Journal of Numerical Cognition

PsychOpen GOLD

Empirical Research

# Effectiveness of a Numeracy Intelligent Tutoring System in Kindergarten: A Conceptual Replication

Ka Rene Grimes[1], Soyoung Park[2], Amanda McClelland[3], Jiyeon Park[4], Young Ri Lee[5], Maryam Nozari[6], Zainab Umer[3], Brenda Zaparolli[3], Diane Bryant[3]

**[1]** *Department of Curriculum and Instruction, Tennessee Technological University, Cookeville, TN, USA.* **[2]** *Department of Special Education, Western Kentucky University, Bowling Green, KY, USA.* **[3]** *Department of Special Education, The University of Texas at Austin, Austin, TX, USA.* **[4]** *Department of Curriculum and Instruction, Eastern Kentucky University, Richmond, KY, USA.* **[5]** *Department of Educational Psychology, The University of Texas at Austin, Austin, TX, USA.* **[6]** *Department of Special Education, The University of Hawai'i at Mānoa, Honolulu, HI, USA.*

Journal of Numerical Cognition, 2021, Vol. 7(3), 388–410, https://doi.org/10.5964/jnc.6931

**Received:** 2020-08-03 • **Accepted:** 2021-03-02 • **Published (VoR):** 2021-11-30

**Handling Editors:** Mojtaba Soltanlou, University of Surrey, Guildford, UK; Krzysztof Cipora, Loughborough University, Loughborough, UK

**Corresponding Author:** Ka Rene Grimes, Tennessee Technological University, Department of Curriculum and Instruction, Box 5042, Cookeville, TN, USA 38505; 1.011.931.372.3100. E-mail: kgrimes@tntech.edu

**Related:** This article is part of the JNC Special Issue "Direct and Conceptual Replication in Numerical Cognition", Guest Editors: Mojtaba Soltanlou & Krzysztof Cipora, Journal of Numerical Cognition, 7(3), https://doi.org/10.5964/jnc.v7i3

**Supplementary Materials:** Data, Materials [see Index of Supplementary Materials]

## Abstract

Intelligent Tutoring Systems are a genre of highly adaptive software providing individualized instruction. The current study was a conceptual replication of a previous randomized control trial that incorporated the intelligent tutoring system Native Numbers, a program designed for early numeracy instruction. As a conceptual replication, we kept the method of instruction, the demographics, the number of kindergarten classrooms (n = 3), and the same numeracy and intrinsic motivation screeners as the original study. We changed the time of year of instruction, changed the control group to a wait-control group, added a maintenance assessment for the first group of participants, and included a mathematical language assessment. Analysis of within- and between-group differences using repeated measures ANOVA indicated gains of numeracy were significant only after using Native Numbers (Partial Eta Square = 0.147). Results of intrinsic motivation and mathematical language were not significant. The effect size of numeracy achievement did not reach that of the original study (Partial Eta Square = 0.622). Here, we compared the two studies, discussed plausible reasons for differences in the magnitude of effect sizes, and provided suggestions for future research.

## Keywords

replication, numeracy, kindergarten, intelligent tutoring system, computer-assisted instruction, tablets, wait-control design

Results from syntheses and meta-analyses of research on difficulties with mathematics indicate that many students who struggle with mathematics at the end of kindergarten continue to struggle throughout their schooling (e.g., Morgan et al., 2016; Nelson & Powell, 2018). Students who have persistent challenges with mathematics in early grades may have or be at risk for mathematics difficulty (MD); and, at some point, be diagnosed as having mathematics learning disability (MLD; Chinn et al., 2017; DSM-5, American Psychiatric Association, 2013). Difficulties associated with identifying MLD, however, are complicated by the current limited understanding of the neural and behavioral mechanisms of numerical cognition, the wide variety of assessments used, and differences in cut scores used across schools and researchers

(Alcock et al., 2016; Lewis & Fisher, 2016; Jordan et al., 2019; Watson & Gable, 2013). Given the long-term negative impact of low numeracy prior to entering the primary grades (Geary, 2015), prevention research in preschool and kindergarten is both logical and recommended to prevent the risk of MD and to help build the emerging body of research identifying characteristics of MD and MLD (Morgan et al., 2016; Nelson & Powell, 2018; Nguyen et al., 2016; Penner et al., 2019).

In the United States, the National Association of Education of Young Children (NAEYC, 2012) and the National Council of Teachers of Mathematics (NCTM, 2010) recommended the use of technology as a tool to help develop conceptual understanding of mathematics for young students; and researchers support the use of technology as one means of differentiated instruction (Deunk et al., 2018; Twyman & Sota, 2016). Reported effectiveness of technology to increase numerical skills, however, has been mixed and may be due to methodological issues, the pedagogy and content of the software, or the software's architecture (Cheung & Slavin, 2013; Harskamp, 2014; Moyer-Packenham et al., 2019; Ok et al., 2020; Twyman & Sota, 2016). Thus, evaluating the effectiveness of mathematics instructional software also requires identifying the software's architecture to help understand potential mechanisms driving academic outcomes (e.g., Larkin & Milford, 2018; Moyer-Packenham et al., 2019; Ok et al., 2020).

## Intelligent Tutoring Systems

The architectural designs of educational software include, for example, models wherein a user works through a series of items in a linear fashion, models that allow users to adjust instruction based on preference, and models that adapt instruction (VanLehn, 2011). The degree to which a program adapts instruction also varies, prompting discussion among experts in the artificial intelligence for education community regarding the need for operational definitions of the term *adaptive* and for standards and the criteria necessary for software to qualify as being labeled *adaptive* (e.g., Aleven, 2015; R. S. Baker, 2016; Ma et al., 2014; Rozo & Real, 2019; Wilson & Scott, 2017; also see https://standards.ieee.org/project/2247_1.html for proposed standards).

An intelligent tutoring system (ITS) is a distinctly different type of adaptive instructional software designed with complex models that embed individualized instruction in tandem with dynamic assessment (Ma et al., 2014; Pirolli, 2014). The term *tutoring*, in an ITS, describes an active, adaptive, student-centered, individualized mode of delivery (Bourdeau & Grandbastien, 2010). Ma et al. (2014) explained that an ITS:

> "computes inferences from student responses, constructs either a persistent multidimensional model of the student's psychological states (such as subject matter knowledge, learning strategies, motivations, or emotions) or locates the student's current psychological state in a multidimensional domain model..." and "uses the student modeling functions...to adapt one or more of the tutoring functions" (p. 902).

Critically, the student model is the software design element that differentiates an ITS from other educational instructional software (Ma et al., 2014; Pavlik et al., 2013). The student model, also referred to as a learner model (Pelánek, 2017), describes the psychological state (i.e., cognitive state) of the user (Ma et al., 2014).

One potential advantage of utilizing an ITS for instruction is that elements required within the software's architecture inherently include practices some researchers have recognized as evidence-based: explicit instruction, adapting instruction to individual needs based on assessment, and providing feedback (e.g., Barnes et al., 2016; Clarke et al., 2015; Dennis et al., 2016; Mononen et al., 2014; Nelson & McMaster, 2019; Wang et al., 2016). That is not to say that other instructional software does not include evidence-based practices, nor suggest that inclusion or exclusion of evidence-based practices alone can predict academic outcomes (e.g., Kiru et al., 2018). Nonetheless, meta-analyses and reviews of ITSs revealed that some ITSs were more effective than small-group instruction and almost as effective as one-to-one tutoring (Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper; 2013; VanLehn, 2011).

## Research Including an Early Numeracy Intelligent Tutoring System

Despite the potential benefits of using an ITS for mathematics instruction, none of the studies included in the published meta-analyses and syntheses of ITS (e.g., Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper; 2013;

VanLehn, 2011) included the use of an ITS designed for early numeracy instruction. In a systematic review of adaptive early numeracy software in grades pre-school through first grade, (authors, in preparation) identified one unpublished dissertation (Dias, 2016) in which the author included an early numeracy ITS: Native Numbers (www.nativebrain.com, Native Brain, 2014), an ITS designed for delivery via iPads.

The location of the Dias (2016) study was in a private school in the United States. Dias randomly assigned participants ($n$ = 57) within three kindergarten classrooms into either the treatment group or the active control group. During regularly scheduled center activity time, both groups worked on supplemental numeracy activities; the treatment group used Native Numbers (Native Brain, 2014) and the active control group worked on paper-and-pencil numeracy enrichment problems with content aligned to Native Numbers' content. The students' classroom teachers monitored all students' activity and provided encouragement or responded to questions regarding technology but did not provide instruction. The student-to-teacher ratio was 1:19. At the individual level, once a student had completed Native Numbers or the enrichment activities, they received post-tests, and moved into business-as-usual mathematics center activities. Dias (2016) reported a substantially large effect size on differences in mathematics outcome scores for participants in the treatment groups compared to the active control groups: $\eta_p^2 = 0.622$.

Based on the need for early numeracy instruction to prevent MD (Geary, 2015; Morgan et al., 2016; Nguyen et al., 2016; Penner et al., 2019), the effectiveness of Native Numbers (Native Brain, 2014) reported by Dias (2016), the call for replication studies (e.g., the *Standards of Evidence for Efficacy, Effectiveness and Scale-up Research in Prevention Science: Next Generation*, Gottfredson et al., 2015), and the lack of studies incorporating ITSs for early numeracy instruction (authors, in preparation), replication of the Dias (2016) study was warranted. The structure of the remainder of the paper includes the purpose of the current conceptual replication study, descriptions and comparisons of the methods and outcomes of the original study and the current study, and discussion of how the results of both studies provide an opportunity to design future research to examine aspects of numerical cognition that are either emerging lines of research or are missing in the literature (e.g., ordinality, mathematical language, multiple representations, number rods).

## Purpose of the Study

The purpose of the current study was to conduct a conceptual replication of the Dias (2016) study. Conceptual replications differ from direct replications in that specific features from the original study are intentionally manipulated to investigate theoretically important variables while maintaining critical elements of the original study (Cai et al., 2018; Coyne et al., 2016). For the current study, we maintained the participant demographics as closely as possible, the treatment instrument (i.e., Native Numbers; Native Brain, 2014), and the two outcome measures. However, we identified four theoretically relevant elements to manipulate:

a. implementing the treatment in the spring
b. providing the same treatment to a wait-control group
c. assessing maintenance of academic gains over time
d. assessing mathematical language

Dias (2016) conducted the study in the fall, at the beginning of the school year. We reasoned staring a study later in the year allowed exploring the variable of maturation and opportunities for students to experience formal mathematics instruction after entering kindergarten. Changing the control group to a wait-control group allowed measuring maintenance of any academic gains the first-treatment group may have had, comparing outcomes of two groups of participants receiving treatment at different points in time, and providing a control for the novelty effect from the use of technology.

