

ABSTRACT

PEDDYCORD-LIU, ZHONGXIU AURORA. Game Learning Analytics and Qualitative Methods for Actionable Change in a Curriculum-Integrated Educational Math Game . (Under the direction of Dr. Tiffany Barnes).

Educational games have become a mainstay of learning for our new generation. As more games emerge, we face unprecedented challenges in evaluating games against their educational goals. Traditional methods fall short. Games are growing in size and complexity, which makes it increasingly inefficient to pinpoint game design flaws through controlled trial experiments. Games are intended to be used over longer periods of time, but pre- and post-tests have limited power in investigating the longitudinal process of gameplay and learning. We need new data-driven methods to understand gameplay and inform game design in an efficient, scalable fashion. The fact is, more games are being used in schools, but pure data-driven methods often ignore the rich context in classrooms. Thus, this work combines game learning analytics and qualitative methods to derive actionable insights for a curriculum-integrated K12 educational game: ST Math.

ST Math is a large-scale curriculum-integrated game used by over 1.2 million students across the U.S. In this dissertation, I have performed four studies on this game's data and users to derive actionable insights. In the first study, I analyzed how performance and learning associated with students' game replay behaviors. This analysis showed that replaying games was not necessarily beneficial, and some patterns may be associated with work-avoidance behaviors. In the second study, I designed data-driven methods to mine predictive relationships between math objectives and to inform the design of ST Math curriculum (sequence of math objectives). This study suggested how math objectives may connect to each other, such as that playing games involving money helps students understand fractions. In the third study, I applied learning curve analyses under different cognitive assumptions. This work pinpointed game levels where students failed to learn or transfer, identifying potential game design changes that could result in better learning. Lastly, I conducted a field study and qualitative analyses to investigate the practical gaps in the design and use of ST Math in classrooms. I identified critical needs and practices of teachers and students, and created a new framework that adds to our understanding of teacher activities around the use of a curriculum-integrated games. I also derived actionable game design feedback and generalizable insights for educational games to be used in classrooms.

My work has made several contributions. First, this dissertation is pioneering work on curriculum-integrated educational games, a type of game integrated into students' learning activities at school with increasing demands and popularity. This dissertation contributes several methods and insights that can benefit the future design and use of curriculum-integrated games that could impact the learning of our future generations. Second, this dissertation contributes to innovative data-driven

methods, educational insights, and actionable game design suggestions that not only benefit ST Math, but are applicable to other educational games and e-learning platforms. Thirdly, this dissertation serves as a model for integrating human insights with data-driven methods, combining qualitative and quantitative research to derive interpretable, actionable, and practical insights to benefit learners.

© Copyright 2018 by Zhongxiu Aurora Peddycord-Liu

All Rights Reserved

Game Learning Analytics and Qualitative Methods for Actionable Change
in a Curriculum-Integrated Educational Math Game

by
Zhongxiu Aurora Peddycord-Liu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2018

APPROVED BY:

Dr. Collin Lynch

Dr. Min Chi

Dr. Teomara Rutherford

Dr. Tiffany Barnes
Chair of Advisory Committee

DEDICATION

I dedicate this dissertation to all the great people who made a positive impact to my life.

My parents Hong Wang and Jie Liu who couldn't speak English, have never studied STEM, or attended graduate school, but supported me both financially and spiritually to pursue my dream in their clumsy way. I'm very lucky to have them as my parents.

Behind every strong woman there's a strong man. I love my incredibly smart, handsome, and loving husband Barry Peddycord III, who made my years of PhD life full of love and passion. He gives me the power of persisting through difficulties, by filling my life with more joys and ice cream than any other flavors. "If one e-mail is a melody, does that make a conversation a duet?"

BIOGRAPHY

Aurora was born in 1993 in Shenyang, China—a very cold city which contributes to her decision of pursuing PhD in the south. She went to the U.S. in 2010 to study math in Worcester Polytechnic Institute, but ended up falling in love with computer science. In 2014, she joined Dr. Tiffany Barnes lab at NC State, where she had the best years of her adult life. During which time, she also found a husband Barry, a cat Cinnamon Bun, and two chihuahuas Cocoa Bean and Eclair Puff. She has traveled to 15 countries and is proud of her art museums' postcard collection.

