/10.1080/19345747.2015.1119229

An Examination of the Promise of the NumberShire Level 1 Gaming Intervention for Improving Student Mathematics Outcomes

Hank Fien[a], Christian T. Doabler[a], Nancy J. Nelson[a], Derek B. Kosty[b], Ben Clarke[a], and Scott K. Baker[a]

ABSTRACT
The purpose of this study was to test the promise of the NumberShire Level 1 Gaming Intervention (NS1) to accelerate math learning for first-grade students with or at risk for math difficulties. The NS1 intervention was developed through the Institute of Education Sciences, Small Business Innovation Research Program (Gause, Fien, Baker, & Clarke, 2011) as a digitally based technology tool to allow educators to intervene early and strategically with students struggling to learn mathematics. This study used a randomized controlled trial design to test the promise of the NS1 intervention. In total, 250 first-grade students were randomly assigned within classrooms to the treatment condition or a control condition. Results indicate significant effects favoring the treatment group on proximal measures of whole-number concepts and skills. Intervention effects were not statistically significant for distal outcome measures. Treatment effects were not moderated by special education or English learner status; however, the condition by initial skill level interaction approached significance. Additionally, there was no relationship between dosage variables and students' response to the intervention. Limitations and future directions for research are discussed.

KEYWORDS
intervention
gaming
math

National assessments indicate that the majority of U.S. students are struggling to acquire knowledge of mathematical concepts and skills necessary to become proficient in mathematics. Results of the 2013 National Assessment for Educational Progress (NAEP) indicate that only 42% of fourth-grade students score at or above *Proficient* in mathematics, and 17% are performing below *Basic*, suggesting that many students struggle to meet grade-level expectations in mathematics (National Center for Education Statistics [NCES], 2014). Difficulties in mathematics are even more pronounced for students with disabilities. An alarming 45% of all fourth-grade students identified with a disability scored below *Basic* on the 2013 NAEP in mathematics (NCES, 2014). Likewise, a staggering 41% of English language learners (ELLs) scored below *Basic* on the 2013 NAEP.

CONTACT Hank Fien ffien@uoregon.edu University of Oregon, Center on Teaching and Learning, 1600 Millrace, Suite 206, 1211 University of Oregon, Eugene, OR 97403-1211, USA.
[a]University of Oregon, Eugene, Oregon, USA
[b]Oregon Research Institute, Eugene, Oregon, USA
Color versions of one or more of the figures in this article can be found online at www.www.tandfonline.com/uree.

Evidence also suggests that students who exhibit mathematics difficulties (MD) early in their formal schooling are likely to continue to struggle throughout the elementary years (Bodovski & Farkas, 2007; Morgan, Farkas, & Wu, 2009). Several recent longitudinal studies have demonstrated that mathematics achievement trajectories are relatively stable and are established as early as kindergarten and first grade (Bodovski & Farkas, 2007; Morgan et al., 2009; Morgan, Farkas, & Wu, 2011). In an analysis of the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) Cohort data set, Morgan et al. (2009) found that students who displayed mathematics difficulties in both fall and spring of kindergarten subsequently had the lowest growth trajectories through the end of fifth grade. Whereas mathematics achievement gaps are evident for important subgroups of students early in their schooling, evidence suggests that these gaps are widening over time. For example, the mathematics achievement gap represented as average scores between ELLs and non-ELLs is 25 points in fourth-grade (219 for ELLs and 244 for non-ELLs). By eighth grade the mathematics achievement gap nearly doubles to a 41-point gap (246 for ELLs and 287 for non-ELLs; NCES, 2014).

Arguably, early intervention is a viable approach to prevent these gaps in mathematics learning and should be orchestrated before these gaps become apparent (Clarke, Baker, et al., 2015; Morgan et al., 2009, 2011), using instructional approaches designed to support students at risk for MD. Converging evidence supports the use of explicit instruction as a viable approach for preventing early mathematics difficulties (Gersten, Beckmann, et al., 2009, National Mathematics Advisory Panel [NMAP], 2008). Researchers and policymakers recommend that such explicit instruction approaches occur in an early detection, prevention-oriented support system, such as models of Response to Intervention (RtI; Gersten, Beckmann, et al., 2009; National Association of State Directors of Special Education [NASDSE], 2006). Typically RtI models promote high-quality instruction and universal screening for all students, with layers of increasingly intensive "tiers" of support for students as the level of student's instructional need increases. Student progress is monitored frequently to determine whether intervention efforts are intensive enough to adequately support students to reach important academic outcomes.

Unfortunately, a major limitation of current instructional practice in mathematics is the significant lack of evidence-based, teacher-directed, explicit instruction for students with or at risk for MD. Although the evidence behind providing students with or at risk for MD with daily, explicit, teacher-directed mathematics instruction is quite strong (Gersten, Chard, et al., 2009), a recent population-based, longitudinal study revealed a limited use of this evidence-based practice in first-grade classrooms (Morgan, Farkas, & Maczuga, 2015). During a month of mathematics instruction, which averaged five days per week and 54 minutes per day, there was an average of only 11.8 instances of teacher-directed instruction, a surprisingly meager dose, considering the instructional practice has strong empirical support to improve the mathematics achievement of struggling learners (Gersten, Beckmann, et al., 2009). Furthermore, Morgan and colleagues (2015) found no relation between the number of students with MD in a classroom and the use of teacher-directed instruction. In other words, teachers, in general, are not differentiating instruction based on the student composition of their particular classroom.

Against this backdrop, we developed the Numbershire Level 1 Gaming Intervention (NS1) as a school-based, digital learning tool designed to accelerate mathematics learning for students struggling to develop an understanding of whole-number concepts and skills. NS1 was conceptualized to serve as a supplemental intervention that can be employed within an RtI service delivery model (e.g., used in Tier 2 as a supplement to core mathematics instruction: Gersten, Beckmann, et al., 2009). We designed NS1 as a tool for teachers to increase the availability of explicit and systematic mathematics instruction for students with or at risk for MD. To increase the mathematics achievement of students with or at risk for MD, we integrated the science of explicit instruction and a focus on critical mathematics content in the early grades (i.e., whole-number concepts) with emerging game-based learning design principles (Doabler & Fien, 2013; Klopfer, Osterweil, & Salen, 2009) that have the potential to engage students in learning opportunities.

Promise of Education Technology and Current Research

There is ostensibly great promise for leveraging both education technology and learning games to improve learning outcomes. Indeed, there is no shortage of theoretical, conceptual, and practical reasons put forth in the literature for the wide-scale adoption of education technology in U.S. classrooms (Atkins et al., 2010). Although there is a major push to integrate technology products with regular classroom practice (Atkins et al., 2010), an important missing element is a solid evidentiary basis for wide-scale dissemination of existing products (Dynarski et al., 2007). Moreover, there are relatively few rigorous programs of research to develop and study technology tools and interventions in the area of early mathematics (Doabler, Fien, Nelson-Walker, & Baker, 2012; Young et al., 2012). For example, the What Works Clearinghouse (WWC) has reviewed nearly 100 elementary mathematics programs to date, a quarter of which are technology programs. Of the technology programs reviewed, only a few have research studies that meet WWC screening criteria to determine the level of evidence available and the degree of program effectiveness. Less than 10% of technology programs reviewed have any research that can be used to evaluate program efficacy. Of that small subset of reviewable programs, only two (Odyssey Math and Dreambox Learning) demonstrated positive or potentially positive effects on mathematics achievement for students in elementary school. This has bearing for schools and teachers looking for supplementary mathematics interventions that effectively differentiate instruction according to the standards they are tasked to address.

In addition, results of large-scale evaluations of education technology tools have been mixed, at best. For example, Dynarski et al. (2007) conducted a large-scale cluster-randomized controlled trial, randomly assigning 428 teachers to one of 16 reading or mathematics technology programs or control classrooms that did not have access to either set of technology programs. The researchers were interested in testing the average effect of teachers having access to education technology tools, not the effect of any particular technology program. Overall, compared to control-group classrooms, test scores were not significantly higher in classrooms using the selected reading and mathematics software programs.

In a follow-up to the Dynarski et al. (2007) study, Campuzano, Dynarski, Agodini, and Rall (2009) sampled 176 classrooms from the original sample of 428 classrooms to further study the effects of 10 technology programs on student learning. Of the 10 programs studied, the researchers found a small, significant effect for only one program in one grade level: Leap Frog for fourth-grade reading (Hedges's $g = .09$). There were no significant effects on mathematics learning for any of the mathematics programs (Campuzano et al., 2009). Together, the Dynarski et al. (2007) and Campuzano et al. (2009) results suggest that, despite the capacity of technology to create dynamic, individualized learning opportunities for students, the potential benefits of technology are not being realized in classrooms, especially those teaching mathematics.