The fourth change to the Dias (2016) study was the inclusion of a mathematical language assessment. Beyond mapping number words to symbolic or non-symbolic notations, mathematical language includes words that describe relationships and space; for example, terms such as "greater" and "below" (Hornburg et al., 2018). Mathematical language is critical for understanding concepts of cardinality, ordinality, sets, and magnitude (Hornburg et al., 2018). Despite the emerging evidence of the importance of mathematical language, to date, few experimental studies have included and measured explicit instruction of mathematical language in early childhood; notable exceptions include studies utilizing storybooks in prekindergarten (Hojnoski et al., 2014; Purpura et al., 2017) and kindergarten (Hassinger-Das et al., 2015; Jennings et al., 1992), and a study using manipulatives, hand gestures, and verbal instruction of mathematical terms in

first grade (Powell & Driver, 2015). Native Numbers (Native Brain, 2014) provides explicit instruction of mathematical relational terms, thus we added a measurement of mathematical language to the conceptual replication.

With the changes to the Dias (2016) study described above, we sought to answer the following research questions:

1. How do gains in students' numeracy compare to the original study?
2. How do the outcomes of intrinsic motivation compare to the original study?
3. Did students' outcomes on a mathematics language screener change as a result of using Native Numbers (Native Brain, 2014)?

# Method

## Participants and Setting

In the United States, the age of kindergarten attendance is around the age of 5. However, kindergarten is not mandatory across all states. As previously described, the Dias (2016) study participants included students from three kindergarten classrooms attending a private school in the northeast of the United States. The sample included 30 girls and 26 boys: 2% African American, 26% Asian, 56% Caucasian, and 16% multi-ethnic. All participants who started the study remained in the study (i.e., zero attrition).

Participants in the current study ($n$ = 46) also attended a private school, but the location was in the United States' southwest region. The teacher-to-student ratio was 1:9, except for one group which had only 17 students. The total number of kindergarten students enrolled at the site was 53 students. However, only 24 girls and 22 boys returned consent forms and provided assent. Ethnicity, as reported by caregivers, indicated the following percentages: 4% Asian; 69% Caucasian; 9% Hispanic; 11% Multi-Ethnicity; 7% preferred not to report. The attrition was also zero.

Although treatment for both studies occurred during regular center-based activity time, prevalent in kindergarten classrooms in the United States, the grouping of students into classrooms was distinctly different in the current study. Specifically, rather than having three separate kindergarten *classrooms*, with individual teachers assigned to each class, the current study site was a large complex without separate classrooms, per se. The building had several smaller rooms used for small group instruction, and each of the smaller rooms opened up to larger common use spaces. Students received core instruction within one of three *groups*, as opposed to classrooms, and each group had two teachers serving as the primary teachers for that group. Students interacted with any of the six teachers in fluid grouping formats throughout the building throughout the day. The teacher-to-student ratio in the current study was approximately 1:9 as opposed to 1:19 in the Dias (2016) study.

Based on the dynamics of the center-based instruction time, the building's design, and the classroom teachers' preference, treatment in the current study occurred in one of the larger multi-purpose rooms, rather than in individual classrooms as in the Dias (2016) study. Thus, in the current study, the physical location and monitoring use of Native Numbers (Native Brain, 2014) was similar to what might occur in a school computer lab.

## Measures

### Intrinsic Motivation

Similar to Dias (2016), we assessed intrinsic motivation via the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990). The inventory was developed for use with students in 1st through 3rd grade as a downward progression of the Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990) and consists of three sub-measures: math, reading, and general overall school experience. Each sub-measure contains 13 statements to which participants are asked to respond whether a statement is very true, a little true, or not true for them. The statements are the same for all sub-measures, only the domain changes for each measure. For example, the statement, "I like to do easy reading work," is changed to "I like to do easy math work." Individual items are assigned a score from 1 to 3 related to the motivation attributed to that item; higher scores represent higher motivation. Scores can include the individual sub-measures or a combination of the sub-measures. The inventory is not available commercially; we received permission for use within this study (Gottfried, personal communication).

Gottfried (1990) reported reliability and validity based on a cross-sectional study and a longitudinal study. Both studies included 7- to 9-year-old participants; although, the ages and number of participants varied for each analysis. Combining the data from both studies, Gottfried reported strong internal consistency for reading motivation ($\alpha$ = 0.82), math motivation ($\alpha$ = 0.84), and motivation for overall school experience ($\alpha$ = 0.82). For the cross-sectional study, two-month test-retest correlations ($n$ = 57) were $r$ = .73 for the reading subscale and $r$ = .73 for the math subscale ($p$ < .001). Correlations between the total score of all scales and researcher developed questions of students' perception of competence were $r(96)$ = .37 and $r(105)$ = .35, $p$ < .001, for the 7- and 8-year-old participants in the cross-sectional and the longitudinal studies respectively. Correlations between teacher reports of student motivation across the subscales ranged from $r$ = .18, $p$ < .05, to $r$ = .25, < .001; correlations by subscale were not reported. Correlations between the Young Children's Academic Intrinsic Motivation Inventory and a modified version of the Children's Academic Anxiety Inventory (Gottfried, 1982, 1985) were $r(96)$ = −.25, $p$ < .01, for participants in the cross-sectional study and $r(103)$ = −.20, $p$ < .01, for 7-year-old participants in the longitudinal study (see Gottfried, 1990, p. 535).

### Numeracy

Similar to Dias (2016), we assessed numeracy via the Number Sense Screener™ (Jordan et al., 2012). The screener includes 29 items, is standardized, and takes approximately 15 minutes to administer one-on-one. Each item answered correctly receives one point. The screener assesses six numerical concepts: counting, number recognition, number comparisons, non-verbal calculation, story problems (oral addition and subtraction), and single-digit number combinations. Additionally, the screener is normed for kindergarten students in the fall and spring and for first-grade students in the fall. Jordan et al. analyzed construct and predictive validity using the Woodcock-Johnson III calculation and applied problem scales (WJ, Woodcock et al., 2007). The correlation between the screener and the WJ composite scores of the subscales produced an average $r$ = .62 for 1st grade and $r$ = .65 for 3rd grade ($N$ = 288). The fall of 1st grade WJ composite scores correlated ($r$ = .72) with the spring of 1st grade WJ composite scores ($n$ = 279) and with the spring of 3rd grade WJ composite scores ($r$ = .70; $n$ = 175). Jordan et al. reported an average internal consistency of $\alpha$ = 0.85.

### Mathematical Language

To our knowledge, a validated stand-alone kindergarten assessment of mathematical language in English does not exist. Therefore, we chose the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015), a researcher-developed measure currently undergoing psychometric evaluation to assess mathematical vocabulary. We obtained permission (Purpura; personal communication) to use the current version of the assessment that includes 16 items drawn from a larger battery (e.g., Purpura & Logan, 2015). The 16 items include six quantitative terms (e.g., more, less) and ten spatial terms (e.g., under, last). Students respond to pictures with oral prompts such as "Point to the box with more dots," where *more* is the target vocabulary term. The assessment takes approximately 10 minutes to administer one-on-one and students receive one point for each correct item. Convergent validity has not been reported; however, this measure has shown strong internal consistency across several studies: $\alpha$ = 0.85 in Purpura and Logan (2015), $\alpha$ = 0.78 in Purpura et al. (2017), $\alpha$ = 0.78 in Hornburg et al. (2018), and $\alpha$ = 0.80 in Purpura et al. (2020).

## Numeracy Instruction

### All Participants: Regular Numeracy Instruction

In both the Dias (2016) study and the current study, all participants received regular classroom mathematics instruction, and the treatment groups and control groups received additional supplemental mathematics instruction. Thus, the control groups for both studies were active-control groups. In the current study, all participants received regular daily mathematics instruction using Bridges Number Corner (see https://www.mathlearningcenter.org/number-corner). Dias did not indicate the specific curriculum used at the school site.

### Treatment Groups: Native Numbers

For both the Dias (2016) study and the current study, participants in the treatment conditions received instruction via Native Numbers (Native Brain, 2014), an application (i.e., app) designed for use on the iPad, available for free, and

downloaded via iTunes. The adaptive algorithm in Native Numbers includes measuring accuracy and, at higher levels of difficulty, response time as a measure of fluency and as an indicator within the architecture to further adapt instruction. Based on the description provided by the parent company of Native Numbers (www.nativebrain.com), the program adapts instruction in real-time, along five levels of academic performance, moving users to more difficult tasks, or returning to earlier concepts when a user experiences difficulty. Once a user has reached the third level, the next set of activities is unlocked. Once unlocked, users can continue working on accuracy and fluency (Levels 4 and 5) on their current activity, move on to the next activity, or move back and forth between any of the unlocked activities.

Native Numbers (Native Brain, 2014) provides instruction for five areas of early numeracy for quantities 1–9: number concepts (mapping the verbal quantity to a visual representation); number relations (instruction of mathematical vocabulary such as *more* and *less*); ordinality; counting (one-to-one, counting on, and counting back); and a set of activities named "Demonstrate Mastery" which requires consolidation of all previous learning. Each of these five areas is further sub-divided into five sets of activities for a total of 25 sets of activities (see Appendix A in the Supplementary Materials).

Native Numbers (Native Brain, 2014) incorporates symbolic representation (i.e., Arabic digits) and two non-symbolic representations of discrete magnitude: (a) sets presented in arrays similar to playing cards and (b) tally marks. Additionally, Native Numbers includes one non-symbolic representation of continuous magnitude: number rods similar to Cuisenaire Rods®. Users learn to map between quantities for each of the different representations through blocked practice and interleaved practice. Also, as part of the Demonstrate Mastery activities, some activities are designed with reversibility (e.g., Simon et al., 2016); for example, given a target quantity and a starting amount, a user must add or remove a quantity to reach the target quantity. Native Numbers, however, does not include symbolic operational signs.

### Control Groups: Numeracy Enrichment Activities

Participants in the Dias (2016) study received additional numeracy instruction through paper and pencil activities aligned to Native Numbers' (Native Brain, 2014) content. Participants in the current study had daily center-based supplemental instruction for approximately 45 minutes in the afternoon. Students worked in small groups with teachers, one-on-one with teachers, or in peer-to-peer activities during this time. The numeracy enrichment activities during this center-based time varied throughout the study, based on the teachers' goals and individual students' needs, and included other instruction such as reading, writing, science, and art (see Appendix B in the Supplementary Materials for a list of the different numeracy activities used during the timeframe of the study).

## Procedures

The current study was approved by the University of Texas at Austin Institutional Review Board (IRB). Teachers and parents/guardians provided written consent, and students provided verbal and written assent. Prior to beginning the assessments, we used an online random number generator to randomize the participants by their primary groups into the first-treatment group or the wait-control group. For ethical and pragmatic reasons, all students had the opportunity to use Native Numbers (Native Brain, 2014); however, analyses included only data from the participants who returned consent forms and provided assent ($n$ = 46). The teachers and parents/guardians received outcome scores from all the assessments, and the teachers received typed narrative reports on student progress.