The author is an reinforcement algorithms who love being adventurous and going out of her comfort zone. She likes traveling, fashion, painting, and of course, Barry. She fancies herself as the lady detective Phryne Fisher. More about her at zhongxiuliu.com .

ACKNOWLEDGEMENTS

I love my advisor Dr. Tiffany Barnes. She is ridiculously cool just as people told me when I started to apply to graduate schools. She believes in me and helps me grow both personally and professionally. I am very proud to be her student. Without her, I would never had the confidence I have today.

I am grateful to my committee: Dr. Collin Lynch, Dr. Min Chi, and Dr. Teomara Rutherford. They sincerely want me success and provide me with support and resources to do so.

I am grateful for people in my lab: Dr. Veronica Catete, Dr. Behrooz Mostafavi, Dr. Andrew Hicks, Dr. Thomas Price, Dr. Michael Eagle, and soon-to-be Doctors Yihuan Dong, Rui Zhi, Christa Cody, Rachel Harred, Mehnak Manikala, Alex Milliken, Nick Lytle, David Warren et al. I am grateful to my friends at school, Dr. Luke Deshotel, soon-to-be Doctors Jennifer Tsan, Sean Mealin and many. Without them, my life outside research would be as bored as a pile of dead water.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	viii
Chapter 1 Introduction	1
1.1 ST Math, its Uniqueness and Generality	3
1.1.1 Data and Population	6
1.2 Research Questions	7
Chapter 2 Literature Review	9
2.1 Educational Games for Math	9
2.1.1 The Design	10
2.1.2 The Assessment	11
2.2 Educational Data Mining and Learning Analytics	12
2.2.1 Mining the Relationships in Educational Content	13
2.2.2 Constructing Learning Curves	14
2.3 Game Learning Analytics	16
2.4 Educational Games in the Classroom	17
2.5 Summary	18
Chapter 3 Study 1: Is Student-initiated Game Replay Worth the Effort?	20
3.1 Introduction	21
3.2 Background	21
3.3 Methods	22
3.3.1 Data & Features	22
3.3.2 Pass Attempt Features	23
3.3.3 Elective Replay Features	24
3.3.4 Student Grouping from ER Features	24
3.4 Results & Discussion	25
3.4.1 Who Engaged in Elective Replay?	25
3.4.2 What Gets Replayed, and When?	27
3.4.3 Is Elective Replay Associated with Gains?	28
3.5 Contribution	31
Chapter 4 Study 2: Inform Curricular Sequencing through Mining Predictive Relationships between Math Contents	33
4.1 Introduction	34
4.2 Background	34
4.3 Method	35
4.3.1 Participants & Data	35
4.3.2 Mining Predictive Relationship	39
4.4 Results & Discussion	40
4.4.1 Categorization of Objectives	40

4.4.2	Categorization of Objective Pairs	42
4.5	Contribution	47
Chapter 5	Study 3: Pinpoint where Learning & Transfer Support is Needed Using Learning Curve Analyses	49
5.1	Introduction	50
5.2	Background	50
5.3	Method	50
5.3.1	Data	50
5.3.2	Fitting Learning Curves	51
5.4	Result & Discussion	52
5.4.1	Analyzing Puzzles in Levels	52
5.4.2	Analyzing Levels in Games	57
5.4.3	Analyzing Games in Objective	59
5.5	Contribution	60
Chapter 6	Study 4: Teacher-focused Field Study on the Classroom Use of a Curriculum-integrated Game	62
6.1	Introduction	63
6.2	Background	63
6.3	Method	64
6.4	How were Teachers Using ST Math?	65
6.4.1	Preparation	66
6.4.2	Integration	67
6.4.3	Intervention	72
6.4.4	Data-Informed Practice	76
6.5	What Did Teachers Suggest?	77
6.6	Contribution	78
Chapter 7	Conclusion	80
7.1	RQ1	81
7.2	RQ2	83
7.3	RQ3	84
7.4	Future Work	86
7.5	Final Canapés	87
Bibliography	88
APPENDIX	102
Appendix A	Math Game Literature	103