Theoretical and Empirical Support for the NS1 Intervention Components

We believe a primary reason for the general lack of positive outcomes for education technology programs to improve student learning is a major disconnect between cutting-edge technology and empirically validated instructional design features. For example, mathematics content and instructional design features that have been demonstrated as efficacious in print-based curricula for students with or at risk for MD could be carefully integrated with education technology as one approach to realize the potential for education technology tools (Baker, Gersten, & Lee, 2002; Clarke, Baker, Chard, Smolkowski, & Fien, 2008; Clarke, Baker, & Fien, 2009; Dede, 2009; National Council of Teachers of Mathematics [NCTM], 2006; NMAP, 2008). In this context, we hypothesize that the careful integration of research-based instructional design principles, early mathematics content, and gaming technology could improve student mathematics achievement. Therefore, NS1 incorporates three design components: (a) evidence-based, explicit instructional design and delivery features (Baker, Fien, & Baker, 2010; Coyne, Kame'enui, & Carnine, 2011; Doabler et al. 2012); (b) critical early mathematics content focused on key whole-number concepts and skills identified in the Common Core State Standards for Mathematical Practice (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; Gersten, Beckmann, et al, 2009); and (c) a highly engaging gaming platform (Dede, 2009). Each of these components is described in detail below.

Evidence-Based Explicit Instructional Design and Delivery Features

A converging body of empirical evidence suggests that explicit, systematic mathematics instruction significantly enhances the mathematics achievement of students with or at risk for difficulties and disabilities and provides them with meaningful access to critical mathematics content (Baker et al., 2002; Gersten, Beckmann, et al., 2009; Gersten, Chard, et al., 2009; Kroesbergen & Van Luit, 2003; NMAP, 2008). For example, Gersten, Chard et al. (2009) conducted a meta-analysis and reviewed 41 studies that focused exclusively on students with learning disabilities and only included intervention studies that employed randomized controlled trials or strong quasi-experimental research designs. The authors found that explicit instruction had the largest impact, $g = 1.22$, 95% CI [0.78, 1.67], among seven dimensions of math instruction.

Likewise, Kroesbergen and Van Luit (2003) conducted a meta-analysis that included interventions that targeted a broader range of students with special needs, including at-risk students, students with learning disabilities, and low-achieving students. In a review that included 58 mathematics intervention studies that employed within-subject (e.g., single-case designs) and between-subject research designs (e.g., RCTs and quasi-experiments), the authors found that teacher-led, direct instruction was one of the most effective means for supporting mathematics learning for kindergarten and elementary-aged students, particularly for learning foundational math concepts and skills.

Baker et al. (2002) conducted a research syntheses using meta-analytic techniques including a set of 15 intervention studies that targeted low-achieving, school-aged students. The authors only included studies that used experimental or quasi-experimental group research designs. Results demonstrated an aggregate weighted effect size of $d = .58$, 95% CI [0.40, .77] for mathematics interventions that used an explicit instructional approach. More recently, the authors of the IES Practice Guide, *Assisting Students Struggling with Mathematics: Response to Intervention for Elementary and Middle Schools* (Gersten, Beckmann, et al., 2009), concluded that "explicit and systematic instruction" in both Tier 2 and Tier 3 contexts had a strong level of evidence. This evidence was based on findings from six randomized controlled trials (RCTs) that demonstrated various components of explicit instruction were present in interventions that were shown to be efficacious for students with and at risk for MD. In summary, a number of meta-analyses, research syntheses, and RCTs have demonstrated positive effects for explicit instruction to effectively teach students across grade levels and across varying samples of students (i.e., low achieving, at risk, students with learning disabilities).

NS1 was carefully developed using explicit instructional design and delivery principles (Coyne et al., 2011) drawn from these meta-analyses and research syntheses focused on low achievers and students with MD (Baker et al., 2002; Gersten, Chard, et al., 2009). The goal of explicit instruction is to present and communicate new information in a manner that is unambiguous and easy to understand (Archer & Hughes, 2011). Explicit instruction makes extensive use of teacher modeling and overt think-alouds to illustrate key concepts and the successful use of skills and strategies. Such approaches also make the otherwise obscure cognitive processes used by proficient learners during skill acquisition apparent to at-risk learners (Gersten, Chard, et al., 2009). The explicit instructional framework of NS1 centers on the following three interrelated design principles.

Utilize Instructional Scaffolding

Instructional scaffolding is a strategy used to provide support to students at critical points during instruction through the systematic introduction and integration of key concepts, skills, and strategies (Chard & Jungjohann, 2006). For example, instruction will initially begin with simpler instructional examples and progress to more complex examples over time as students demonstrate conceptual understanding. As students progress in their understanding of concepts, supports are progressively and systematically faded to promote learner independence (Coyne et al., 2011). In NS1, for example, instruction on new and complex mathematics content begins with virtual representations (e.g., place-value models) and transitions to more abstract representations (i.e., numbers) as students demonstrate proficiency.

Provide Opportunities for Guided Practice, Feedback, and Review

Often, students have difficulty performing procedural skills because they have not yet mastered prerequisite concepts in the instructional sequence (Hudson & Miller, 2006) and have not been taught to a high criterion level of performance (Coyne et al., 2011). Therefore, to become proficient in the application of newly taught skills and strategies, students need multiple opportunities to practice with immediate, highly specific, academic feedback (Clarke, Doabler, Nelson, & Shanley, 2015; Doabler et al., 2015; Gersten, Beckmann, et al., 2009). Research on mathematics instruction indicates that frequent, well-designed, guided practice opportunities help students develop conceptual knowledge and attain automaticity with critical skills and procedures (Kilpatrick, Swafford, & Findell, 2001). In NS1, guided and independent practice opportunities are also deliberately sequenced to gauge initial learning and ongoing retention of new and previously learned concepts and skills (Coyne et al., 2011). Practice opportunities, as operationalized in NS1, consist of a student responding to a mathematics-related question or task. For example, in an activity focused on using symbols (i.e., $<$, $>$, $=$) to compare two-digit numbers, NS1 would record a practice opportunity for every attempt made by a student to determine the relative magnitude of a targeted number.

Provide Differentiated Support to Learners

The benefits of providing differentiated and intensive support to students who struggle with early mathematics concepts has strong empirical support (Clarke, Doabler, et al., 2015; Doabler & Fien, 2013; Gersten, Beckmann, et al, 2009). NS1 employs a Differentiated Learning Pathway (DLP) to individualize and intensify instruction for students struggling to master mathematics content. The DLP precisely calibrates instruction matched to students' skill levels based on metrics of their mathematics performance, operationalized by latency and accuracy scores in game activities. The DLP has the capacity to make within- and between-activity adjustments based on student performance data. These adjustments reroute students to optimal activities for improving procedural skills and building conceptual knowledge. For example, if the player reaches a performance criterion of 90% or greater on an independent practice activity, she continues on the *default gameplay pathway*, following the standard scope and sequence. If the player scores 75%–89% accuracy, she is routed to the *additional practice pathway*, which provides the player with additional practice on a previously mastered, prerequisite skill designed to build knowledge required to access the target skill in the default pathway. If the player scores less than 75% accuracy, she is routed to the *additional instruction and guided practice pathway*, a more intensive, individualized experience that provides reteaching and practice with the target skill.

Critical Early Mathematics Content Focused on Whole-Number Concepts

There is consensus among experts that mathematics interventions in the early grades should focus intensely on building understanding of whole-number concepts (Gersten, Chard, et al., 2009). Therefore, NS1 was strategically designed to build number sense and facilitate proficiency in three whole-number concept domains specified by the CCSS-M (NGA Center for Best Practices & CCSSO, 2010): Counting and Cardinality, Number and Operations in Base Ten, and Operations and Algebraic Thinking.

Counting and Cardinality

Counting and cardinality refers to students' knowledge of the relationship between numbers and quantities and thus lays the foundation for building students' number sense (NGA Center for Best Practices & CCSSO, 2010). As a broader construct, number sense encompasses a child's fluidity with and flexibility in using and manipulating numbers, ability to perform mental mathematics, and capability to make quantitative comparisons without difficulty (Berch, 1998; Gersten & Chard, 1999). Number sense in combination with numeration skills (e.g., number identification, one-to-one correspondence, counting, understanding of the number line) sets the stage for students to solve foundational arithmetic problems. The early lessons of NS1 target topics from the Counting and Cardinality domain to prime students' background knowledge and better ensure their success with more advanced topics addressed in the later lessons. For example, in the initial lessons of NS1, students learn how to "count on" from numbers other than 1 to gain proficiency with efficient and sophisticated counting strategies for solving computational problems, such as number combinations (Gersten, Jordan, & Flojo, 2005; Fuchs et al., 2010).