### Baseline and Pre-Tests

In the Dias (2016) study, additional school staff helped facilitate the assessments. The teachers at the site of the current study requested that the researchers conduct all the assessments. Pre-testing in the current study occurred in either the large multipurpose rooms or smaller areas throughout the building. Unlike Dias, we included an assessment of mathematical language and combined this assessment with the Number Sense Screener™ (Jordan et al., 2012) into one sitting, administered one-on-one. After completing all the assessments, participants began instruction in their separate groups; however, due to the large number of participants, we phased in the first-treatment group over three days.

Assessment of intrinsic motivation via the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990) differed significantly in the current study from the Dias (2016) study. Specifically, during a pilot administering

the inventory to nine kindergarten students, we found that students had difficulty understanding the directions for responding to negative statements. For example, we included three practice statements as pre-training to ensure participants understood the format and how to respond. For negative statements such as "I do not like ice-cream," participants responded "very true" when in fact, the students meant the opposite response; they did like ice-cream. This confusion occurred across the negative response statements of the practice statements and the statements in the inventory, despite explicit modeling and feedback with the practice statements. Additionally, facilitating the pilot of the inventory took more than 30 minutes for only two of the three sub-measures.

Concerned about participant testing fatigue and reliability of responses, we removed the overall schoolwork sub-measure and we removed two questions each from the reading and mathematics sub-measures (the same questions for both content areas): one indicating negative intrinsic motivation and one indicating positive intrinsic motivation. Also, due to time constraints and researcher availability, we chose to implement the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990) in small groups of three, rather than one-on-one. Additionally, as our question of interest was whether the use of Native Numbers (Native Brain, 2014) increased intrinsic motivation, we changed the timing of the pre-test for each group such that we administered the inventory immediately before using Native Numbers (i.e., the pre-test assessment for the wait-control group was later than the first-treatment group, occurring just prior to using Native Numbers).

### Treatment

Dias (2016) reported participants in both groups (i.e., treatment and control) worked for up to 30 minutes and were allowed to quit at any time; the range of hours participants used Native Numbers (Native Brain, 2014) was from 5–12, for one to three days per week. Like the Dias (2016) study, in the current study, the number of days per week participants used Native Numbers varied and averaged to three days per week. Participants in the current study were also permitted to quit working at any time, although we encouraged participants to work for at least 10 minutes and announced when 10 minutes had elapsed. Each day we were at the site, the classroom teachers provided us names in advance for the treatment participants that they needed to work with for non-mathematics instruction (e.g., writing, reading, or group projects requiring intact groups of students). Due to the dynamic structure of this activity time, tracking the exact time participants entered or exited the room where Native Numbers' instruction occurred was difficult. Therefore, we recorded only whether a participant logged on to Native Numbers or not. The average time available for working was up to about 30 minutes, similar to that of Dias.

Also similar to Dias (2016), we designed the current study such that as each participant completed all 25 sets of activities in Native Numbers (Native Brain, 2014) to the 5$^{th}$ level, we administered the post-tests individually, after which the participant transitioned to a business-as-usual control group status. Although we designed the study with a goal of a clear line of separation between when the wait-control group would switch to using Native Numbers, we explicitly planned for, and included in the IRB proposal, the possibility that some students in the first-treatment group would not reach the 5$^{th}$ level on all 25 activities before the wait-control group entered into the treatment phase. We monitored the progress of the first treatment group and determined when to enter the wait-control group based on the number of days remaining in the academic year; that is, a point at which we could reasonably expect that all participants would have the same minimum number of days available in the treatment phase. As the study progressed, we set the minimum number of days at 22 (i.e., both groups had the opportunity to use Native Numbers for at least 22 days). The study's total elapsed time was 18 weeks, including holidays, field trips, professional development days, pre-testing, and post-testing.

One week before the minimum 22 days possible for the wait-control group to use Native Numbers (Native Brain, 2014) we assessed the wait-control group participants on motivation and facilitated the second assessment of the Number Sense Screener™ (Jordan et al., 2012) and the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015). After completing the three assessments, similar to the first treatment group's entry, participants in the wait-control group transitioned to the treatment phase over three days. As each participant in the wait-control group reached the 5$^{th}$ level on all 25 activities of Native Numbers, we administered the numeracy, mathematical language, and intrinsic motivation post-tests; then, these participants returned to business-as-usual center activities. Two weeks before the end of the school year, we began administering maintenance assessments of numeracy and mathematical language to the first treatment group participants who had completed Native Numbers a minimum of four weeks earlier. One

week before the end of the school year, we administered the numeracy, mathematical language, and intrinsic motivation post-tests to the wait-control participants who had not yet finished all 25 activities ($n$ = 6).

### Fidelity

All researchers received training on administering and coding the assessments prior to the start of the study. We randomly selected 33% of the assessments for each of the three testing time points and coded fidelity of the researchers' implementation of the assessments using checklists based on the measures' scripted instructions. Fidelity was 99.8% for the numeracy and mathematical language assessments, and fidelity for the intrinsic motivation was 99.3%, with the modifications described earlier. Additionally, we double coded all assessment data during grading and when coding into the Excel spreadsheet, ensuring a research team member who did not facilitate the assessment conducted the second coding. We verified participant usage of Native Numbers (Native Brain, 2014) by cross-checking the attendance records provided by the teachers and by data collected on the teacher dashboard associated with Native Numbers.

The iPads belonged to the school site. Each participant was assigned an iPad for use throughout the day. To minimize the risk that a student could log on to Native Numbers (Native Brain, 2014) while one of the researchers was not present, we took daily screenshots of the dashboards to verify participants had not logged on. To further minimize the risk that participants might use Native Numbers when it was not their assigned treatment phase, we installed Native Numbers only on the iPads of the participants currently assigned to use Native Numbers, and we removed Native Numbers from each iPad after a participant completed all 25 activities to the 5$^{th}$ level. According to the dashboards, none of the participants used Native Numbers outside of the time set for the study. Furthermore, results from a survey sent home with the consent forms indicated that none of the participants had used Native Numbers prior to starting the current study.

### Analyses

Analyses included raw scores for all measures and, with the exception of a robust test of imputed data described later, were conducted using SPSS for Mac, Version 26. Syntax for all analyses are provided in Appendices E–G of the Supplementary Materials. Additionally, before describing the primary analyses, in the following section, we described the methods used to examine the scores of 13 participants, seven from the first-treatment group and six from the wait-control group, whose use of Native Numbers (Native Brain, 2014) varied compared to the other participants.

**Preliminary Analysis: Difference in Days of Use** — Seven participants from the first-treatment group had not completed all 25 activities to the 5$^{th}$ level of Native Numbers (Native Brain, 2014) when it was time for the wait-control group to enter the treatment phase. Therefore, these participants continued to work, potentially introducing bias into the analysis. Importantly, the threat of bias was due to the difference in the number of days afforded these seven participants, not due to contamination. That is, the participants in the first-treatment group were not in both conditions at the same time.

Visual inspection of the dashboard data indicated that six participants in the wait-control group either did not complete all 25 activities ($n$ = 2) or did not reach the 5$^{th}$ level on one or more activities ($n$ = 4). An examination of attendance records indicated that for 11 of these 13 participants, the discrepancies in the number of days using Native Numbers (Native Brain, 2014) were due to absences, participating in reading or writing assessments during the period offered to use Native Numbers, or work involving projects requiring intact groups (e.g., end-of-the-year presentations).

To assess the potential impact the additional elapsed time the seven first treatment group participants had to use Native Numbers (Native Brain, 2014) compared to the six wait-control group participants, we first inspected each participants' total number of days of use (see Table 1). Overall, the data suggested that the number of days of use was not a driving factor of changes from the pre-test to the post-test of numeracy. For example, only two of the seven participants in the first treatment group worked more than the 22 days afforded to both groups. Notably, one participant in the first-treatment group who worked 28 days, and two wait-control group participants who worked only 15 days, had the same pre-test scores and almost identical post-test scores; the two wait-control group participants' post-test scores were one point higher than the participant in the first-treatment group.

**Table 1**

*Days of Use and Outcome Scores for Participants Who Did Not Complete All 25 Activities to the Highest Level by Day 22*

| Group: Student | NSS[a] Pre-Test Score | NSS Post-Test Score | NSS Change | Days at 22[a] | Total Days | Days > 22[b] |
|---|---|---|---|---|---|---|
| First: 1 | 19 | 22 | 3 | 17 | 28 | 6 |
| First: 2 | 20 | 26 | 6 | 19 | 26 | 4 |
| First: 3 | 17 | 23 | 6 | 21 | 22 | 0 |
| First: 4 | 20 | 27 | 7 | 19 | 22 | 0 |
| First: 5 | 18 | 29 | 11 | 18 | 21 | 0 |
| First: 6 | 14 | 25 | 11 | 14 | 20 | 0 |
| First: 7 | 22 | 23 | 1 | 12 | 17 | 0 |
| Wait:1 | 25 | 27 | 2 | 17 | 17 | 0 |
| Wait: 2 | 18 | 22 | 4 | 16 | 16 | 0 |
| Wait: 3 | 25 | 26 | 1 | 15 | 15 | 0 |
| Wait: 4 | 19 | 23 | 4 | 15 | 15 | 0 |
| Wait: 5 | 19 | 23 | 4 | 15 | 15 | 0 |
| Wait: 6 | 22 | 27 | 5 | 11 | 11 | 0 |

*Note.* Group = first or wait-control; Student = participants not completing 25 activities in Native Numbers (Native Brain, 2014) to the 5th level by the 22nd day; NSS = The Number Sense Screener™ (Jordan et al., 2012). [a]Days worked up through the 22nd day available.
[b]Days worked above 22.

To further analyze the extent to which these 13 participants varied by outcome scores, we conducted an independent sample *t*-test. The results indicated the means of the numeracy outcome scores for the seven participants in the first treatment group ($M$ = 25.00, $SD$ = 2.52) compared to the six in the wait-control group ($M$ = 24.67, $SD$ = 2.25), were not significantly different, $t(1, 11) = 0.250$, $p = .81$, 95% CI [−2.605, 3.272]. Given the lack of significant difference in outcome scores, we did not consider that the 13 participants presented a risk of bias based on the number of days they used Native Numbers (Native Brain, 2014).

**Preliminary Analysis: Missing Data** — Although the seven participants in the first treatment group did not appear to introduce bias, they were missing the third assessment, the maintenance assessment, because they did not have a minimum of four weeks remaining in the year to examine retention of academic gains. We considered two options for analyzing repeated measures for within- and between-group analyses: repeated measures mixed ANOVA (RMM ANOVA) and linear mixed model (LMM) regression (Verma, 2015). LMM's are robust to unbalanced designs containing missing data (e.g., Muth et al., 2016), whereas RMM ANOVAs are sensitive to unbalanced groups and SPSS will list-wise delete all data from subjects missing even one data point. On the other hand, while statisticians have proposed models for reporting significant effects for LMMs, not all statisticians agree on the procedures (e.g., Meteyard & Davies, 2020; Rights & Sterba, 2020). In keeping with Meteyard and Davies (2020) recommendation for reporting why a particular analysis was chosen over another, we chose to replace the missing data and run the analyses with a RMM ANOVA because (a) the number of participants in our study was small, thus risking non-convergence for all the possible decisions of random effects in a LMM, and (b) the point of the analyses was not to fit a best model for our data, per se, but to compare the differences in the magnitude of effects between the current study and that of the Dias (2016) study.

