NumberShire 1 Pilot Study Results Summary

Widespread concern has been expressed about the persistent low mathematics achievement of students in the US, particularly for students from low-income and minority backgrounds and students with disabilities. For example, results of the 2011 National Assessment for Educational Progress (NAEP) indicate that only 40% of 4th graders score at or above Proficient in math. Difficulties in mathematics achievement are particularly severe for students from low income and minority backgrounds and those with disabilities. For instance, nearly half of all 4th graders identified with a disability scored Below Basic on the 2011 NAEP. Mounting evidence also suggests that students who perform poorly in mathematics in the early grades are likely to continue to struggle throughout elementary school (Bodovski & Farkas, 2007; Morgan et al., 2009). Thus early intervention designed to support the needs of a range of learners is vital.

Instructional gaming technology, when designed and fictionalized well, has the potential to improve the motivation and mathematics achievement of students with or at-risk for mathematics difficulties (MD). Advanced gaming technology can provide a foundation to increase instructional intensity and serve as a motivational component for students who have experienced a long line of failure and frustration (Gersten et al., 2009b). Instructional gaming can, for example, facilitate the instructional interactions that deeply engage at-risk learners in the critical content of mathematics. Technology-based programs are also well suited to serve as targeted or intensive, supplemental interventions within a response-to-intervention framework, because of their capacity to differentiate instruction for a range of learners.

Despite these potential advantages, the research base is scant for efficacious technology tools in early mathematics (Dynarski et al., 2007). The National Education Technology Plan (NETP: Atkins et al., 2010) indicates that technology should be exploited to make learning experiences more meaningful, engaging, and accessible for students struggling to acquire academic proficiency. However, few of the numerous products available on the current market are grounded in research and development efforts that can fully address the agenda of the NETP and adequately meet the instructional needs of students at risk for MD. For example, in the area of early math instruction, the What Works Clearinghouse (WWC) has reviewed a total of 75 elementary programs to date, 22 of which are technology programs. Of these 22 programs, only two products at the elementary level have research studies that meet WWC screening criteria. In other words, less than 10% of the subset of technology programs reviewed have any research that could be used to evaluate their efficacy. Of the two reviewable programs, WWC ratings of impact on student outcomes are “potentially positive” and “mixed.”

Moreover, few existing technology products infuse in their design research-based instructional and technological design principles that support students struggling to learn academic content. Many technology-based mathematics programs lack explicit modeling for teaching new and complex concepts, and fail to provide guided practice opportunities to facilitate student learning (Doabler & Nelson, 2013). In addition, existing technology programs often fail to limit extraneous information, teach key vocabulary, and provide clear instructional examples. A need clearly exists for more efficacious technology tools specific to early mathematics instruction (Dynarski et al., 2007).

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

Purpose and Research Questions

Project NumberShire supports the development of in-depth knowledge of whole number concepts for students with or at risk for MD in grades K-2, a focus recommended by mathematics education experts (NMAP, 2009; Gersten et al., 2009a). NumberShire is a browser-based, educational video game in which players build an idyllic fairytale village by learning and applying whole number knowledge in three domains of the Common Core State Standards for Mathematics (CCSSO, 2010). The purpose of the NumberShire pilot study was to test the feasibility and promise of the NumberShire first grade interventions in authentic school settings. Specifically, we aimed to answer four research questions: (1) Is NumberShire feasible to implement in authentic school settings, delivered by school staff, using existing school equipment? (2) What are the statistical and practical effects of the NumberShire intervention on student proximal and distal outcomes?

Method

NumberShire Intervention

NumberShire is a fully featured, integrated learning and assessment system designed to support students with or at risk for MD develop proficiency with whole numbers in three domains represented in the K-2 Common Core Standards for Mathematics (CCSS-M: CCSSO, 2010): (a) Counting and Cardinality, (b) Number and Operations in Base Ten, and (c) Operations and Algebraic Thinking. Each version of NumberShire (i.e., NumberShire K, NumberShire 1, and Number Shire 2) consists of 6-12 hours of individualized instructional game play, comprised of 15-minute sessions, delivered 4 times per week. In addition, student mastery data reports are generated that depict student game performance in independent practice activities according to the CCSS-M, and are available to teachers for use in data based decision-making.

In the game, players assume the role of a young member of a Renaissance-style village in the fairytale-inspired medieval kingdom of Tally-ho, where the village elder is stepping down and handing over the mantle of leadership to the player. Sim-style game mechanics allow the player to click on village buildings to trigger mini-games targeted at whole number concepts, such as composing and decomposing teen numbers, and word problem solving.