Number and Operations in Base Ten

Key to understanding whole numbers is recognizing the convention and structure of the base ten system (Cawley, Parmar, Foley, Salmon, & Roy, 2001; NGA Center for Best Practices & CCSSO, 2010; Van de Walle, 2001). To develop knowledge of place value, NS1 teaches concepts such as ten-to-one relationships, the position of digits in two-digit numbers to determine their value, and the groupings of ones and tens to compose and decompose two-digit numbers. Mastery of these place-value concepts and skills represents a critical bridge to related mathematics topics. For instance, NS1 explicitly teaches students how to compose and decompose two-digit numbers (e.g., 16 is made up of 1 ten and 6 ones) so that they can acquire the conceptual groundwork for solving multidigit addition and subtraction problems.

Operations and Algebraic Thinking

The third domain emphasized in NS1 focuses on Operations and Algebraic Thinking. To extend students' understanding of whole numbers, NS1 promotes fluency of number combinations within 20, understanding of the equal sign and properties of operations, and solving word problems. Because students with or at risk for MD typically struggle to solve mathematics word problems (Bryant & Bryant, 2008), a primary focus of NS1 is to promote a robust understanding of word problem solving by teaching the structural features underlying different problem types. For example, students are taught how to solve word problems that require one- and two-step solutions and involve *unknowns* (e.g., addends) in all positions.

Gaming Platform

Motivation and engagement techniques for fostering the development of mathematics proficiency are particularly important for young students with and at risk for MD because they have often experienced a long line of "failure and frustration with math" (Gersten, Beckmann, et al., 2009, p. 44). Expert panels, therefore, recommend that Tier 2 and 3 mathematics interventions include motivational strategies such as (a) reinforcing or praising students for their effort and engagement in mathematics lessons and (b) rewarding student accomplishment (Gersten, Beckmann, et al., 2009; Woodward et al., 2012). The NS1 gaming platform was designed to increase students' engagement and motivation to learn mathematics by situating mathematics learning experiences within a rich and engaging narrative arc. For example, NS1 sessions are set in a Renaissance-style, fairy tale-inspired village called NumberShire. In the game, players assume the role of a young member of the village and engage in brief mathematics activities focused on building proficiency with whole numbers.

Gameplay allows players to interact with key NS1 characters and receive effort- and performance-based rewards as they succeed in solving mathematics problems, such as individualizing the attributes and attire of their gameplay character (i.e., avatar).

Purpose of the Study

Our NS1 research and development work has been enveloped in a robust line of incremental research activities (Gause et al., 2011). In the early years of our Small Business Innovative Research (SBIR) project, a primary aim was to develop prototypes of the intervention using an iterative design framework (Clements, 2007; Doabler et al., 2015). Once the lessons were fully developed and compiled into a complete intervention, we then assessed the feasibility and usability of NS1 through a series of implementation studies (Nelson et al., 2014). Various data sources, such as professional and preferential feedback from teachers and students, were collected during these implementation studies and used to guide major revisions of the intervention. Our next objective, and the focus of the current study, was to test the promise of NS1 under rigorous experimental conditions. Due to the need to conduct a final round of intervention revisions at the end of this pilot study and our SBIR funding, we were limited to testing only a portion of the NS1 intervention (i.e., 8 weeks of the full 12-week intervention). Therefore, the primary goal of this manuscript is to report the promise of an abbreviated version of NS1 to improve student proximal and distal mathematics outcomes.

Research Questions and Hypotheses

This study was guided by three research questions:

1. *What are the statistical and practical effects of NS1 on student mathematics outcomes?* We hypothesized that students in the NS1 treatment condition would outperform their peers in the control group on the proximal mathematics measure. Additionally, we anticipated that mean trends would favor the treatment group on the distal mathematics measures, but those differences would not reach statistical significance. The rationale for this hypothesis was based on the abbreviated dose of the intervention (8 weeks instead of the full 12 weeks) and the relatively low to modest statistical power for the targeted distal mathematics measures.
2. *Is treatment effect moderated by prior achievement or English language learners (ELLs) status?* NS1 was developed as a supplemental Tier 2 intervention; therefore, we hypothesized that it would benefit the entire treatment sample but be particularly beneficial for students at the upper end of the at-risk sample. Because the NS1 intervention carefully controls for mathematics language and vocabulary, we anticipated that ELLs would benefit commensurately with their non-ELL peers. In other words, we expected no moderation effect for ELL status on the treatment effect.
3. *Given variability in treatment dosage, is there a relationship between treatment exposure and response to the intervention, or in other words, is there evidence of dose response?* We expected a positive relationship between treatment exposure (i.e., measured as number of student practice opportunities in the NS1 game) and student responsiveness to the NS1 intervention.

Method

Design

This study used a randomized controlled trial (RCT) design to test the promise of the NS1 intervention. In total, 250 first-grade students were randomly assigned within classrooms to the treatment condition ($n = 125$) or a control condition ($n = 125$). Students assigned to the treatment condition received the NS1 intervention in addition to the core mathematics instruction provided in their first-grade classroom. For students randomly assigned to the control condition, schools provided core mathematics instruction and a variety of supplemental mathematics interventions. Student mathematics achievement data was collected at pretest and posttest and approximately every two weeks during intervention gameplay. Data from direct observations and project surveys were also analyzed.

Participants and Procedures

Schools

In the fall of 2013, the principal investigators invited nine schools from two school districts in different regions of Oregon to participate in the study. All of the first-grade classrooms ($n = 26$) from the nine schools expressed interest in participating in the study. Eleven of the classrooms were set in five Title I schools, located in a suburban school district (District A) in the second largest city in the state (Eugene, Oregon). In District A, 57% of students received free or reduced-price lunch, 19.6% received special education services, 3.1% were considered ELLs, and 24.1% identified as ethnic minorities. The remaining 15 classrooms were set in four schools in a suburban school district (District B) located in the Portland metropolitan area. Of the four schools in District B, three received Title 1 funding. In District B, 35.5% of students received free or reduced-price lunch, 13.2% received special education services, 14.8% were considered ELLs, and 39.3% identified as ethnic minorities.

Teachers

Twenty-six certified teachers delivered core mathematics instruction in the participating first-grade classrooms. Teachers, who were predominantly female and White, reported an average of 16.1 years of experience teaching mathematics and 9.5 years providing interventions for students at risk for MD. Teachers had formal training in education, with all teachers reporting coursework in mathematics education and nearly half reporting other graduate-level coursework in mathematics. Teachers indicated that they had a range of experience using technology in the classroom, including computers and laptops. Of the 26 teachers, only two reported using technology-based mathematics interventions in their current classrooms.

Interventionists

NS1 was facilitated by 10 interventionists, of which nine were district-employed instructional assistants, and one was a regular parent volunteer at a participating school. Participating interventionists were predominantly female, and just over half were White, while the remaining interventionists were Hispanic or Latino. They reported an average of 6.9 years of experience working in schools and 3.3 years providing interventions for students at risk for MD. The majority of interventionists did not have formal training in mathematics or education. Interventionists indicated that they had varied experiences using technology, with nearly all reporting that they used computers to support instruction. One-third of interventionists reported using technology-based mathematics interventions prior to participating in the study.

Students

Eligibility for the NS1 intervention used a three-stage process. First, 632 consented first-grade students enrolled in the 26 participating classrooms were screened in fall 2013. Students were screened using the fall benchmark of the easyCBM-Common Core State Standards (easyCBM-CCSS; Alonzo, Tindal, Ulmer, & Glasgow, 2006) assessment. In each classroom, the 10 students with the lowest scores on the easyCBM-CCSS were identified as NS1-eligible and then matched in pairs according to their scores. Each participating classroom had five pairs of NS1-eligible students. For example, the two lowest performers on the easyCBM-CCSS formed the first pair, while the students who were rank-ordered in the ninth and tenth positions formed the fifth pair. A total of 250 students were determined eligible for the intervention. According to easyCBM CCSS national norms, approximately 60% of the sample of the eligible students ($n = 151$) scored at or below the 25th percentile, and 97% of the eligible students ($n = 243$) scored at or below the 50th percentile. The remaining seven students scored between the 50th and 75th percentile.