We also considered two different methods of replacing the missing data of the seven participants in the first treatment group's maintenance scores: multiple imputation and the Last Observation Carried Forward (LOCF). Generally, statisticians highly discourage using the LOCF; however, for the current study, we considered the LOCF the most parsimonious and valid procedure (e.g., Overall et al., 2009). First, the missing data were the maintenance scores for the seven first treatment group participants who did not have a full month remaining post-treatment, not data from their pre-test or post-test scores. Second, we did not need the maintenance scores to compare differences in the effect of treatment between the first group and the wait-control group after each had used Native Numbers (Native Brain, 2014). Last, we did not need the maintenance scores to compare the magnitude of differences in the effect between the

treatment and control group in the Dias (2016) study to the effect between the first treatment group and the wait-control group in the current study prior to the wait-control group receiving treatment. We needed the missing data to (a) examine the first-treatment group's maintenance of gains after a minimum of one month and (b) to allow SPSS to run a RMM ANOVA without deleting the seven participants entirely.

Nonetheless, given the unfavorable view of statisticians to use the LOCF, as a robust test we ran five imputations and examined differences in the means and pooled means of the imputed data to the mean of the LOCF (see Appendix C in the Supplementary Materials for descriptive statistics). Although across the five imputations the means of the imputed data were slightly higher than the mean of the LOCF, the difference in pooled means were not significant: $t(40) = 1.634$, $p = .110$. The degree of freedom of the pooled result was corrected using an approximation suggested by van Ginkel and Kroonenberg (2014). Based on the results of the robust test and the justifications provide above, we retained the LOCF as the means for replacing the missing data on the third measure of the seven first treatment group participants when conducting the RMM ANOVA. Additionally, we removed these seven participants when examining the first treatment group's maintenance gains.

**Preliminary Analysis: Wait-Control Group Pre-Test Scores** — Due to the length of time from the first numeracy assessment to the second assessment, approximately six weeks, we used the second testing point as the wait-control group's pre-test score and considered the first assessment score as their baseline. Further justification for using the second assessment point as the wait-control group's pre-test was that results from a paired sample *t*-test indicated that the wait-control group had growth during the time between the first testing point and the second testing point, although the growth was not statistically significant: $t(1, 21) = -1.47$, $p = .157$, 95% CI [−0.738, 0.119]. Nonetheless, rather than taking an average of the wait-control group's first assessment and their second assessment, we considered their second test to be the most accurate reflection of their growth and used it as their pre-test.

**Primary Analysis** — We analyzed within- and between-group differences of the numeracy scores, across the three testing points, via a two-way RMM ANOVA. Two-way RMM ANOVAs are suitable for designs with two independent groups when looking at both between-group and within-group differences on a dependent variable measured repeatedly across both groups (Verma, 2015). We also conducted two separate *t*-tests to (a) examine the growth of the wait-control group after they used Native Numbers (Native Brain, 2014) and (b) to determine if the first-treatment group maintained any potential gains a minimum of one month after their post-test.

Initial descriptive statistics examining Q-Q Plots, Boxplots, and the Shapiro-Wilk test indicated the scores of the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015) were not evenly distributed. At the pre-test, approximately 80% of the participants scored at or close to ceiling. Therefore, we conducted separate non-parametric tests for within- and between-group differences using Mann-Whitney U Tests for between-group analysis (e.g., Nachar, 2008) and Wilcoxon Sign tests (e.g., Harris & Hardin, 2013) to analyze within-group differences.

Similar to the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015), the initial descriptive statistics of the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990) indicated a violation of assumptions of normal distribution for both reading and mathematics intrinsic motivation. Therefore, we conducted separate non-parametric tests for within- and between-group differences via a Mann-Whitney U Test for differences between groups and a Wilcoxon Sign test for differences within groups.

## Results of the Current Study Compared to Dias (2016)

We included the same numeracy outcome measure and the same intrinsic motivation measure as Dias (2016): The Number Sense Screener™ (Jordan et al., 2012) and the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990). Based on the significant modifications we made when implementing the intrinsic motivation inventory, we did not compare the results from the current study to those reported by Dias (2016). Furthermore, the results from both the mathematical language assessment and the intrinsic motivation inventory indicated insignificant differences between the first-treatment group and the wait-control group from pre-test to post-test scores; therefore, we presented these findings first.

## Intrinsic Motivation Results

We ran two Mann-Whitney U tests to determine if there were differences between the groups for intrinsic motivation for reading and intrinsic motivation for math. Our visual inspection indicated that although the distributions of scores for the two groups were not similar, differences between the groups' levels of intrinsic motivation for reading were not significantly different at the pre-test ($U$ = 291.50, $z$ = 0.607, $p$ = .544), nor the post-test ($U$ = 262.50, $z$ = −0.033, $p$ = .974). Similarly, between group differences for intrinsic motivation for math were not significantly different at the pre-test ($U$ = 231.50, $z$ = −7.20, $p$ = .471) nor the post-test ($U$ = 233.50, $z$ = −0.691, $p$ = .490).

We ran a series of Wilcoxon Signed Rank Tests to determine if there were significant changes from the pre-test to the post-test for reading motivation and for mathematics motivation within each group. For the first-treatment group, the median difference from the pre-test to the post-test for reading motivation was not significant ($z$ = 1.513, $p$ = .130); however, there was a statistically significant median increase in mathematics motivation from the pre-test to the post-test for mathematics motivation ($z$ = 1.990, $p$ = .047). For the wait-control group, the median differences from the pre-test to the post-test of reading motivation ($z$ = 0.526 $p$ = .599) and mathematics motivation ($z$ = 1.709, $p$ = .087) were not significant. Appendix D in the Supplementary Materials includes a table of the means, medians, and standard deviations of the intrinsic motivation analyses.

## Mathematical Language Results

Initial descriptive statistics of the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015) indicated that almost all the participants across both groups were at ceiling on the pre-test. Mann-Whitney U tests confirmed the median difference between the groups was not significant for either the first test, the second test, or the third test ($U$ = 186.00, $z$ = −1.834, $p$ = .067; $U$ = 281.50, $z$ = 0.437, $p$ = .662, $U$ = 183.50, $z$ = −0.119, $p$ = .922, respectively). Similarly, the Wilcoxon Signed Rank tests also indicated that within-group differences between the first testing period and the second testing period were not significant for the first-treatment group ($z$ = 0.000, $p$ = 1.00. However, changes from the first test to the second test were statistically significant for the wait-control group ($z$ = 2.289, $p$ = .022); that is, changes from their baseline to their pre-test were significant, although the median change was less than one point (e.g., 0.59). Wilcoxon Signed Rank tests also indicated that within-group differences between the second testing period and the third testing period were not significant for the first-treatment group ($z$ = 0.333, $p$ = .739) or the wait-control group ($z$ = 0.791, $p$ = .429). Appendix D of the Supplementary Materials contains a table of the means, medians, and standard deviations of the language analyses.

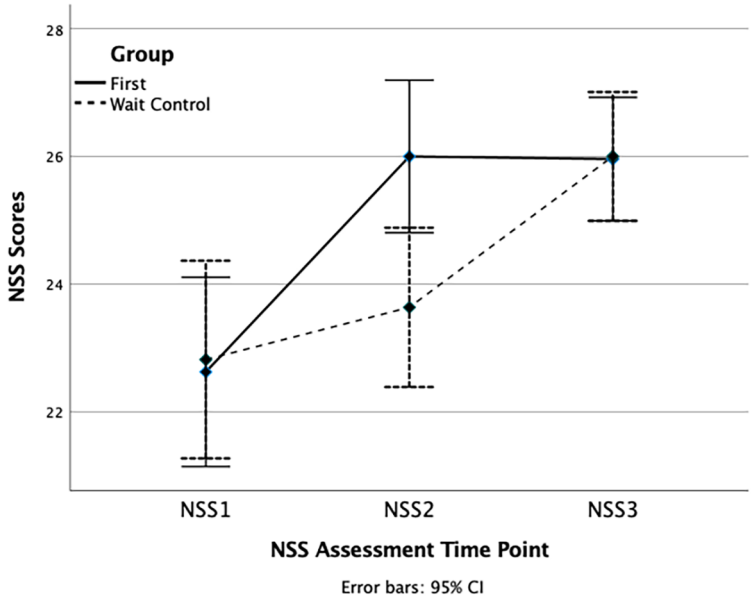
## Numeracy Results

Results of the RMM ANOVA indicated that the mean difference (0.723) of numeracy scores between the groups was not significant ($F_{1,44}$ = 0.918, $p$ = .343, 95% CI [−0.798, 2.245]); however, there was a significant interaction between Group and Testing Periods ($F_{2, 88}$ = 6.432, $p$ = .002). These results are in line with a wait-control design in which all participants receive the instruction, but at different points in time. That is, if both groups received treatment, collapsed across the testing periods, the between-group differences should not be different if the treatment was effective for both groups.

Follow up univariate analyses for each of the testing periods indicated the mean difference (0.193) between the first-treatment group's numeracy scores and the wait-control group's scores was not significant on the pre-test/baseline ($F_{1, 44}$ = 0.033, $p$ = .857, 95% CI [−1.950, 2.336]). Similarly, the mean difference (1.176) on the third test, after removing the seven participants who did not have the third test, was not significant ($F_{1, 37}$ = 2.847, $p$ = .100, 95% CI [−0.236, 2.589]). However, the mean difference (2.36) of the numeracy scores between the groups was significantly different on the second test, the post-test for the first-treatment group ($F_{1, 44}$ = 7.603, $p$ = .008, $\eta_p^2$= 0.147, 95% CI [0.636, 4.091]; see Figure 1).

**Figure 1**

*Scores of the Number Sense Screener™ by Group Across Time*




*Note.* Participants were tested three times using the Number Sense Screener™ (Jordan et al., 2012). The first test (NSS1) served as the first-treatment group's pre-test and the wait-control group's baseline. The second test (NSS2) was the first-treatment group's post-test and the wait-control group's pre-test. The third test (NSS3) was the wait-control group's post-test and a maintenance test for the first-treatment group.

To determine the effect of using Native Numbers (Native Brain, 2014) for the wait-control group, we conducted a paired sample *t*-test from the second testing point to the third testing point, their pre-test to post-test period. Results indicated the use of Native Numbers elicited a mean increase of 2.364 ($SE$ = 0.339) in the wait-control group's numeracy scores, $t(6.973)$, $p < .001$, $d = 1.48$, 95% CI [1.659, 3.609]). Thus, after using Native Numbers both groups significantly increased their numeracy scores, as measured by the Number Sense Screener™ (Jordan et al., 2012).