LIST OF TABLES

Table 1.1	The First Six (out of 26) Objectives from Grade 3, 2012-2013 Curriculum, and the Total Number of Games, Levels, Puzzles and Lives in each Objective. NOTE: Numbers of Lives and Puzzles are Unevenly Distributed across Game Levels.	5
Table 1.2	Populations' Demographics Information.	7
Table 3.1	Example of Elective Replay and Pass Attempts.	23
Table 3.2	Elective Replay (ER) Features and Their Descriptive Statistics Among Students who Electively Replayed, Collapsed to the Student Granularity.	25
Table 3.3	Mann-Whitney U Tests Comparing Gameplay Characteristics Between Student Groups of Different Elective Replay Patterns.	26
Table 3.4	Decision Trees to Predict Levels whose Pass Attempts were Interrupted or Followed by Elective Replay.	28
Table 3.5	% of Data Observations with Gains, No Gains, and Percentage Dropped for the Three Gain Types.	29
Table 3.6	Mann-Whitney U Tests Comparing Gains Between Student Groups of Different Elective Replay Pattern.	31
Table 4.1	A Description of Objectives in the Curriculum and Objectives with Predictive Power Over Each From Regression Analyses.	35
Table 4.2	Statistics of the Sequence of Objectives 3rd Grade Students Played in ST Math.	38
Table 5.1	Learning Curve Plots and AFM Statistics.	53
Table 5.1	(continued).	54
Table 5.2	Learning Models Under Different Assumptions of KC Interactions.	60
Table 6.1	Categories of Teacher Activities to Orchestrate a Curriculum-Aligned Digital Game in Classrooms.	66
Table A.1	Articles of Studies on Math Educational Games.	104

LIST OF FIGURES

Figure 1.1	ST Math Content and Examples.	4
Figure 3.1	Decision Tree to Predict Whether a Student will Gain in State Standardized Math Test.	30
Figure 4.1	Objectives, the # of Objectives They Predict, and the # of Objectives that Predict Them.	41
Figure 4.2	Pairwise Predictive Relationship between Objectives.	42
Figure 5.1	An Example of a Too Difficult Level with Flat Learning Curve.	56
Figure 5.2	An Example Where Students Failed to Transfer. The Above Four Types of Puzzles Showed Four Disjointed Learning Curves.	57
Figure 5.3	Hierarchical Combinations of Game 6 Levels that Led to Models with Different BICs. BIC Similar to Baseline Indicates that the Combined Levels Share a Similar KC.	58

CHAPTER

1

INTRODUCTION

Educational games have become a mainstay of learning for our current generation. From literature reviews, decades of research have shown that educational games positively impact the outcomes of learning, and motivation and attitude towards learning [Ke09; BH13; Boy16]. Consequently, there is increasing popularity of educational games worldwide [Adk17]. For example, Ambient Insight reported that global game-based learning revenue was expected to achieve 20.2% growth, reaching \$8.1 billion in the next five years [Adk17]. A recent national survey reported that more teachers have started to adopt educational games in teaching [Fin15], with curriculum-integrated games identified as a key market demand [Ric13a; TV14]. For example, ST Math, the game we are studying, is integrated into the curriculum of 3,900 schools and used by over 1.2 million students according to its official website. As more and more educational games emerge and are utilized in schools, we face unprecedented challenges in designing and using educational games effectively to benefit our next generations.

One primary challenge is to provide efficient feedback to game designers. In practice, game designers may lack exposure to the educational context of their games, or the consequences of young children's gameplay behaviors. For example, prior work has found game designs that are misaligned with educational goals [Sch14; Har14], game mechanisms that allow shallow learning behaviors [HA15], children's unexpected gameplay behaviors [Lin11], and uneven game level progression that discourages students [Hic16]. Pinpointing such game design flaws is even more beneficial in large-scale games such as ST Math [Rut14], yet more difficult. These games usually contain hundreds of

mini-games or game levels that are intended to be played over a long period of time. Evaluating such games through traditional controlled experiments or user studies can be expensive and inefficient. As educational games grow in size and scale, we need data-driven methods to inform game design in an efficient and scalable fashion. Such data-driven methods need to be used in combination with human interpretations, to derive actionable insights given the educational content and specific context of learning.