Next, all NS1-eligible students were pretested on a battery of pretest measures, including the fall benchmark of the easyCBM-National Council of Teachers of Mathematics (easyCBM-NCTM; Alonzo et al., 2006) assessment, Group-Administered Missing Number (GA-MN), and Group-Administered Quantity Discrimination (GA-QD) measures (Doabler et al., 2015), and a researcher-developed mastery assessment of the NS1 intervention. Following pretesting, pairs of NS1-eligible students within each classroom were randomly assigned to one of two conditions: treatment (i.e., NS1 intervention) or a control condition (i.e., "business as usual" mathematics interventions). One classroom was unable to comply with random assignment and was dropped from the study. Consequently, randomization resulted in 125 students in each condition.

Table 1 displays student demographic information by condition, reported by school districts at the beginning of the study. Across both conditions, participating students were predominately White, and the average age of students was 6.5 years. As displayed in Table 1, treatment and control conditions were similar in demographic characteristics.

Mathematics Instruction

Core Mathematics Instruction

All treatment and control students continued to receive district-approved core mathematics instruction during the eight-week study. Teachers reported that core mathematics instruction was delivered, on average, one hour per day, five days per week. Across the participating first-grade classrooms, the types of instructional materials varied. Of the 26 classrooms, 16 used a commercially available mathematics program as a primary mode of instruction. The most popular curricula used were Everyday Mathematics ($n = 12$) and Saxon Mathematics ($n = 2$). Six of the classrooms used instructional materials developed by their respective school districts. Teachers reported that core mathematics instruction focused primarily on concepts and skills from the CCSS-M domains of Operations and Algebraic Thinking and Numbers and

Table 1. Treatment and control condition comparison.

Project conditions	Treatment	Control
Distinguishing features		
Supplemental Intervention	NumberShire Level 1	Business-as-Usual (BAU)
Setting	Computer lab	Push-in and pull-out group settings
Management	Facilitated by instructional assistants	Taught by teachers and instructional assistants
Delivery	Gaming platform	In-person
Materials	Technology-based	Print-based with incidental use of technology
Format	Individual	Individual and small group
Content	CCSS-M, whole number concepts	CCSS-M, variety of skills
Focus	Explicit instruction	Other methods
Minutes	15 per session	15–30 per session
Frequency	4 times per week	4 times per week
Support	Initial facilitation training, plus ongoing support as needed	Typical district supports for initial training
Shared features		
Treatment and Control	Core math instruction delivered by classroom teacher Consent, random assignment, incentives All project assessments at all time points	

Operations in Base Ten. The primary instructional formats used for core mathematics instruction included teacher-led instruction, peer and group work, and independent student work.

NumberShire Level 1 Intervention

NS1 is a game-based, Tier 2 mathematics intervention designed to support first-grade students with or at risk for MD in developing proficiency with whole-number concepts and skills. Three domains of whole numbers represented in the CCSS-M (NGA Center for Best Practices & CCSSO, 2010) are targeted in the intervention: (a) Counting and Cardinality, (b) Number and Operations in Base Ten, and (c) Operations and Algebraic Thinking. Specifically, NS1 provides explicit instruction in rational counting, decomposition of numbers, sophisticated counting strategies, properties of operations, number combinations, multidigit addition and subtraction, and word problem solving.

NS1 consists of 48 sessions, themed into 12 weeks of instruction (four sessions per week). Each session is designed to provide 15 minutes of instruction. In total, the intervention offers students 12 hours of individualized instructional gameplay.

To promote mathematics proficiency among at-risk learners, NS1 utilizes an explicit instructional framework. The intervention offers explicit modeling to demonstrate exactly what students are expected to learn, scaffolded instruction to guide students through the learning process, and independent practice to facilitate learner independence. Each session is organized into four mathematics activities: Warm-Up, Teaching Event, Assessment Event, and Wrap-Up. The purpose of the Warm-Up is to allow students to practice a previously mastered concept or skill. The Teaching Event introduces new whole-number concepts and skills that are central to the session's learning objective. For this part of the session, NS1 characters provide vivid demonstrations of the targeted concept or skill and offer clear explanations of how students are expected to interact with the activity. Students then practice the skill with support from NS1 characters, before practicing the skill independently. The third and fourth activities, Assessment Event and Wrap-Up, provide extended practice to review whole-number concepts and skills. All four activities include a variety of virtual mathematical representations (e.g., number lines, base-10 blocks), frequent practice opportunities, and high-quality academic feedback to facilitate students' procedural fluency and build conceptual understanding of whole-number concepts.

NS1 Session Delivery. The intent of this study was to deliver the NS1 intervention four days per week for eight weeks. Students received NS1 in addition to the core mathematics instruction provided in their respective first-grade classrooms. Trained interventionists were paid by the project to facilitate the NS1 intervention in school computer labs. Based on the availability of interventionists and space in the computer labs, the number of treatment students who received NS1 at one time varied by school, with intervention groups ranging between 5 and 25 students. On average, treatment students received 15 minutes of NS1 instruction per session.

At the start of each NS1 session, interventionists had students use a project-assigned password to sign on to the intervention. Once logged on, students selected their NS1 avatar and then completed the session's Warm-Up, which lasted an average of two minutes per session. The Warm-Up activities, which take place in the avatar's village house, have a primary focus on building students' fluency with basic number combinations. During the number combination activities, students received academic feedback about their accuracy with the targeted problems.

After the Warm-Up, students encountered a 5–7 minute Teaching Event. To enact this activity, the student's avatar selected an object or building within the village of Tally-Ho. The Teaching Event entailed an NS1 character introducing students to a new concept or skill by offering step-by-step demonstrations and detailed explanations. For example, in the seventh session, Thatcher Tom initially demonstrates how the value of digits in teen numbers depends on their place in the target number (16 is made up of 1 ten and 6 ones). To keep students engaged in these teaching moments, students were tasked with assisting NS1 characters in setting up the demonstrations (e.g., by identifying the underlying problem structures in Teaching Events focused on word problem solving).

Following the Teaching Event, students spent approximately 3–5 minutes in the Assessment Event and 2 minutes in the Wrap-Up activity. Combined, these final activities offered students additional practice to build deep understanding of whole-number concepts and skills. The Assessment Event reviewed the concept or skill introduced in the previous session. Primary aims of the Wrap-Up activities included strategic counting and number writing. At the conclusion of each session, the student's avatar was allowed to select a virtual reward and use it enhance the aesthetics of their personalized Tally-Ho village. Figure 1 provides screenshots of the Tally-Ho Village and characters from the game as well math models used within the game.

NS1 Training. Prior to the start of the study, all interventionists received four hours of professional development comprised of a two-hour training presentation delivered by research staff and two one-hour, site-based meetings with the project team. Professional development focused on preparing interventionists to (a) efficiently facilitate intervention groups, (b) troubleshoot and solve technical problems with the computers (e.g., resetting a student's computer during a session), and (c) monitor student progression during gameplay. It is important to note that interventionists did not provide any instructional assistance during the gameplay sessions. Interventionists were shown how to help students with the sign-on process at the start of intervention sessions and actively monitor students during gameplay.

Figure 1. Screenshots from the NumberShire Level 1 online interactive mathematics intervention game.

Project staff also familiarized interventionists with the structure of NS1, describing the mathematical content contained within the intervention and the rationale behind the use of an explicit instructional framework and gaming technology to teach mathematics.

Fidelity of Implementation. Project staff directly observed each NS1 intervention group once during the eight-week study using the Technology Observation Tool (TOT; Nelson & Doabler, 2013). The TOT is a researcher-developed, standardized protocol designed to assess fidelity of implementation of the NS1 intervention. Project staff observed and rated each of the intervention session sites (e.g., computer labs) on six items of implementation fidelity: (a) use of effective procedures at start of gameplay, (b) students use headphones during gameplay, (c) student engagement, (d) active monitoring and classroom management, (e) troubleshooting of technological issues, (f) use of effective procedures at end of gameplay. All items were rated on a 4-point scale (1 = *not present*, 4 = *highly present*) and were averaged to compute an overall implementation fidelity score. The average fidelity ratings for interventionists' use of effective procedures at the start of the session, active monitoring during student gameplay, and use of effective procedures at the conclusion of the session were 3.5 ($SD = 1.1$), 3.6 ($SD = 0.7$), and 3.2 ($SD = 1.2$), respectively. Interventionists also received an average rating of 2.6 ($SD = 1.1$) for troubleshooting technology issues during sessions. Observers rated students' engagement during gameplay and use of headphones, a critical component of NS1, as 3.7 ($SD = 0.7$) and 3.2 ($SD = 0.8$), respectively. The average overall fidelity score was 3.3 ($SD = 0.8$), indicating moderate overall fidelity with substantial variability between NS1 groups.