Removing the seven participants who did not have a maintenance test, we conducted a *t*-test to examine if the first treatment group maintained their previous gains. The results indicated that the first-treatment group maintained their growth, and slightly increased their numeracy scores ($M$ = 0.765, $SE$ = 0.689), although the changes were not statistically significant, $t(16) = 1.110$, $p = .283$, 95% CI [0.696, 2.225]. See Table 2 for the descriptive statistics for the current study and Table 3 for the Dias (2016) study.

**Table 2**

*Descriptive Statistics of the NSS Scores for the Current Study*

| Group / Test | *N* | *M* | *SD* | *SE* |
|---|---|---|---|---|
| First Group Pre-Test | 24 | 22.63 | 3.62 | 0.74 |
| Wait-Control Baseline | 22 | 22.82 | 3.58 | 0.76 |
| First Group Post-Test | 24 | 26.00 | 2.36 | 0.48 |
| Wait-Control Pre-Test | 22 | 23.64 | 3.40 | 0.73 |
| First Group Maintenance | 17[a] | 27.18[a] | 1.94 | 0.47 |
| Wait-Control Post-Test | 22 | 26.00 | 2.31 | 0.49 |

*Note.* NSS = The Number Sense Screener™ (Jordan et al., 2012).
[a]This number reflects removing the seven participants from the first group who did not have the maintenance assessment.

**Table 3**

*Descriptive Statistics of the NSS Scores for the Dias (2016) Study*

| Group / Testing | *N* | *M* | *SD* |
|---|---|---|---|
| Treatment Group Pre-Test | 27 | 16.63 | 4.84 |
| Control Group Pre-Test | 30 | 19.05 | 5.10 |
| Treatment Group Post-Test | 27 | 22.04 | 3.87 |
| Control Group Post-Test | 30 | 18.87 | 5.13 |

*Note.* NSS = The Number Sense Screener™ (Jordan et al., 2012).

## Discussion

The purpose of this conceptual replication study was to examine changes in numeracy achievement, intrinsic motivation, and mathematics language when instruction was provided by Native Numbers (Native Brain, 2014), an ITS. As a conceptual replication, we aimed to use the same assessment instruments, grade level, number of classrooms, and participant demographics as the original study (Dias, 2016). For pragmatic reasons, we chose to conduct the study in a geographical location close to the authors. However, we manipulated five distinct aspects of our study: the time of year of the study, the inclusion of a wait-control group, the inclusion of an assessment of mathematical language, the setting within the school where participants used Native Numbers, and who facilitated and monitored the participants' assessments and use of Native Numbers. The last two changes came about through collaborative discussions with the participating teachers prior to the study's start and were included in the IRB proposal as potentially malleable components to change based on the teachers' preferences.

Our first research question was whether we could replicate similar effects as Dias (2016) on numeracy outcomes (e.g., $\eta_p^2 = 0.622$) given our changes in design. While we did not replicate the same effect size as Dias (2016), the effect size in the current study was moderate to high: $\eta_p^2 = 0.147$. Within-group gains for the first-treatment group and the wait-control group were significant after receiving instruction via Native Numbers (Native Brain, 2014); and the first group maintained their achievement for at least one month.

From a design perspective, we matched participant demographics as closely as possible regarding the type of school (i.e., private) and socio-economic status. One plausible explanation for differences in effect sizes is that the studies occurred at different times of the year: fall versus spring. Initial pre-treatment scores differed between the Dias (2106) groups and our groups. The post-test scores for the Dias treatment group were close to the pre-test scores in our sample: $M = 22.04$ for Dias compared to $M = 22.72$ for the entire sample in the current study. Even without treatment, the participants in our entire sample had close to the same scores in the middle of February as the Dias treatment group had post-treatment in the fall. We inferred that kindergarten participants in our sample eventually had the same scores as the Dias' participants had in the fall; however, the participants in our study did not reach this level of achievement until well past the middle of the school year. We did not have the beginning of the year data for the current sample as a direct comparison to the Dias pre-test scores for the same time period, nor end of the year data for the Dias' participants; however, one implication is higher numeracy earlier in the year may allow for additional growth in numeracy for the remainder of the year.

Two additional changes in the current study from the Dias (2016) study that may explain differences in the effect sizes were who facilitated the supplemental numeracy activities and where the participants engaged in these supplemental activities. Specifically, having more students working on Native Numbers (Native Brain, 2014) in a computer lab-type setting, as in the current study, leaving a lower teacher-to-student ratio to work with the wait-control group, could have provided the wait-control group an advantage over the control group in the Dias study, as well as over the first-treatment group in the current study. The regular teacher-to-student ratio in our sample was 1:9 as opposed to the 1:19 ratio in the Dias study. With rare exception, only one of the researchers in the current study monitored all the treatment participants who were gathered in one room. Thus, the *researcher*-to-student ratio average was 1:27 for the treatment group, but the *teacher*-to-student ratio was 1:5 for the wait-control group. The *teacher*-to-student ratio for the

Dias study participants remained 1:19 as the classroom teachers monitored both the treatment and the control groups at the same time. Despite the differences in the adult-to-participant ratios, both groups in the current study significantly increased their numeracy scores after using Native Numbers. Therefore, having more students together in a larger room did not appear to negatively impact the first treatment group's outcome scores compared to the wait-control group. Regardless, whether the differences in the magnitude of effect sizes were due to differences in the adult-to-student ratios is unknown.

Our second research question was whether we could replicate similar intrinsic motivation findings like those in the Dias (2016) study. Results of the current study indicated that the use of Native Numbers (Native Brain, 2014) did not significantly increase between-group differences for intrinsic motivation for mathematics or reading. The only significant change in motivation was for the first treatment group for mathematics. Given the small sample size and lack of between-group differences, we do not have an explanation for why the first treatment group increased their intrinsic motivation for mathematics and the wait-control group did not.

One thing to note is that, as an indicator of being intrinsically motivated to read or to engage in mathematics, the highest score possible was 33. A score of 22 would indicate that a participant was neither intrinsically motivated nor unmotivated to read or do mathematics; an 11 would indicate a participant was generally unmotivated to read or engage in mathematics. For both groups, the medians for reading and mathematics leaned towards being more intrinsically motivated; for example, the lowest median (24.50) was on the pre-test for mathematics. The significant modifications we made when implementing the motivation inventory and the difficulty the participants had interpreting how to respond to the negative questions may explain this null finding. However, Dias (2016) noted that even though the treatment group's intrinsic motivation scores increased significantly compared to the control group, intrinsic motivation scores did not predict numeracy outcomes.

Our last research question looked at mathematical language changes, a measure not included in the Dias (2016) study. We did not find differences in mathematical language achievement after using Native Numbers (Native Brain, 2014). The null results are partially explained by the fact that nearly all participants were at or near ceiling at the pre-test. The Preschool Assessment of Mathematical Language (Purpura & Logan, 2015) was designed for pre-kindergarten, not kindergarten. Based on the results of our analyses, we cannot report an impact on mathematical language after using Native Numbers.

## Limitations

The current study had notable limitations preventing generalization to different contexts and demographics. Importantly, both studies included small sample sizes and were conducted in private schools with predominately Caucasian participants. Additionally, unlike the teacher-to-student ratio in the Dias (2016) study (i.e., 1:19), our participants had an average teacher-to-student ratio of 1:9; a ratio unlikely to occur in public school settings in the United States. Furthermore, a researcher monitored the use of Native Numbers (Native Brain, 2014) in the current study, compared the participants' teachers in the Dias (2016) study. While the ratio of adults-to-participants provided a proof of concept that Native Numbers can be utilized in a computer lab type of setting, or with entire classrooms at one time, how the role of a researcher monitoring use of Native Numbers impacted outcomes (i.e., compared to a lab instructor or classroom teacher) is unknown and was not a research question explored in the current study. Importantly, however, the researchers did not provide additional instruction or help other than problem-solving technology issues or providing encouragement for participants to think through a problem when they were stuck.

Next, the mathematical language measure used in the current study was not designed for kindergarten. We caution readers not to use the results from our study to compare the results to other mathematical language studies using the Preschool Assessment of Mathematical Language (Purpura & Logan, 2015) with pre-kindergarten students, the age group for which the assessment was designed. Likewise, the Young Children's Academic Intrinsic Motivation Inventory (Gottfried, 1990) was not designed for kindergarten. The lack of grade-level-appropriate standardized measures for both intrinsic motivation and mathematical language in kindergarten impedes our understanding of these two roles in numerical cognition for kindergarten aged participants.

Additionally, one of the primary limitations preventing direct comparison of the current study to the Dias (2016) study was that all the participants in the Dias study completed all 25 Native Numbers' (Native Brain, 2014) activities to the highest level; the participants in the current study did not. The time of year we started the study limited the number of days available for instruction. Starting earlier in the year may have allowed all participants to finish. However, this may have confounded the variable of interest related to allowing participants in the current study more time attending formal schooling before using Native Numbers. Nonetheless, including extra days in the current study was not possible.

Because of the highly adaptive nature of ITSs, we planned in advance for the possibility that some students would finish quickly, while others would take more time. We considered several different options. For example, before entering the wait-control group into treatment status, we could have administered the post-assessments to the first treatment group participants who had not yet completed all 25 activities to the fifth level and then moved them into the business-as-usual status. Alternatively, we could have kept the participants who had finished and had received post-tests in the same room with the participants who were still working on Native Numbers (Native Brain, 2014). A third option was to shorten the study's duration by setting a firm number of days for each phase. We did not consider any of the above options as being consistent with an ecologically valid design.

First, we were concerned about a potentially demotivating impact for both those who finished early and those still working if we kept the students who had finished together in the same room. If we had kept the treatment participants together, we would have needed an additional activity to keep those who finished early occupied, introducing a confounding variable. Conversely, removing the participants who had not yet finished might have been seen as unfair if they were engaged and wanted to finish. We planned for and allowed participants to choose how long they wanted to work, not only to maintain ecological validity but also to acknowledge and honor the participants' agency.

We also chose not to limit the number of days participants used Native Numbers (Native Brain, 2014) as this would have: (a) prohibited a direct comparison of outcome scores from the current study to those of the Dias (2016) study, and (b) ignored results from other studies that included adaptive software where the authors noted that shorter lengths of time for the interventions might not have been adequate for some participants (e.g., Garduno, 2016; Salminen et al., 2015). At the design phase of this study, we explicitly chose to monitor how the first treatment group participants progressed and then decided when to enter the wait-control group based on the number of days remaining in the year. Importantly, the conceptual changes for this replication were driven by the question of whether the time of year was a potential variable for differences in effect sizes, if effect sizes were replicated. The addition of the wait-control group was an additional measure of validity for any significant outcomes between the first treatment group and the wait-control group but was not required as a replication of the Dias study.

Furthermore, we theorized that if participants in the first treatment group needed additional time, these students could potentially represent a subset of students at risk for MD, and analyzing their data could potentially inform future studies of MD. Results from our study did not indicate risk for MD. The outcome scores for the first treatment group participants who had not finished all activities to the $5^{th}$ level by the $22^{nd}$ day were not significantly different from the outcome scores of the wait-control group participants who did not finish the 25 activities to the highest level. Except for two participants in the first group who needed additional days, the discrepancies in the number of days *available* for working and the number of days students actually worked were due to absences, the classroom teachers pulling the students to conduct other classroom assessments, or participants working in groups with teachers for class-wide projects requiring intact groups.