Sessions utilize an explicit instructional format and contain three instructional phases: explicit modeling, supported practice, and independent practice. Embedded within each session are four mathematics mini-games, including a Teaching Event (i.e., a mini-lesson targeting a new instructional objective), Assessment Event (i.e., review of a previously mastered objective), Warm-up, and Wrap-up. Mini-games include clear explanations to introduce new material and high quality feedback, and a differentiated learning pathway is used to direct students to additional instruction and practice activities when game performance indicates a need for support. A variety of virtual mathematical representations (e.g., number lines, base-10 blocks) and frequent practice opportunities are employed to facilitate procedural fluency and build a robust and enduring conceptual understanding of whole number concepts.

Participants and Procedures

In fall 2013, schools in two school districts in different regions in Oregon were invited to participate in a randomized-controlled pilot study. All of the first grade classrooms (n= 26) in nine schools expressed interest in participating in the study. Eleven of the participating classrooms were set in five Title I schools, located in a suburban school district (District A) in Eugene, Oregon. District-wide, 57% of students received free or reduced price lunch, 19.6% received special education services, 3.1% were considered English Learners, and 24.1%

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

identified as ethnic minorities. The remaining 15 classrooms were set in four schools in a suburban school district (District B) in the Portland metropolitan area. In District B, 35.5% of students received free or reduced lunch, 13.2% received special education services, 14.8% were considered English Learners, and 39.3% identified as ethnic minorities.

After obtaining teacher and parent consent, all assenting students in each first grade classroom completed a math screening assessment (EasyCBM-CCSS, Fall benchmark). The ten students with the lowest scores on the screening assessment in each classroom were identified as NumberShire-eligible. Subsequently, NumberShire-eligible students were assessed using EasyCBM-NCTM (Fall benchmark), Early Numeracy CBM (EN-CMB), and a comprehensive proximal NumberShire Common Core State Standard assessment (PN) at pre-test. After students completed assessment, students were randomly assigned within classroom to the NumberShire treatment or a wait-list control condition. One classroom was unable to comply with random assignment and was dropped from the study, resulting in 125 students in each condition.

School-employed educational assistants were trained and paid by the project to facilitate NumberShire intervention groups, comprised of NumberShire treatment students from each school building. Schools determined intervention group size (range = 5-25 students per group; median = 10 students) based on availability of hired staff. Students participated in the NumberShire intervention 4-5 days per week for eight weeks (i.e., the pilot study tested the first eight weeks of the NumberShire first grade intervention). Every two weeks, NumberShire-eligible students were administered a mastery test based on game content taught during the previous two weeks. At the end of the eight-week study (post-test), students were assessed using EasyCBM-NCTM (Winter benchmark), EN-CBM, and the comprehensive proximal NumberShire assessment.

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 also evaluated using ANCOVAs adjusting for pretest proximal NumberShire (PN) total score as a covariate. 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. Non-nested analyses were appropriate for this study given that students were both the unit of randomization and the unit of analysis. All analyses were conducted with SPSS 21 and alpha was set to $p < .05$, two-tailed, for all tests.

Hedges' *g* was reported as a metric of intervention effect size (What Works Clearinghouse, 2008; .2, .5, and .8 are considered small, medium, and large effects). Hedges' *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.

Results

Baseline Equivalency and Attrition

The expectation of baseline equivalency 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 1 for demographic descriptive information). Compared to control students, treatment students performed significantly better on pretest en-CBM quantity discrimination ($M = 21.1$, $SD = 7.5$ vs. $M = 19.0$, $SD = 7.6$; $t[235] = 2.18$, $p =$

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

.030) and the proximal NumberShire assessment (M = 39.2, SD = 17.0 vs. M = 34.6, SD = 16.7; t[246] = 2.15, p = .032). To control for baseline non-equivalencies, these measures were included as additional covariates in all outcome analyses.

The extent to which attrition threatened the internal and external validity of the 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 8 (6.4%) of the treatment participants did not complete a posttest assessment compared to 4 (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 proximal NumberShire assessment (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 2 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 primary proximal NumberShire assessment (p < .001, partial η^2 = .063, Hedges' g = 0.30) and the 2-week interim proximal NumberShire assessment (p = .025, partial η^2 = .022, Hedges' g = 0.22).