Metrics gathered during NS1 gameplay served as an additional measure of fidelity of implementation, including number of sessions completed, number of items completed, and latency and accuracy in responding. Between pretest and posttest treatment students, on average, completed 18.6 game sessions or 4.7 weeks of gameplay ($SD = 8.1$ sessions, range = 2 to 33 sessions) and repeated 12.8 game sessions ($SD = 5.9$, range = 2 to 24). During gameplay treatment students completed an average of 499.8 practice opportunities ($SD = 269$, range = 41 to 1,156) and completed 69% of the practice opportunities correctly ($SD = 12\%$, range = 37% to 90%). Project staff used gameplay metrics to track student progress through game sessions on a weekly basis and corresponded regularly with interventionists to provide support when needed (e.g., when a student's gameplay progression deviated from the standard schedule of four sessions per week).

Control Condition

Students who were randomly assigned to the control condition received "business-as-usual" Tier 2 mathematics intervention supports in addition to core instruction for the duration of the eight-week study. Reported mathematics interventions for control students included several commercially available mathematics programs, a district-developed core program, teacher-developed materials, and other intervention resources. Everyday Mathematics was reported as the supplemental intervention program implemented in six classrooms. TouchMath and SRA Explorations and Applications were each used as supplemental intervention programs in two participating classrooms. Seven of the classrooms reported using a district-developed core program as their Tier 2 intervention. Teachers from the remaining classrooms reported using a host of intervention resources, including mathematics facts worksheets and a program that focused on calendar concepts to provide supplemental intervention to students in the control group.

Collectively, interventions for control students emphasized instruction in the CCSS-M, addressing the domains of Operations and Algebraic Thinking, Counting and Cardinality, and Number and Operations in Base Ten. Twelve of the teachers reported that the control group interventions prioritized teaching students the names of numbers and the appropriate count sequence. Other commonly taught skills were counting to tell the number of objects, comparing numbers, and understanding concepts of addition and subtraction. Working with numbers 11–19 to gain foundations for place value was the least prioritized topic reported for the control condition. All control-group interventions were teacher-led, but many involved peer or independent work as part of the intervention. Teachers reported that control students in three classrooms received additional support with technology-based interventions, including the software program *IXL* and games available through a leading publishing company (SRA). Students in the control condition received an average of 24 minutes of supplemental mathematics intervention per day, four times per week (see Table 1 for distinguishing features of the NS1 treatment and control conditions).

Student Measures

Trained project staff administered a battery of student mathematics assessments during the course of the study. Researcher-developed measures were used to assess mathematics learning proximal to the NS1 intervention, while other established measures of procedural fluency and conceptual understanding were administered to assess general mathematics performance. Researcher-developed mastery tests were administered approximately every two weeks to treatment and control students according to treatment students' progress in the intervention. Student mathematics achievement data were collected at pretest and posttest.

easyCBM Mathematics

EasyCBM Mathematics (Alonzo et al., 2006) is a standardized, individualized, computer-administered assessment for students in kindergarten through eighth grade. There are currently two versions available commercially—an early version of the assessment developed to assess the National Council of Teachers of Mathematics Focal Points (easyCBM-NCTM), and a subsequent version of the assessment developed to measure performance in the CCSS-M (easyCBM-CCSS). EasyCBM-CCSS was administered at the beginning of the study as a screening assessment to determine eligibility for NS1 because the majority of participating schools were already using this assessment for benchmarking classroom-wide. Once NS1-eligible students were identified, the easyCBM-NCTM was administered as a pretest and posttest measure of general mathematics performance.

easyCBM-CCSS

The CCSS version (Wray, Alonzo, & Tindal, 2014) of the assessment at each grade level includes 48 items developed to assess the CCSS-M. Measures are untimed, but the estimated administration time is 18–30 minutes. Results from K–8 studies of the technical adequacy of the easyCBM-CCSS indicate measures have strong internal consistency ($\alpha = .90$) and split-half reliability (.80 first half, .86 second half: Wray et al., 2014).

easyCBM-NCTM

Each grade level of the NCTM version of the assessment includes 15 items to assess each of three NCTM focal points at each grade level. In first grade, the focal points are Numbers and Operations; Numbers, Operations, and Algebra; and Geometry. Measures are untimed, but the estimated administration time is 18–30 minutes. The internal consistency of the mathematics measures in Grade 1 is strong (α range = .78–.89) and the concurrent validity correlation with the Terra Nova in first grade is .73 (Anderson et al., 2010).

Group-Administered Missing Number (GA-MN) and Group-Administered Quantity Discrimination (GA-QD)

The Group-Administered Missing Number (GA-MN) and Group-Administered Quantity Discrimination (GA-QD) measures (Doabler et al., 2015) were administered at pretest and posttest. GA-MN and GA-QD represent modified versions of the Early Numeracy–Curriculum Based Measurement subtests: Missing Number and Quantity Discrimination (Clarke & Shinn, 2004). The GA-MN and GA-QD are one-minute, fluency-based measures that assess strategic counting and magnitude comparison, respectively. Unlike traditional mathematics CBMs, which require verbal student responses, GA-MN and GA-QD have students record their responses through a written format. The GA-MN requires students to write in the missing number among a string of three numbers (0–10), with the first, middle, or last number of the string missing (e.g., 5_7). The GA-QD requires students to circle the number in a pair (numbers 0–10) with the higher value. Test–retest reliability for GA-MN and GA-QD is reported, respectively, at .85

($p < .001$) and .87 ($p = .003$). Predictive validity coefficients for GA-MN and GA-QD range from .45 to .67 with a first-grade mathematics measure (Doabler et al., 2015).

ProFusion-Revised (ProFusion-R)

The ProFusion-R is a proximal measure of student mathematics learning that was developed and utilized in one of our previous intervention development projects (Clarke et al., 2014; Doabler et al., 2015). A combination of group and individually administered items assess student understanding of concepts in place value, decomposing and composing numbers, word problem solving, and single and multi-digit addition and subtraction, among other skills. Predictive validity correlations between ProFusion-R and GA-MN and GA-QD measures (Doabler et al., 2015) range from .49 to .61. Concurrent validity with the Stanford Achievement Test, 10th Edition Mathematics subtest is .68 (Clarke, Baker, et al., 2015). For the nontimed ProFusion-R items, the standardized item alpha = .93; for the timed ProFusion items, the average item-total correlation was .51. The ProFusion-R was administered at pretest and posttest to assess procedural fluency and conceptual understanding in mathematics aligned with CCSS-M standards and domains taught during the NS1 intervention. At pretest, students were characterized as having low initial skills if they scored between 0 and 33 on the Profusion-R measure and having high initial skills if they scored equal to or higher than 34 on the measure.

NumberShire Level 1 Mastery Tests (NMT)

Mastery tests were developed by the research team to assess student learning every two weeks during the NS1 intervention. Items included in each NMT were constructed to measure only the standards taught and practiced in the two-week period preceding administration. Approximately three items per standard were included in each NMT. NMTs were administered every two weeks of the study, up through week six.

Surveys

A survey was also administered during the study to gather data about participant experiences and document practices employed in core and supplemental mathematics instruction delivered to students in the treatment and control groups.

Demographics and Perceptions

At the end of the study, teachers and interventionists completed a demographics survey to document background information, previous teaching experience, and perceptions of NS1. The survey asks interventionists to self-report training and experience teaching mathematics and using technology (e.g., level of education, mathematics education certifications, frequency of use of technology interventions in the classroom).

Instructional Practices

At the end of the study, teachers were also asked to describe core and supplemental mathematics instruction provided to all NS1-eligible students. The survey asks teachers to describe the programs used, their content focus, the type of instructional strategies used during instruction, and the frequency and duration of instruction.

Statistical Analysis

Univariate effects of intervention condition on posttest outcome measures were examined using between-subjects analysis of covariance (ANCOVA), adjusting for pretest scores. Intervention effects on the two-, four-, and six-week interim mastery tests were evaluated using ANCOVAs, adjusting for pretest ProFusion-R total score as a covariate. Next, we tested for differential effects of condition by Special Education (SPED) status, ELL status, initial skill level, and student engagement. For tests of differential effects of condition, we extended the primary ANCOVA models to include the main effects and cross product of condition and the proposed moderator. Pearson's *r* correlation coefficients were used to explore associations between number of sessions completed and change in outcomes from pretest to posttest among students assigned to the treatment condition. All analyses were conducted with SPSS 21, and alpha was set to $p < .05$, two-tailed, for all tests.

Hedges's *g* was reported as a metric of intervention effect size (What Works Clearinghouse, 2008; effects of .25 and above are considered "substantively important"). Hedges's *g* was computed as the difference between the covariate adjusted means of the two groups at posttest divided by the posttest pooled standard deviation of the outcome. For our primary outcome analyses, we also reported partial η^2, the proportion of variance explained by condition, to facilitate interpretation and future meta-analyses.