While we monitored the days students used Native Numbers (Native Brain, 2014) and cross-checked the days with the attendance records and daily screen shots, we did not track the number of minutes each student used Native Numbers. This was an intentional decision on our part based partially on the logistics (e.g., the lack of researchers available for notation and IRB requirements for videotaping participants) and partially due to the dynamics of the naturalistic setting: students interacting with their peers showing their progress (e.g., Lim, 2012), the freedom to move about the room or to leave to get a drink of water, and entering the room later or leaving sooner to work with their classroom teachers on other projects. Likewise, we did not ask the developers to extract the number of minutes from the log-file data because we would have needed additional data to verify the number of minutes logged into Native Numbers equated to the minutes the students were actively engaged (e.g., Bond & Bedenlier, 2019; Fincham et al., 2019).

Beyond the limitations related to tracking minutes of active engagement, the range of the difference in the number of days the participants required to reach the highest level across all the activities is an example of the logistical challenges of incorporating highly individualized, adaptive instruction (e.g., Bondie et al., 2019), whether by practitioners or researchers. That is, determining how to balance instruction at the group level, while also providing instruction for students who finish well ahead, or behind, their peers, is a problem of practice. On the other hand, highly adaptive software theoretically affords the individualized instruction that students need.

## Future Directions for Classroom Use and Research

While the current study contributes to the scant literature of replication studies of numerical cognition, based on the limitations described above, we outlined several possibilities for future research: the timing of and need for additional support while using technology, the generalization of concepts, and specific future numeracy research possibilities using Native Numbers (Native Brain, 2014), other ITS, or other software.

First, the use of Native Numbers (Native Brain, 2014), regardless of the time of year initiated, brought the majority of students' levels of achievement up to the highest percentiles, according to the Number Sense Screener™ (Jordan et al., 2012). Where Dias (2016) reported a between-group difference of $\eta_p^2 = 0.622$ at the beginning of the school year, our between group-difference in late spring was well above a moderate effect ($\eta_p^2 = 0.147$). Although none of the participants in our study would have met the definition of MD if using a <30% cut-off score, some participants were close to that cut-score at the pre-test (e.g., the 34$^{th}$ percentile) and required additional time to complete all the activities. Thinking about *who* is at risk for MD, and that the National Assessment of Educational Progress (National Center for Education Statistics [NAEP], 2017) report indicated 60% of 4$^{th}$ grade students in the United States performed at or below proficiency, the use of a cut score below the 30th percentile for determining who needs additional assistance may not capture all students needing support. Students at the edge of any cut score, or performing above but not to proficiency, are not considered at risk for MD, yet may not have a solid conceptual understanding of numeracy. The 60% of students in the United States performing at or below proficiency likely needed additional, or different, instruction well before 4$^{th}$ grade, even if their performance was not below the 30$^{th}$ percentile, or other percentiles used for tiered response to intervention support.

Together, the findings from the current study and the Dias (2016) study provided evidence that the use of Native Numbers (Native Brain, 2014) as a supplement to instruction can increase numeracy achievement, even for students performing at or above-average performance; although, the sample sizes of both studies do not support a definite conclusion. One question of interest is how the instruction of Native Numbers, which only includes quantities of 1–9, allowed participants to significantly increase their numeracy scores on the Number Sense Screener™ (Jordan et al., 2012) for items that assessed content not provided by Native Numbers. The Number Sense Screener™ contains 29 items; ten directly aligned with the content and range of quantities instructed by Native Numbers. At the middle of the year pre-test, 40% of the kindergarten participants performed above the 90$^{th}$ percentile. At the end of the study that number increased to 74% and only one participant fell below the 80$^{th}$ percentile. Tools such as Native Numbers, which provide intelligent tutoring, may offer what Fuchs, Fuchs, and Compton (2012) described as "secondary prevention" or "booster lessons" (Bryant et al., 2008). Dias (2016) reported changes from the pre-test to the post-test for the treatment group moved participants from the 68$^{th}$ percentile to 91st percentile on the Number Sense Screener™. Implementing highly individualized technology brought the scores of most participants using Native Numbers, in both studies, up to the highest levels of performance. However, the highly individualized tutoring also introduced challenges related to the number of days each student needed to reach the highest level in Native Numbers. Future studies are warranted to investigate Native Numbers' pedagogy for the potential mechanisms of generalization to concepts not included in Native Numbers' instruction. Additionally, different study designs such as staggered entry or single-subject studies with multiple baselines may afford ways to navigate the logistical challenges of research incorporating adaptive software.

A second research opportunity is to analyze the log-file data produced when using technology within an experiment (see a suggested framework by Heffernan & Heffernan, 2014). All technology can track interactions, affording fine-grained analyses of performance (e.g., Heffernan & Heffernan, 2014; Larkin & Milford, 2018; Rau et al., 2013). Native Numbers (Native Brain, 2014) has features that are well suited to the types of micro-genetic studies that log-file

data can provide such as how individuals vary when acquiring flexible use of different representations of quantities and the relationships between shared underlying numeric structures (e.g., Goldwater & Schalk, 2016; Parviainen, 2019). For example, Native Numbers includes three sets of five activities using the different representations for quantity recognition, ordinality, and relational language. Studies of relational language and ordinality with young participants are underexplored areas of numerical cognition, as are studies using number rods, which are introduced as the first activity (e.g., Coles & Sinclair, 2018; Goffin & Ansari, 2016; Schalk et al., 2016; Morsanyi et al., 2018; Venenciano et al., 2012).

Additionally, Dias (2016) noted participants required between 1000–3000 tasks to reach mastery of the 25 sets of numeracy activities. In our study, that range was approximately 1200–7700. While these numbers may seem incredibly high, at face value, the numbers align to a finding in a recent approximate number system training study with adult participants conducted by Cochrane et al. (2019). Some (expert) adults required more than 20 days of training and over 8000 trails to increase approximate number system performance. Cochran et al. noted the long, slow progression of change for a novel task. In a study with prekindergarten participants using a numeracy app, Broda et al. (2019) noted that when participants completed 400 practice opportunities, their accuracy increased 80%. In another study, although not with technology, Doabler et al. (2019) examined the ratio of teacher instruction-to-student practice opportunities and noted that a ratio of 1:3 significantly increased kindergarten participants' achievement; that is for every one instance of teacher delivered explicit instruction, participants needed three opportunities to practice.

Why would the number of tasks or practice attempts matter? A commonly accepted theory in early numeracy is that acquiring mastery of the counting system takes years (e.g., Geary et al., 2019). However, how many years and how many tasks (i.e., attempts, practices) an individual needs are unknown. Like the literature on mathematical relational vocabulary in kindergarten, the literature on how much practice students need to master numerical skills is virtually silent. Nonetheless, schools implement pacing guides and standards of practice assuming a normal progression of mastery without a clear understanding of how long a normal progression to mastery takes. McLean and Rusconi (2014) argued, "no empirically-based normative developmental trajectory for mathematics learning has been yet established, and the non-problematic range of variation for age-appropriate levels is currently untested" (p. 1). Educators, policymakers, and researchers would benefit from knowing how many tasks, and therefore, how much time or differentiation is required to reach fluency, as well as how long this information remains in long-term memory. The use of log-file data from software like Native Numbers (Native Brain, 2014) can help us answer questions like those above because of the fine-grained levels of data produced while learning (e.g., R. Baker, 2016; U.S. Department of Education, Office of Educational Technology, 2012).

However, tracking the number of tasks alone is likely insufficient to capture the process of learning, just as tracking the number of minutes the program is running is insufficient to capture learning. Despite the significant gains in the current study, not all participants had the same amount of growth. As discussed previously, we did not examine possible reasons why some participants did not finish all 25 activities to the highest level within the 22-day timeframe. Research on how kindergarten students, both at risk and not at risk for MD, present with different degrees of difficulty is necessary to help researchers and teachers understand possible correlations between general domain abilities, self-determination, resilience, and goal orientations; specifically in the domain of mathematics, but also when using technology (e.g., Higgins et al., 2019; Hohol et al., 2017; Keller & Libertus, 2015; Matusz et al., 2019; Peng et al., 2016; White et al., 2017). To understand the process of learning likely requires mixed-methods studies incorporating multiple data sources (e.g., Baccaglini-Frank et al., 2020; Fincham et al., 2019). One option is to use the screen capturing capabilities of iPads while the students are using Native Numbers (Native Brain, 2014), or other software. Audio and video data may provide information such as meta-cognitive self-talk, social dialogue, and potentially different problem solving strategies for the specific numeracy skills designed into Native Numbers. Audio and video data captured by the iPads during use also reduces the need for the number of observers and additional equipment required for an observational study.

Finally, neither our study nor the Dias (2016) study considered how using an alternative software program, whether an ITS or not, may affect differences in outcome scores, persistence, or motivation. Native Numbers (Native Brain, 2014) provides valuable lab space to investigate different types of numerical representation, different instructional methods such as blocked or distributed learning and reversibility problems (e.g., Barzagar Nazari & Ebersbach, 2018), as well as relational language and ordinality. Given the above opportunities to advance knowledge of early numeracy,

future studies are warranted using either adapted versions of the current Native Numbers model, incorporating another software into the design of a study, or using another ITS.

Classroom teachers need viable, economically feasible means of providing individualized instruction. However, not all software, or other curricular resources, are designed with substantial pedagogy, and the role of the teacher is critical to the success of any curriculum, supplemental or otherwise (e.g., Nye, 2014; Wen et al., 2019). The Dias (2016) study was an initial confirmation of Native Numbers' (Native Brain, 2014) effectiveness, and before embarking on the type of research described above, we sought to replicate effectiveness. Although our effect size was not as substantial as that of the Dias study, both studies' effect sizes were moderate to extremely high for a small homogeneous sample of 73 participants combined. Nonetheless, the Dias study and our conceptual replication provided emerging evidence that using Native Numbers for supplemental instruction holds promise as a useful tool, a one-to-one tutor in the room, and as a tool for researching numerical cognition.

**Funding:** The authors have no funding to report.

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

**Competing Interests:** The authors have declared that no competing interests exist.

**Data Availability:** A synthetic data set is available online (Grimes et al., 2021a). Submission of actual data was prohibited by the Institutional Review Board approval for this study.