Another challenge is how to support the use of serious games in authentic classroom settings [BH13]. Research has shown that games-based learning can be enhanced when combined with classroom instruction [Wou13; Row17; Bak15]. However, a recent literature review concluded that the crucial role of teachers has been neglected or marginalized, in both research and game-design [Mol17]. Teachers not only have limited knowledge to choose educational games, but limited time to play and prepare for the use of educational games before teaching [San06; Mol17; Che14; Mif13; ES13; Lim11; DT10]. Teachers may be unaware of how a game's content or features affect learning outcomes. This means common classroom decisions, such as which game content to skip or play, could affect learning unexpectedly. We need to use data to derive educational insights on children's gameplay and learning to empower both game designers and practitioners. These data-driven insights should be enhanced with field studies to understand the practical gaps in the design and use of educational games in classrooms.

However, there are few studies that address these challenges. The majority of math-games-research limits assessments to pre- and post-tests. Some researchers have conducted longitudinal empirical studies [Ros03; Rut10; Rut14; Sch14; Bai12a; Bak15] to prove their games' effectiveness, as shown in Appendix I. However, pre- and post-tests alone cannot provide sufficient feedback for game design if the process of gameplay and learning itself remains in a black box. On the other hand, researchers have increasingly applied data mining and analytics methods to gameplay data [Loh15]. However, most of these studies were conducted in pre-designed experimental conditions, few analyzed data from authentic classroom settings or discussed how these analytics could help the classroom use of educational games. Moreover, few studies looked into the rich contextual information of the classroom or heard from teachers. Educational games are increasing in scale and being utilized by more schools. We need to join the forces of data-driven methods and qualitative insights to derive actionable, practical, and generalizable insights to the design and use of educational games.

In this dissertation, I developed and applied data-driven methods to address the stated challenge, combined with field study of classroom use. I intend to make the below contributions:

- Pioneering research on curriculum-integrated educational math games, a type of game integrated into school activities with increasing market demand and popularity,
- Data-driven and qualitative methods and insights that inform the game design of ST Math,

which could potentially benefit over 1.2 million students currently using ST Math across the U.S.,

- Data-driven and qualitative methods and insights that are applicable to educational games and other e-learning platforms, benefiting students beyond ST Math.

Together, these contributions support a new methodology of integrating human insights with data-driven methods, combining qualitative and quantitative research to derive actionable, practical insights.

1.1 ST Math, its Uniqueness and Generality

ST Math [Rut14], designed by the MIND Research Institute, teaches K-12 math concepts through solving puzzles to help a penguin 'Jiji'. The design is centered around the following educational hypotheses: that students can gain conceptual understanding along with procedural and computational skills through intuitive spatial relationships and the exposure to multiple representations of math problems; that English language learners would struggle less in a creative and language-light environment; that new math skills are built upon previously mastered skills and students will learn better by progressing from simpler to more complicated skills at their own paces [Rut14].

ST Math includes content mapped to the Common Core curriculum and to relevant state standards and is designed as a supplement to a school's existing math curriculum. ST Math teaches math concepts through visualizations and spatial reasoning, minimizing written instructions. Every student answer leads to animated feedback. For example, a puzzle practices multiplication of 3×4 by showing 3 dogs with missing shoes. Students calculate to select the number of shoes needed; animated feedback will show shoes missing (if the answer is less than 12), that the shoes fit, or how many shoes are left (if the answer is more than 12). Students are supposed to learn and correct their answers based on the feedback, consequently developing their understanding of multiplication. Students may struggle to get the answer, but the process itself (e.g., spatial-reasoning, reflection on wrong answers) develops student skills and mindset—productive struggle is a key component of ST Math. Figure 1.1 illustrates the hierarchy of ST Math content. Table 1.1 presents the Grade 3, 2012-2013 curriculum, and the amount of associated game content.