Despite common misconceptions that multilevel modeling is required for education intervention research, the student-level analyses reported here are appropriate given that students were the unit of randomization and that the intervention was delivered at the student level. A recent IES guide to nested randomized controlled trials (Lohr, Schochet, & Sanders, 2014) acknowledged the confusion regarding when one must account for clustering effects. Consistent with Bloom, Bos, and Lee (1999), Roberts and Roberts (2005), Rubin (1974), and others, Lohr et al. indicate that individual randomization "essentially cancels out the pre-existing clustering effect from the original schools, just as it cancels out pre-existing effects from unobserved connections between the students such as belonging to the same church, softball team, or play group" (p. 34). Similarly, Raudenbush and Sadoff (2008), describe how individual randomization removes the effects of clusters on the average treatment effect. Because higher levels of nesting have no effect on the average effect estimator or its standard error for this study design (e.g., Bloom et al., 1999; Raudenbush & Sadoff, 2008), and consequently no effect on the Type I or Type II error rates (Murray, 1998), we have not included these levels in our analysis.

Results

Baseline Equivalence and Attrition

The expectation of baseline equivalence due to random assignment of groups was examined. The treatment and control groups were compared on demographic characteristics and outcome measures collected at pretest. Contingency table analyses and *t* tests were conducted on categorical and continuous measures, respectively. The groups did not significantly differ on any demographic characteristics (see Table 2 for demographic descriptive information). Compared to control students, treatment students performed significantly better on pretest GA-QD ($M = 21.1$, $SD = 7.5$ vs. $M = 19.0$, $SD = 7.6$; $t[235] = 2.18$, $p = .030$, $d = 0.28$) and the ProFusion-R ($M = 39.2$, $SD = 17.0$ vs. $M = 34.6$, $SD = 16.7$; $t[246] = 2.15$, $p = .032$, $d = 0.27$). To control for nonequivalence at baseline, these measures were included as additional covariates in all outcome analyses.

Table 2. Demographic characteristics by condition.

	Treatment (n = 125)	Control (n = 125)
Race/Ethnicity *n* (%)		
Asian	5 (4)	5 (4)
Black	6 (5)	10 (8)
Latino	26 (21)	29 (23)
Multiracial	10 (8)	5 (4)
White	77 (62)	76 (61)
Female *n* (%)	64 (51)	61 (49)
SPED *n* (%)	11 (9)	12 (10)
ELL status *n* (%)	31 (25)	28 (22)
Age *M* (*SD*)	6.5 (0.5)	6.5 (0.5)

Notes. *M* = Mean, *SD* = Standard deviation. Age was computed as of the beginning of the study (10/1/2013.)

The extent to which attrition threatened the internal and external validity of this study was evaluated using contingency table analyses and analysis of variance. Participants who completed a posttest assessment were compared to those who did not with respect to demographic characteristics and pretest outcome measures. We also conducted analyses to test whether outcome variables were differentially affected across conditions by attrition. These latter analyses examined the effects of condition, attrition status, and their interaction on pretest outcomes. Examination of attrition between pretest and posttest revealed that eight (6.4%) of the treatment participants did not complete a posttest assessment compared to four (3.2%) of the control participants. Attrition rates did not significantly differ by condition or demographic characteristics. Compared to students who completed a pretest and posttest assessment, students who did not complete a posttest assessment performed significantly worse on the ProFusion-R assessment at pretest ($M = 23.6$, $SD = 7.2$ vs. $M = 37.4$, $SD = 17.0$; $t[246] = 2.55$, $p = .011$). We found no statistically significant interactions between attrition and condition predicting baseline outcomes, suggesting that attrition was not systematic.

Intervention Effects

Table 3 provides means and standard deviations for each outcome by assessment time and condition, along with results of the outcome analyses. Statistically significant effects of treatment over control were obtained on the ProFusion-R ($p < .001$, partial $\eta^2 = .063$, Hedges's $g = 0.30$) and the two-week NMT ($p = .025$, partial $\eta^2 = .022$, Hedges's $g = 0.22$). Intervention effects were not statistically significant for other study outcomes.

We extended the primary ANCOVA models to test for differential effects of condition on primary outcomes (i.e., ProFusion-R, GA-MN, and GA-QD, and easyCBM-NCTM) by SPED status, ELL status, and initial skill level based on the pretest ProFusion-R assessment scores. We found no statistically significant interactions with condition ($ps > .224$), indicating no differential response to the intervention as a function of baseline student characteristics.

Dose Response

Pearson's r correlation coefficients were used to explore associations between program implementation metrics and change in outcomes from pretest to posttest among students assigned to the treatment condition. Change in outcomes was not significantly correlated with the

Table 3. Descriptive statistics and ANCOVA results for the outcome measures.

	Pretest	Posttest		Condition effect			
Outcome measure/condition	*M* (*SD*)	*M* (*SD*)	*Adj M*	*F*	*p*	Partial eta^2	Hedges's *g*
easyCBM-NCTM total raw				1.29	.257	.006	−0.13
Treatment	21.9 (5.3)	23.9 (6.8)	23.2				
Control	21.6 (4.4)	23.8 (6.8)	24.1				
Group EN-CBM quantity discrimination				0.32	.570	.001	0.07
Treatment	21.1 (7.5)	26.3 (7.5)	25.5				
Control	19.0 (7.6)	24.5 (6.7)	25.0				
Group EN-CBM missing number				0.47	.495	.002	0.08
Treatment	8.8 (5.1)	11.9 (5.3)	11.4				
Control	7.7 (4.5)	10.5 (5.0)	11.0				
Profusion-R				15.0	<.001	.063	0.30
Treatment	39.2 (17.0)	60.2 (19.5)	58.2				
Control	34.6 (16.7)	51.1 (19.3)	52.3				
2-week interim mastery test				5.07	.025	.022	0.22
Treatment	na	36.0 (11.0)	34.5				
Control	na	31.0 (10.8)	32.1				
4-week interim mastery test				2.57	.110	.011	0.17
Treatment	na	26.3 (12.2)	25.1				
Control	na	22.4 (10.3)	23.2				
6-week interim mastery test				0.24	.624	.001	0.06
Treatment	na	30.0 (9.3)	28.9				
Control	na	27.6 (10.2)	28.3				

Notes. *M* = Mean, *SD* = Standard Deviation, *Adj* = Adjusted. Baseline differences between conditions were observed on the en-CBM quantity discrimination and the Profusion-R assessment; therefore, all analyses included pretest scores on these measures as covariates in addition to the pretest score on the target measure. Analyses involving interim mastery test assessments included the primary proximal Profusion-R pretest score as a covariate because the interim measure was not assessed at pretest. na = not assessed at pretest.

number of sessions completed (*r*s ranged from −.10 to .07, *p*s > .270), number of repeated sessions (*r*s ranged from −.09 to .09, *p*s > .338), number of practice opportunities (*r*s ranged from −.04 to .18, *p*s > .070), or item response accuracy (*r*s ranged from .02 to .16, *p*s > .111).

Discussion

The present study was designed to test the promise of the NS1 intervention for improving the mathematical outcomes of students at risk for MD. Recognizing the shortage of rigorous studies of education technology tools, we chose to employ a randomized controlled trial to determine the statistical and practical effects of NS1, our first research aim. We hypothesized that at-risk students in the NS1 treatment condition would significantly outperform their at-risk counterparts in the control condition on proximal measures of whole-number concepts and skills that were directly targeted in the intervention. This hypothesis was based on our attention to incorporating evidence-based instructional design elements (Coyne et al., 2011) with an engaging gaming platform.

The results of the ANCOVA analyses suggest that NS1 treatment students outperformed the control students on the primary proximal measure of whole-number concepts and skills, demonstrating a moderate and practically important effect (approximately a third of a standard deviation difference). Treatment students also significantly outperformed their controls on the first of three interim mastery tests, demonstrating a .22 effect size difference. Although trends favored the treatment students on the second and third interim mastery tests, these effects were not significant. This finding is interesting and worth critical reflection. It may be that the NS1 intervention was less engaging over time, or it may be that the intervention was less effective in teaching treatment students to master skills and concepts in the last half of the program. It is also plausible that item difficulty was not constant across the interim researcher-developed mastery tests and warrants further research in subsequent studies of the NS1 intervention.