## Supplementary Materials

The Supplementary Materials contain the following items (for access see Index of Supplementary Materials below):

- **A deidentified synthetic data set in Excel format**
- **The following five appendices:**
  1. Objectives and sequence of instruction in Native Numbers
  2. List of activites the active control group completed during math centers
  3. Descriptive statistics and syntax of analyses of multiple imputation of missing data
  4. Descriptive statistics for motivation and mathematical language measures
  5. Syntax of analyses for the differences in performane of the 13 particpants who did not complete all 25 activities to the highest level and syntax for the number sense screener

### Index of Supplementary Materials

Grimes, K. R., Park, S., McClelland, A., Park, J., Lee, Y. R., Nozari, M., Umer, Z., Zaparolli, B., & Bryant, D. (2021a). *Supplementary materials to "Effectiveness of a numeracy intelligent tutoring system in kindergarten: A conceptual replication"* [Synthetic data and codebook]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5234

Grimes, K. R., Park, S., McClelland, A., Park, J., Lee, Y. R., Nozari, M., Umer, Z., Zaparolli, B., & Bryant, D. (2021b). *Supplementary materials to "Effectiveness of a numeracy intelligent tutoring system in kindergarten: A conceptual replication"* [Appendices]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5233

Journal of Numerical Cognition. (Ed.). (2021). *Supplementary materials to "Effectiveness of a numeracy intelligent tutoring system in kindergarten: A conceptual replication"* [Open peer-review]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5226

## References

Alcock, L., Ansari, D., Batchelor, S., Bisson, M. J., De Smedt, B., Gilmore, C., . . . Jones, I. (2016). Challenges in mathematical cognition: A collaboratively-derived research agenda. *Journal of Numerical Cognition, 2*, 20-41. https://doi.org/10.5964/jnc.v2i1.10

Aleven, V. (2015). A is for Adaptivity, but what is Adaptivity? Re-defining the field of AIED. In K. PorayskaPomsta, G. McCalla, & B. du Boulay (Eds.), *Proceedings of the workshop at the 17th International Conference on Artificial Intelligence in Education (AIED 2015)*

*on Les Contes du Mariage: Should AI stay married to Ed?* (Vol. 4, pp. 11–20). Retrieved from http://ceur-ws.org/Vol-1432/ai_ed_proc.pdf

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC, USA: American Psychiatric Publishing.

Baccaglini-Frank, A., Carotenuto, G., & Sinclair, N. (2020). Eliciting preschoolers' number abilities using open, multi-touch environments. *ZDM Mathematics Education, 52*, 779-791. https://doi.org/10.1007/s11858-020-01144-y

Baker, R. (2016). Using learning analytics in personalized learning. In M. Murphy, S. Redding, & J. Twyman (Eds.), *Handbook on personalized learning for states, districts, and schools* (pp. 165–174). Philadelphia, PA, USA: Center on Innovations in Learning. Retrieved from https://files.eric.ed.gov/fulltext/ED568173.pdf#page=181

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education, 26*(2), 600-614. https://doi.org/10.1007/s40593-016-0105-0

Barnes, M. A., Klein, A., Swank, P., Starkey, P., McCandliss, B., Flynn, K., . . . Roberts, G. (2016). Effects of tutorial interventions in mathematics and attention for low-performing preschool children. *Journal of Research on Educational Effectiveness, 9*(4), 577-606. https://doi.org/10.1080/19345747.2016.1191575

Barzagar Nazari, K., & Ebersbach, M. (2018). Distributed practice: Rarely realized in self-regulated mathematical learning. *Frontiers in Psychology, 9*, Article 2170. https://doi.org/10.3389/fpsyg.2018.02170

Bond, M., & Bedenlier, S. (2019). Facilitating student engagement through educational technology: Towards a conceptual framework. *Journal of Interactive Media in Education, 1*, Article 11. https://doi.org/10.5334/jime.528

Bondie, R. S., Dahnke, C., & Zusho, A. (2019). How does changing "one-size-fits-all" to differentiated instruction affect teaching? *Review of Research in Education, 43*(1), 336-362. https://doi.org/10.3102/0091732X18821130

Bourdeau, J., & Grandbastien, M. (2010). Modeling tutoring knowledge. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in intelligent tutoring systems* (Studies in computational intelligence, Vol. 208, pp. 123–143). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-14363-2_7

Broda, M., Tucker, S., Ekholm, E., Johnson, T. N., & Liang, Q. (2019). Small fingers, big data: Preschoolers' subitizing speed and accuracy during interactions with multitouch technology. *The Journal of Educational Research, 112*(2), 211-222. https://doi.org/10.1080/00220671.2018.1486281

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008). Mathematics intervention for first-and second-grade students with mathematics difficulties: The effects of tier 2 intervention delivered as booster lessons. *Remedial and Special Education, 29*(1), 20-32. https://doi.org/10.1177/0741932507309712

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., & Hiebert, J. (2018). The role of replication studies in educational research. *Journal for Research in Mathematics Education, 49*(1), 2-8. Retrieved from https://doi.org/10.5951/jresematheduc.49.1.0002

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K–12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88-113. https://doi.org/10.1016/j.edurev.2013.01.001

Chinn, S., Ashcroft, R., & Ashcroft, R. E. (2017). *Mathematics for dyslexics and dyscalculics: A teaching handbook* (4th ed.). Chichester, United Kingdom: John Wiley & Sons.

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*(5), 257-265. https://doi.org/10.1177/1053451214560888

Cochrane, A., Cui, L., Hubbard, E. M., & Green, C. S. (2019). "Approximate number system" training: A perceptual learning approach. *Attention, Perception & Psychophysics, 81*, 621-636. https://doi.org/10.3758/s13414-018-01636-w

Coles, A., & Sinclair, N. (2018). Re-thinking 'normal' development in the early learning of number. *Journal of Numerical Cognition, 4*(1), 136-158. https://doi.org/10.5964/jnc.v4i1.101

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education, 37*(4), 244-253. https://doi.org/10.1177/0741932516648463

Dennis, M. S., Sharp, E., Chovanes, J., Thomas, A., Burns, R. M., Custer, B., & Park, J. (2016). A meta-analysis of empirical research on teaching students with mathematics learning difficulties. *Learning Disabilities Research & Practice, 31*(3), 156-168. https://doi.org/10.1111/ldrp.12107

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31-54. https://doi.org/10.1016/j.edurev.2018.02.002

Dias, L. (2016). *An adaptive computer-based kindergarten curriculum for number sense* (Unpublished doctoral dissertation). Rivier University, Nashua, NH, USA.

Doabler, C. T., Gearin, B., Baker, S. K., Stoolmiller, M., Kennedy, P. C., Clarke, B., . . . Smolkowski, K. (2019). Student practice opportunities in core mathematics instruction: Exploring for a goldilocks effect for kindergartners with mathematics difficulties. *Journal of Learning Disabilities, 52*(3), 271-283. https://doi.org/10.1177/0022219418823708

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J. P., & Gašević, D. (2019). Counting clicks is not enough: Validating a theorized model of engagement in learning analytics. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 501-510). https://doi.org/10.1145/3303772.3303775

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children, 78*(3), 263-279. https://doi.org/10.1177/001440291207800301

Garduno, A. E. (2016). *Preschool and educational technology: Evaluating a tablet-based math curriculum in Mexico City* (Doctoral dissertation, Harvard University, Cambridge, MA, USA). Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:2711271

Geary, D. C. (2015). Preschool children's quantitative knowledge and long-term risk for functional innumeracy. In S. Chinn (Ed.), *The international handbook for mathematical difficulties and dyscalculia* (pp. 235–242). Abingdon, United Kingdom: Routledge.

Geary, D. C., Vanmarle, K., Chu, F. W., Hoard, M. K., & Nugent, L. (2019). Predicting age of becoming a cardinal principle knower. *Journal of Educational Psychology, 111*(2), 256-267. https://doi.org/10.1037/edu0000277

Goffin, C., & Ansari, D. (2016). Beyond magnitude: Judging ordinality of symbolic number is unrelated to magnitude comparison and independently relates to individual differences in arithmetic. *Cognition, 150*, 68-76. https://doi.org/10.1016/j.cognition.2016.01.018

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin, 142*(7), 729-757. https://doi.org/10.1037/bul0000043

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

Gottfried, A. E. (1982). Relationships between academic intrinsic motivation and anxiety in children and young adolescents. *Journal of School Psychology, 20*, 205-215. https://doi.org/10.1016/0022-4405(82)90050-4

Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology, 77*, 631-645. https://doi.org/10.1037/0022-0663.77.6.631

Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology, 82*, 525-538. https://doi.org/10.1037/0022-0663.82.3.525

Harris, T., & Hardin, J. W. (2013). Exact Wilcoxon signed-rank and Wilcoxon Mann–Whitney rank sum tests. *The Stata Journal, 13*(2), 337-343. https://doi.org/10.1177/1536867X1301300208

Harskamp, E. (2014). The effects of computer technology on primary school students' mathematics achievement: A meta-analysis. In S. Chinn (Ed.), *The Routledge international handbook of dyscalculia* (pp. 383–392). Abingdon, United Kingdom: Routledge.

Hassinger-Das, B., Jordan, N. C., & Dyson, N. (2015). Reading stories to learn math: Mathematics vocabulary instruction for children with early numeracy difficulties. *The Elementary School Journal, 116*(2), 242-264. https://doi.org/10.1086/683986

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*, 470-497. https://doi.org/10.1007/s40593-014-0024-x

Higgins, K., Huscroft-D'Angelo, J., & Crawford, L. (2019). Effects of technology in mathematics on achievement, motivation, and attitude: A meta-analysis. *Journal of Educational Computing Research, 57*(2), 283-319. https://doi.org/10.1177/0735633117748416

Hohol, M., Cipora, K., Willmes, K., & Nuerk, H.-C. (2017). Bringing back the balance: Domain-general processes are also important in numerical cognition. *Frontiers in Psychology, 8*, Article 499. https://doi.org/10.3389/fpsyg.2017.00499

Hojnoski, R. L., Columba, H. L., & Polignano, J. (2014). Embedding mathematical dialogue in parent–child shared book reading: A preliminary investigation. *Early Education and Development, 25*(4), 469-492. https://doi.org/10.1080/10409289.2013.810481

Hornburg, C. B., Schmitt, S. A., & Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *Journal of Experimental Child Psychology, 176*, 84-100. https://doi.org/10.1016/j.jecp.2018.07.005

Jennings, C. M., Jennings, J. E., Richey, J., & Dixon-Krauss, L. (1992). Increasing interest and achievement in mathematics through children's literature. *Early Childhood Research Quarterly, 7*(2), 263-276. https://doi.org/10.1016/0885-2006(92)90008-M

Jordan, N. C., Glutting, J. J., & Dyson, N. (2012). *Number sense screener™(NSS™) User's guide, k–1, research edition.* Baltimore, MD, USA: Brookes Publishing.