Program Utilization, Implementation Fidelity, and Quality of School Technology

Between pretest and posttest, treatment students, on average, completed 18.6 game sessions (SD = 8.1, range = 2 to 33) and repeated 12.8 game sessions (SD = 5.9, range = 2 to 24). A correlation analysis was used to explore associations between number of sessions completed and change in outcomes from pretest to posttest among students assigned to the treatment condition. There were no statistically significant correlations between change in outcomes from pretest to posttest and the number of sessions completed (r's ranged from -.10 to .07, p > .200 for all tests), or the number of repeated sessions (r's ranged from -.13 to .09, p > .170 for all tests).

Project staff observed and rated each of the nine participating schools on the quality of implementation and school technology (see Table 3). Implementation quality was measured by items that assessed instructor procedures at the start of gameplay, the extent to which students wore headphones and were engaged during gameplay, instructor monitoring and classroom management, instructor troubleshooting abilities, and delivery of tangible rewards at the end of the session. Items were rated on a four-point scale (1 = Not Present, 4 = Highly Present) and were averaged to compute an overall implementation quality score. The average overall quality score was 3.1 (SD = 0.7, range = 1.7 to 3.8), indicating moderate overall quality with substantial variability between schools.

Project staff also rated each school on the number of computers that had to be reset during the observed session (M = 1.2, SD = 1.6), the number of computers that experienced long wait or loading times (M = 1.4, SD = 1.7), and the average duration of session log-in process (M = 5.8 minutes, SD = 6.3). Instructor impressions of the typicality of the observed session was also rated on a four-point scale (1 = Not At All Typical, 4 = Highly Typical) and, on average, was 3.1 (SD = 1.3), indicating that the observed session was similar to the typical session.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

Table 1
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)

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

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

Table 2

Descriptive Statistics and ANCOVA Results for the Outcome Measures

Outcome Measure/ Condition	Pretest	Posttest		Condition Effect			
	M (*SD*)	*M* (*SD*)	*Adj M*	*F*	*p*	Partial Eta^2	Hedges' *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				
easyCBM-NCTM total percentile				1.59	.208	.008	-0.15
Treatment	29.2 (25.0)	20.4 (20.0)	18.2				
Control	26.0 (19.8)	19.9 (20.0)	21.2				
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				
Primary proximal NumberShire				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				

Note. *M* = Mean, *SD* = Standard Deviation, *Adj* = Adjusted. Baseline differences between conditions were observed on the en-CBM quantity discrimination and the Proximal NumberShire 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 NumberShire pretest score as a covariate, as the interim measure was not assessed at pretest. na = not assessed at pretest.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.

Table 3

Descriptive Statistics for School-level Implementation Quality and Quality of School Technology

Measure	*M* (*SD*)	Min	Max
Implementation quality[1]			
Instructor used effective procedures at the start of game play	3.5 (1.1)	1.0	4.0
Students wore headphones during game play	3.7 (0.7)	2.0	4.0
Students were engaged in the game session	3.2 (0.8)	2.0	4.0
Instructor actively monitored students during game play	3.6 (0.7)	2.0	4.0
Instructor was able to troubleshoot technology issues during session	2.6 (1.1)	1.0	4.0
Instructor used a procedure at the conclusion of the session to manage students that finished sessions early	3.2 (1.2)	1.0	4.0
Instructor provided tangible rewards at the end of the session	1.6 (1.9)	0.0	4.0
Implementation quality summary score	3.1 (0.7)	1.7	3.8
Quality of school technology			
Number of computers that had to be reset during the session	1.2 (1.6)	0.0	5.0
Number of computers that experienced long wait or loading times	1.4 (1.7)	0.0	5.0
Duration of session log-in process	5.8 (6.3)	1.0	20.0
Network quality[2]	2.4 (.5)	2.0	3.0
Computer quality[2]	1.9 (1.2)	1.0	4.0
Overall quality[2]	4.4 (1.7)	3.0	7.0

Note. *M* = Mean, *SD* = Standard Deviation, Min = minimum, Max = maximum. [1]Implementation quality items were rated on a four-point scale (1 = not present, 4 = highly present) and were averaged to compute the implementation quality summary score. [2]Network quality, computer quality, and overall quality of technology items were rated on a four-point scale (1 = poor, 4 = optimal).

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120071 to the University of Oregon and SBIR funds to Thought Cycle, LLC. The opinions expressed are the authors and do not represent views of IES or USDE.