As hypothesized, we found no significant differences between treatment and control students on the distal outcome measures. Although the trend favored NS1 treatment students on two of the three distal measures, none of the effects were significant, nor were the magnitude of effects large enough to be considered potentially promising effects, as deemed by the WWC (e.g., Hedges's $g > .25$). Our hypothesis for null findings on the distal measures was based on three factors: (a) students received an abbreviated version of the full NS1 intervention, (b) our study was relatively small and underpowered to detect effects on measures with historically low to medium-sized effects in mathematics intervention research, and (c) the distal measures assess concepts that do not directly align with the content targeted in the NS1 intervention (e.g., measurement, geometry).

Although we do not want to overinterpret the results from this promise study, we do find it interesting that our positive results on proximal outcomes stand in stark contrast to the largely negative results from previous rigorous evaluation of education technology tools (Dynarski et al. 2007; Campuzano et al., 2009). However, we do believe that the findings of the current promise study do align with the slate of intervention studies that demonstrate the positive effect of explicit instruction approaches on the mathematics learning of students with early signs of math difficulties (Baker et al., 2002; Gersten, Beckmann, et al., 2009; Gersten, Chard, et al.. 2009; Kroesbergen & Van Luit, 2003). Further, the current findings support our hypothesis that features of explicit instruction that have been demonstrated as efficacious in print-based curricula for students with or at risk for MD could be carefully integrated with education technology to realize the potential for education technology tools (Baker et al., 2002; Clarke et al., 2008; Clarke et al., 2009; Nelson, Fien, Doabler, & Clarke, in press).

Our second research aim was to test the potential moderating effects of prior mathematics achievement and ELL status on treatment effects. It was hypothesized that ELL status would not moderate treatment effect because we believed ELLs, like their non-ELL peers, would benefit from the NS1 intervention due to our careful attention to precise mathematical language and vocabulary within gameplay. As predicted, results indicated that ELL status did not moderate treatments effects for any of the outcomes, suggesting that ELL students equitably responded to the treatment relative to their non-ELL peers. Our second moderation hypothesis posited that at-risk students at the upper end of the at-risk pretest distribution would benefit more from the NS1 than at-risk students at the lower end of the pretest distribution. We anticipated that the entire at-risk sample would benefit from NS1, but because we designed the intervention as a Tier 2 supplemental intervention, we conjectured that students at moderate risk might benefit more than students at significant risk for MD. However, this hypothesis was not supported by the analysis and we found no evidence for pretest score moderating treatment impact. Thus, NS1 appeared to be equally effective across the pretest distribution.

Our third and final research aim was to examine dose response patterns or the degree to which intervention gains were related to number of sessions completed, number of sessions repeated, or number of practice opportunities completed across NS1 gameplay. We hypothesized a positive relation between practice opportunities and strength of intervention gains. Surprisingly, we found no statistical relationship between any of the dosage variables and students' change in any of the outcomes from pretest to posttest. We believe our operational definition of dosage might be too crude, as currently defined, and that we may need to reconceptualize "dosage" and make it more precise to include the vast amount of gameplay metrics.

In summary, we believe that the NS1 has demonstrated some promise for improving student mathematical outcomes related to whole-number concepts and skills. These promising outcomes were found in a rigorous research design and in comparison to a relatively strong business-as-usual control condition—all participating schools had fairly sophisticated Tier 2 mathematics systems of support for at-risk students. As with any randomized experiment, it is important to consider the nature of the counterfactual in this study. Critical for understanding the magnitude of an observed treatment effect in randomized experimental designs is grasping the instructional events that occur in the comparison or control condition (Flay et al., 2005; Gersten, Baker, & Lloyd, 2000; Gersten et al., 2005; Shadish, Cook, & Campbell, 2002). The measurement of the comparison condition can provide critical information for generating and justifying causal inferences. It is therefore particularly important for researchers who use these types of designs to become knowledgeable about the comparison condition and examine the degree to which implementation of instruction differs between treatment and control conditions. Lemons, Fuchs, Gilbert and Fuchs (2014) noted that such implementation comparisons will likely lie somewhere on a continuum between no differences and stark overlap.

Consider an RCT study that tests the efficacy of a print-based, Tier 2 mathematics program (Intervention X) with 200 at-risk kindergarten students. The researchers assign 100 of the students to the treatment condition (Intervention X) and the remaining 100 to a control condition, which provided students with no mathematics intervention services. Using the continuum model suggested by Lemons and colleagues (2014), a comparison of instructional implementation would reveal no overlap between the two conditions. In this case, if Intervention X was soundly designed and implemented with high fidelity, one would expect treatment students to significantly outperform their control peers on most mathematics outcomes. This is largely in part because the control students were essentially assigned to a no-treatment condition.

However, if the researchers were interested in conducting a more rigorous study of Intervention X, they might increase the robustness of the comparison condition. For instance, they might have students assigned to the comparison condition receive a competing evidence-based, Tier 2 mathematics program (Intervention Z) that targets identical mathematics content and incorporates similar instructional design components to that included in Intervention X. Given the nature of Intervention Z, one could argue that there is at least modest overlap between it and the treatment intervention. Because of this overlap, the researchers would likely characterize the study's findings in a different light than studies that use a no-treatment condition. For example, if treatment effects were obtained, even minimal ones, then those results could arguably be considered substantively important given that Intervention Z has a previously established evidentiary basis for improving student mathematics outcomes. If treatment students performed equivalent to or no worse than their control peers, then the researchers might likewise characterize the findings as educationally meaningful.

While the control condition in our study implemented a host of intervention materials and instructional practices, one could argue that it represents a more robust comparison than a no-treatment control condition. For example, six of the classrooms used Everyday Mathematics, a program the WWC has deemed as having "potentially positive effects" on student mathematics achievement. The noted mathematics content overlap between NS1 and the control condition encourages us to use a more molecular approach in our future work to distinguish the similarities and differences between NS1 and the mathematics interventions commonly used in today's classrooms. We would expect these procedures to help us better understand the meaning of our observed effects.

Limitations and Directions for Future Research

We have several limitations worth noting in the present study. First, the purpose of the study was to examine the promise of the program—and not to document program efficacy. Although we employed a rigorous experimental design, we included a relatively modest sample and were sufficiently powered to detect effects on proximal measures, but not effects on distal measures. Because this promise study was nested within an iterative design framework, we continued making revisions to the NS1 intervention after the conclusion of the study to ready it for subsequent formal efficacy testing. Therefore, the positive, significant outcomes should only be viewed as preliminary support for the NS1 supplemental intervention. In addition, our moderation results must likewise be viewed as preliminary findings and further testing is warranted to examine if important subgroups respond similarly, or not, to the NS intervention.

A second factor that affects the generalizability of the findings is the setting in which the current study took place. We tested the promise of NS1 in nine schools from two school districts in the Pacific Northwest and the sample of students was not entirely representative of U.S. schoolchildren. Although the percentage of White students (61% in study, 60% U.S.) and Asian students (5% study sample, 4% U.S.) was commensurate with the larger population of U.S. students, we had a lower than average representation of Black students (5% study sample, 17% U.S.) and higher than average representation of Hispanic students (21% study sample, 17% U.S.; NCES, 2014). Therefore, it is unknown whether similar results would be found in a more representative sample of U.S. students or a sample that included higher than average Black students, for example.

A final limitation of the current study is that we have only begun to sort through the large amount of gameplay metrics to pose interesting research questions (e.g., patterns of play that predict responsiveness) and to verify certain assumptions for gameplay sessions (e.g., treatment integrity and adherence). For example, there may be interesting patterns of student engagement that may associate with responsiveness (or nonresponsiveness) that are not readily identified by direct observation. Our initial foray into operationalizing and measuring dosage as practice opportunities within gameplay sessions was not predictive of pre–post gains, and therefore, a more sophisticated conceptualization of dosage may be necessary to examine dose response or other interesting research questions related to responsiveness. In addition, conceptualizing and documenting such important constructs as treatment fidelity and treatment adherence could be transformed in the context of educational gaming interventions. Research studies commonly focus on teachers alone when collecting information on the fidelity of intervention implementation. Rather than regarding students as passive recipients of treatment, we encourage the field to begin to view students as active participants of interventions, particularly technology-based interventions.

To address some of these limitations, we propose to further extend these research activities by conducting a series of rigorous studies to document the efficacy of NS1 across more diverse samples and settings of students. Toward that end, we are submitting an Efficacy and Replication application to the Institute of Education Sciences National Center on Special Education's Education Technology topic area. We propose to implement the fully featured 12-week NS1 intervention with a much larger number of students from more diverse schools set in three regions of the United States (i.e., Las Vegas, NV, Boston, MA, and Portland, OR). We will be adequately powered to detect both proximal and distal outcome measures. Further, we propose to examine whether the significant outcomes demonstrated in the current study extend to diverse settings, and we propose to examine whether the treatment effect varies for important subgroups of students (e.g., ELL students, students with MD). In this way, we hope to (a) increase the use of rigorous methods endorsed and promulgated by the WWC and (b) provide invaluable information for practitioners and researchers seeking to implement evidence-based supplementary technology-based interventions in mathematics for students at risk for mathematics difficulties.