Jordan, N. C., Rinne, L., & Hansen, N. (2019). Mathematical learning and its difficulties in the United States: Current issues in screening and intervention. In A. Fritz, V. Hasse, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 183–199). https://doi.org/10.1007/978-3-319-97148-3

Keller, L., & Libertus, M. (2015). Inhibitory control may not explain the link between approximation and math abilities in kindergarteners from middle class families. *Frontiers in Psychology, 6*, Article 685. https://doi.org/10.3389/fpsyg.2015.00685

Kiru, E. W., Doabler, C. T., Sorrells, A. M., & Cooc, N. A. (2018). A synthesis of technology-mediated mathematics interventions for students with or at risk for mathematics learning disabilities. *Journal of Special Education Technology, 33*(2), 111-123. https://doi.org/10.1177/0162643417745835

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42-78. https://doi.org/10.3102/0034654315581420

Larkin, K., & Milford, T. (2018). Mathematics apps—Stormy with the weather clearing: Using cluster analysis to enhance app use in mathematics classrooms. In N. Calder, K. Larkin, & N. Sinclair (Eds.), *Using mobile technologies in the teaching and learning of mathematics.* (Mathematics Education in the Digital Era, Vol. 12, pp. 11–30). https://doi.org/10.1007/978-3-319-90179-4

Lewis, K. E., & Fisher, M. B. (2016). Taking stock of 40 years of research on mathematical learning disability: Methodological issues and future directions. *Journal for Research in Mathematics Education, 47*(4), 338-371. https://doi.org/10.5951/jresematheduc.47.4.0338

Lim, E. M. (2012). Patterns of kindergarten children's social interaction with peers in the computer area. *International Journal of Computer-Supported Collaborative Learning, 7*(3), 399-421. https://doi.org/10.1007/s11412-012-9152-1

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918. https://doi.org/10.1037/a0037123

Matusz, P. J., Merkley, R., Faure, M., & Scerif, G. (2019). Expert attention: Attentional allocation depends on the differential development of multisensory number representations. *Cognition, 186*, 171-177. https://doi.org/10.1016/j.cognition.2019.01.013

McLean, J. F., & Rusconi, E. (2014). Mathematical difficulties as decoupling of expectation and developmental trajectories. *Frontiers in Human Neuroscience, 8*, Article 44. https://doi.org/10.3389/fnhum.2014.00044

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, Article 104092. https://doi.org/10.1016/j.jml.2020.104092

Mononen, R., Aunio, P., Koponen, T., & Aro, M. (2014). A review of early numeracy interventions for children at risk in mathematics. *International Journal of Early Childhood Special Education, 6*, 25-54. https://doi.org/10.20489/intjecse.14355

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Who is at risk for persistent mathematics difficulties in the United States? *Journal of Learning Disabilities, 49*, 305-319. https://doi.org/10.1177/0022219414553849

Morsanyi, K., van Bers, B. M., O'Connor, P. A., & McCormack, T. (2018). Developmental dyscalculia is characterized by order processing deficits: Evidence from numerical and non-numerical ordering tasks. *Developmental Neuropsychology, 43*(7), 595-621. https://doi.org/10.1080/87565641.2018.1502294

Moyer-Packenham, P. S., Lommatsch, C. W., Litster, K., Ashby, J., Bullock, E. K., Roxburgh, A. L., . . . Clarke-Midura, J. (2019). How design features in digital math games support learning and mathematics connections. *Computers in Human Behavior, 91*, 316-332. https://doi.org/10.1016/j.chb.2018.09.036

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative models for small samples in psychological research: Applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educational and Psychological Measurement, 76*(1), 64-87. https://doi.org/10.1177/0013164415580432

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology, 4*(1), 13-20. https://doi.org/10.20982/tqmp.04.1.p013

National Association for the Education of Young Children. (2012). *Technology and interactive media as tools in early childhood programs serving children from birth through age 8.* Retrieved from https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/resources/topics/PS_technology_WEB.pdf

National Center for Education Statistics. (2017). *NAEP results*. Retrieved from https://www.nationsreportcard.gov/math_2017/nation/achievement?grade=4

National Council of Teachers of Mathematics. (2010). *Strategic use of technology in teaching and learning mathematics*. Retrieved from https://www.nctm.org/Standards-and-Positions/Position-Statements/Strategic-Use-of-Technology-in-Teaching-and-Learning-Mathematics

Native Brain. (2014). Native Numbers [Mobile Application Software]. Retrieved from www.nativebrain.com

Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology, 111*(6), 1001-1022. https://doi.org/10.1037/edu0000334

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities, 51*(6), 523-539. https://doi.org/10.1177/0022219417714773

Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly, 36*(3), 550-560. https://doi.org/10.1016/j.ecresq.2016.02.003

Nye, B. D. (2014). Barriers to ITS adoption: A systematic mapping study. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014: Proceedings* (pp. 583–590). https://doi.org/10.1007/978-3-319-07221-0_74

Ok, M. W., Bryant, D. P., & Bryant, B. R. (2020). Effects of computer-assisted instruction on the mathematics performance of students with learning disabilities: A synthesis of the research. *Exceptionality, 28*(1), 30-44. https://doi.org/10.1080/09362835.2019.1579723

Overall, J. E., Tonidandel, S., & Starbuck, R. R. (2009). Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts. *Social Science Research, 38*(2), 492-503. https://doi.org/10.1016/j.ssresearch.2009.01.004

Parviainen, P. (2019). The development of early mathematical skills–A theoretical framework for a holistic model. *Journal of Early Childhood Education Research, 8*(1), 162-191. Retrieved from https://jecer.org/fi/wp-content/uploads/2019/10/Parviainen-issue8-1.pdf

Pavlik, P. I., Brawner, K., Olney, A., & Mitrovic, A. (2013). Review of student models used in intelligent tutoring systems. In R. Sollilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 1. Learner modeling* (pp. 39–68). Adpehi, MD, USA: U.S. Army Research Laboratory.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*(3–5), 313-350. https://doi.org/10.1007/s11257-017-9193-2

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology, 108*(4), 455-473. https://doi.org/10.1037/edu0000079

Penner, M., Buckland, C., & Moes, M. (2019). Early identification of, and interventions for, kindergarten students at risk for mathematics difficulties. In K. Robinson, H. Osana, & D. Kotsopoulos (Eds.), *Mathematical learning and cognition in early childhood* (pp. 57–78). https://doi.org/10.1007/978-3-030-12895-1_5

Pirolli, P. (2014). Computer-aided instructional design systems. In H. Burns, J. W. Parlett, & C. L. Redfield (Eds.), *Intelligent tutoring systems: Evolutions in design* (pp. 105–126). New York, NY, USA: Psychology Press.

Powell, S. R., & Driver, M. K. (2015). The influence of mathematics vocabulary instruction embedded within addition tutoring for first-grade students with mathematics difficulty. *Learning Disability Quarterly, 38*(4), 221-233. https://doi.org/10.1177/0731948714564574

Purpura, D. J., King, Y. A., Rolan, E., Hornburg, C. B., Schmitt, S. A., Hart, S. A., & Ganley, C. M. (2020). Examining the factor structure of the home mathematics environment to delineate its role in predicting preschool numeracy, mathematical language, and spatial skills. *Frontiers in Psychology, 11*, Article 1925. https://doi.org/10.3389/fpsyg.2020.01925

Purpura, D. J., & Logan, J. A. (2015). The nonlinear relations of the approximate number system and mathematical language to early mathematics development. *Developmental Psychology, 51*(12), 1717-1724. https://doi.org/10.1037/dev0000055

Purpura, D. J., Napoli, A. R., Wehrspann, E. A., & Gold, Z. S. (2017). Causal connections between mathematical language and mathematical knowledge: A dialogic reading intervention. *Journal of Research on Educational Effectiveness, 10*(1), 116-137. https://doi.org/10.1080/19345747.2016.1204639

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction, 23*, 98-114. https://doi.org/10.1016/j.learninstruc.2012.07.003

Rights, J. D., & Sterba, S. K. (2020). New recommendations on the use of R-squared differences in multilevel model comparisons. *Multivariate Behavioral Research, 55*(4), 568-599. https://doi.org/10.1080/00273171.2019.1660605

Rozo, H., & Real, M. (2019). Pedagogical guidelines for the creation of adaptive digital educational resources: A review of the literature. *Journal of Technology and Science Education, 9*(3), 308-325. https://doi.org/10.3926/jotse.652

Salminen, J., Koponen, T., Leskinen, M., Poikkeus, A.-M., & Aro, M. (2015). Individual variance in responsiveness to early computerized mathematics intervention. *Learning and Individual Differences, 43*, 124-131. https://doi.org/10.1016/j.lindif.2015.09.002

Schalk, L., Saalbach, H., Grabner, R. H., & Stern, E. (2016). Relational quantitative reasoning in kindergarten predicts mathematical achievement in third grade. *Journal of Numerical Cognition, 2*(2), 77-90. https://doi.org/10.5964/jnc.v2i2.29

Simon, M. A., Kara, M., Placa, N., & Sandir, H. (2016). Categorizing and promoting reversibility of mathematical concepts. *Educational Studies in Mathematics, 93*(2), 137-153. https://doi.org/10.1007/s10649-016-9697-4

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970-987. https://doi.org/10.1037/a0032447

Twyman, J., & Sota, M. (2016) Educational technology and response to intervention: Affordances and considerations. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 493–517). https://doi.org/10.1007/978-1-4899-7568-3_29

U.S. Department of Education, Office of Educational Technology. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Retrieved from http://www.ed.gov/technology

van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research, 49*, 78-91. https://doi.org/10.1080/00273171.2013.855890

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Venenciano, L., Dougherty, B., & Slovin, H. (2012). *The Measure Up program, prior achievement, and logical reasoning as indicators of algebra preparedness.* Paper presented at the 12th International Congress on Mathematical Education (ICME-12), Seoul, South Korea.

Verma, J. P. (2015). *Repeated measures design for empirical researchers.* Hoboken, NJ, USA: John Wiley & Sons.

Wang, A. H., Firmender, J. M., Power, J. R., & Byrnes, J. P. (2016). Understanding the program effectiveness of early mathematics interventions for prekindergarten and kindergarten environments: A meta-analytic review. *Early Education and Development, 27*(5), 692-713. https://doi.org/10.1080/10409289.2016.1116343

Watson, S. M., & Gable, R. A. (2013). Unraveling the complex nature of mathematics learning disability: Implications for research and practice. *Learning Disability Quarterly, 36*(3), 178-187. https://doi.org/10.1177/0731948712461489

Wen, Z. A., Amog, A. L. S., Azenkot, S., & Garnett, K. (2019). Teacher perspectives on math e-learning tools for students with specific learning disabilities. Paper presented at the *21st International ACM SIGACCESS Conference on Computers and Accessibility.* Pittsburg, PA, USA. Retrieved from https://dl.acm.org/citation.cfm?id=3354607

White, R. E., Prager, E. O., Schaefer, C., Kross, E., Duckworth, A. L., & Carlson, S. M. (2017). The "Batman effect": Improving perseverance in young children. *Child Development, 88*(5), 1563-1571. https://doi.org/10.1111/cdev.12695

Wilson, C., & Scott, B. (2017). Adaptive systems in education: A review and conceptual unification. *International Journal of Information and Learning Technology, 34*(1), 2-19. https://doi.org/10.1108/IJILT-09-2016-0040

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson tests of achievement.* Riverside Publishing. Retrieved from http://www.riverpub.com/products/wjIIIComplete/index.html

*Journal of Numerical Cognition* (JNC) is an official journal of the Mathematical Cognition and Learning Society (MCLS).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.