Conflict of Interest

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Hank Fien, Christian T. Doabler, Nancy J. Nelson, and Scott K. Baker are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain Numbershire-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services. An independent external evaluator and coauthor of this publication completed the research analysis described in the article.

Funding

This research was supported in part by Project NumberShire 1, Grant No. EDIES11C0026, a subcontract with ThoughtCycle through the U.S. Department of Education, Institute of Education Sciences, Small Business Innovation Research Program. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

ARTICLE HISTORY

Received 22 May 2015
Revised 3 November 2015
Accepted 9 November 2015

EDITORS

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

References

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *EasyCBM online progress monitoring assessment system*. Eugene: University of Oregon.

Anderson, D., Lai, C.-F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E.,… Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level mathematics measures (grades K-2), 2009–2010 version* (Technical Report No. 1006). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Archer, A. L., & Hughes, C. A. (2011). Exploring the foundations of explicit instruction. In A. L. Archer & C. A. Hughes (Eds.), *Explicit instruction: Effective and efficient teaching* (pp. 1–22). New York, NY: The Guilford Press.

Atkins, D. E., Bennett, J., Brown, J. S., Chopra, A., Dede, C., Fishman, B.,… Williams, B. (2010). *Transforming American education: Learning powered by technology*. Retrieved from http://www.ed.gov/sites/default/files/netp2010.pdf.

Baker, S. K., Fien, H., & Baker, D. L. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of Tier 1 and Tier 2 instructional supports. *Focus on Exceptional Children*, *42*(9), 1–20.

Baker, S. K., Gersten, R. M., & Lee, D.-S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal*, *103*, 51–73.

Berch, D. B. (1998, April). *Mathematical cognition: From numerical thinking to mathematics education*. Paper presented at the National Institute of Child Health and Human Development, Bethesda, MD.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review*, *23*, 445–469. doi:10.1177/0193841x9902300405

Bodovski, K., & Farkas, G. (2007). Do instructional practices contribute to inequality in achievement? The case of mathematics instruction in kindergarten. *Journal of Early Childhood Research*, *5*, 301–322.

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

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts* (Executive summary No. NCEE 2009-4042). Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Cawley, J. F., Parmar, R., Foley, T., Salmon, S., & Roy, S. (2001). Arithmetic performance of students: Implications for standards and programming. *Exceptional Children*, *67*, 311–328.

Chard, D. J., & Jungjohann, K. (2006). *Scaffolding instruction for success in mathematics learning, intersection: Mathematics education sharing common grounds*. Houston, TX: Exxon-Mobil Foundation.

Clarke, B., Baker, S. K., Chard, D., Smolkowski, K., & Fien, H. (2008). *Early learning in mathematics: Efficacy in kindergarten classrooms* (Institute of Education Sciences Grant). Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=599.

Clarke, B., Baker, S. K., & Fien, H. (2009). *Foundations of mathematical understanding: Developing a strategic intervention on whole number concepts* (Institute of Education Sciences Grant). Retrieved from https://ctl.uoregon.edu/research/projects/foundations-mathematical-understanding-developing-strategic-intervention-whole

Clarke, B., Baker, S. K., Smolkowski, K., Doabler, C. T., Strand Cary, M., & Fien, H. (2015). Investigating the efficacy of a core kindergarten mathematics curriculum to improve student mathematics learning outcomes. *Journal of Research on Educational Effectiveness*, *8*, 303–324. doi:10.1080/19345747.2014.980021

Clarke, B., Doabler, C. T., Nelson, N. J., & Shanley, C. (2015). Effective instructional strategies for kindergarten and first-grade students at risk in mathematics. *Intervention in School and Clinic*, *50*, 257–265. doi:10.1177/1053451214560888

Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D. B., Baker, S. K., Fien, H., & Smolkowski, K. (2014). Preliminary evaluation of a Tier-2 mathematics intervention for first grade students: Utilizing a theory of change to guide formative evaluation activities. *School Psychology Review*, *43*, 160–177.

Clarke, B., & 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. (2007). Curriculum research: Toward a framework for "research-based curricula." *Journal for Research in Mathematics Education*, *38*, 35–70.

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

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*, 66–69. doi:10.1126/science.1167311

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal*, *115*, 303–333. doi:10.1086/679969

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

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

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L.,… Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort* (Report to Congress No. NCEE 2007–4005). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Flay, B. R., Biglan, A., Boruch, R. F., González Castro, F., Gottfredson, D. C., Kellam, S.,… Jil, 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., Powell, S. R., Seethaler, P. M., Fuchs, D., Hamlett, C. L., Cirino, P. T., & Fletcher, J. M. (2010). A framework for remediating number combination deficits. *Exceptional Children*, *76*, 135–156.

Gause, M., Fien, H., Baker, S. K., & Clarke, B. (2011). *Project NumberShire I: A game-based integrated learning and assessment system to target whole number concepts*. (Subcontract with Thought Cycle, U.S. Department of Education, Institute of Education Sciences, Small Business Innovation Research Program, 2011–2013, Funding Number: EDIES11C0026, $1,049,954).

Gersten, R. M., Baker, S. K., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *Journal of Special Education*, *34*, 2–18.

Gersten, R. M., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools* (Report No. NCEE 2009–4060). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED504995

Gersten, R. M., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education*, *33*, 18–28.

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

Gersten, R. M., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*, 293–304. doi:10.1177/00222194050380040301

Hudson, P., & Miller, S. P. (2006). *Designing and implementing mathematics instruction for students with diverse learning needs*. Boston, MA: Pearson Education.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: Mathematics Learning Study Committee, National Research Council.

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward*. Cambridge, MA: The Education Arcade.
Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial & Special Education*, *24*, 97–114. doi:10.1177/07419325030240020501
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
Lohr, S., Schochet, P. Z., & Sanders, E. A. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis* (NCER No. 2014–2000). Retrieved from http://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf
Morgan, P. L., Farkas, G., & Maczuga, S. (2015). Which instructional practices most help first-grade students with and without mathematics difficulties? *Educational Evaluation and Policy Analysis*, *37*, 184–205. doi:10.3102/0162373714536608
Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*, 306–321. doi:10.1177/0022219408331037
Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*, 472–488. doi:10.1177/0022219411414010
Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.
National Association of State Directors of Special Education. (2006). *Response to intervention: Policy considerations and implementation*. Retrieved from http://www.nasdse.org/publications-t577/response-to-intervention-policy-considerations-an.aspx
National Center for Education Statistics. (2013). *A first look: 2013 mathematics and reading* (NCES 2014–451). Washington, DC: National Center for Education Statistics, Institute of Education Sciences.
National Center for Education Statistics. (2014). *The condition of education 2014* (NCES 2014–083). Retrieved from http://nces.ed.gov/pubs2014/2014083.pdf.
National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Retrieved from http://www.nctm.org/standards/focalpoints.aspx?id=282
National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematical Practice*. Retrieved from http://www.corestandards.org/the-standards/mathematics
National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.
Nelson, N. J., & Doabler, C. T. (2013). *TOT: Technology observation tool*. Eugene: University of Oregon.
Nelson, N. J., Fien, H., Doabler, C. T., & Clarke, B. (in press). Considerations for realizing the promise of educational technology. *TEACHING Exceptional Children*.
Nelson, N. J., Fien, H., Doabler, C. T., Clarke, B., Gause, M., & Baker, S. K. (2014, September). *Assessing the feasibility and promise of an individualized mathematics gaming intervention for first grade students in a randomized-controlled trial*. Presented at the Improving Early Math Outcomes for Students with Disabilities through Intensive Intervention Symposium, Fall Conference of the Society for Research on Educational Effectiveness, Washington, DC.
Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*, 138–154. doi:10.1080/19345740801982104
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
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
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Van de Walle, J. A. (2001). *Elementary and middle school mathematics: Teaching developmentally*. New York, NY: Longman.

What Works Clearinghouse. (2008). *What works clearinghouse: Procedures and standards handbook (Version 2.0)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A.,… Ogbuehi, P. (2012). *Improving mathematical problem solving in grades 4 through 8* (NCEE 2012–4055). Retrieved from http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch

Wray, K. A., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM© CCSS math measures: Grades K-8* (Technical Report No. 405). Retrieved from http://www.brtprojects.org/publications/dl/1111

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B.,… Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, *82*, 61–89. doi:10.3102/0034654312436980