ST Math games are structured at the top level by objectives, which are broad learning concepts. Within each objective, individual games teach more targeted concepts through presentation of puzzles, which are grouped into levels for students to play. When they first use ST Math, students start by completing a series of training games on the use of the ST Math platform and its features. Students are then guided to complete the first available objective in their grade-level curriculum, such as "Multiplication Concepts." Students can only see this objective. and must complete a pre-test before beginning the content. Once students have completed the pretest, they can start

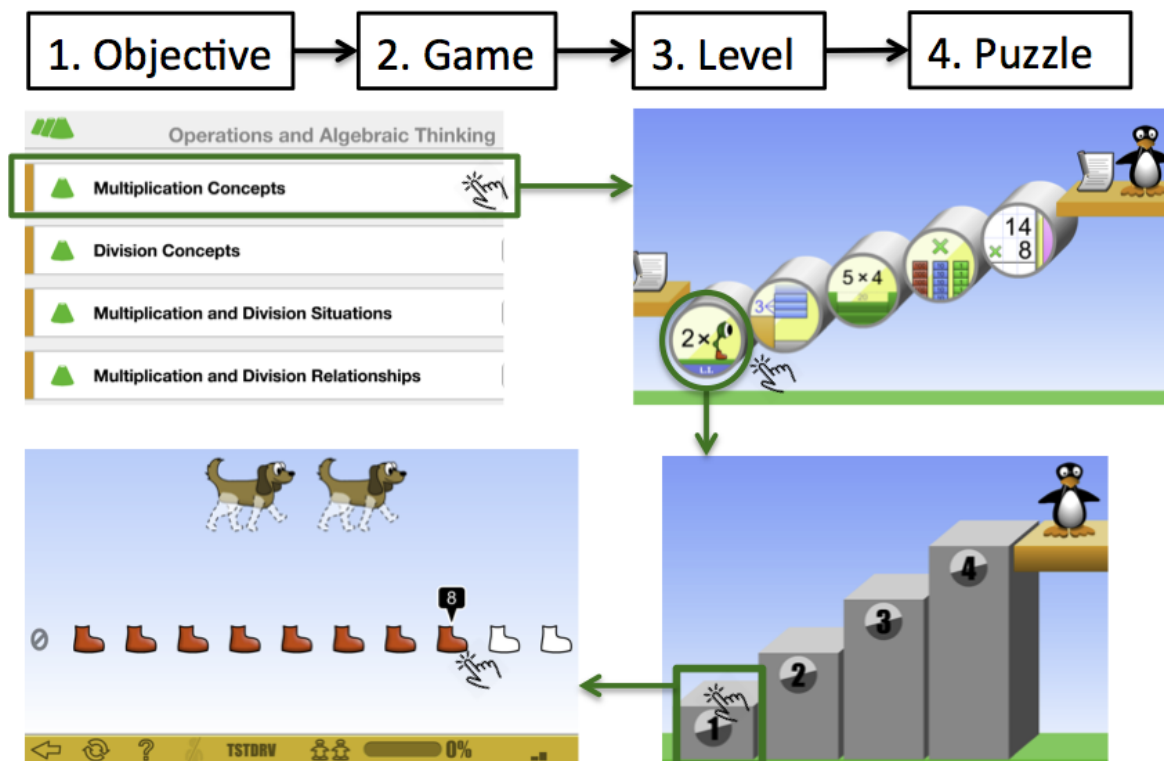


Figure 1.1 ST Math Content and Examples.

the first sub-objective (game). Games represent scenarios for problem-solving using a particular mathematical concept, such as finding the right number of boots for X animals with Y legs as shown in the puzzle picture in Figure 1.1 Each game contains one to ten levels, which follow the same general structure of the game, but with increasing difficulty.

As with many games, students are given a set number of 'lives' per level. Every time they fail to complete a puzzle correctly, they lose one life. If all of their lives for a given level are exhausted, then they will be required to re-attempt the level before they can move on to new content. Once a student has passed a level, they can elect to replay it at any time. Once a student has passed every level within an objective, then they will be permitted to take the objective-level post-test. The objective pre- and post-tests consist of five to ten multiple choice questions related to the objective, and are parallel in both the question format and the content's difficulty. Students cannot progress to the next objective until they have completed the latest objective's post-test.

One thing to note about ST Math is that the When playing a level, students do not all answer the same puzzles in the same order. MIND creates ST Math puzzles following three types of templates: randomly generated, randomly selected, and randomly ordered. To better understand the context for our data-driven studies, we report the number and percentage of levels created using each

Table 1.1 The First Six (out of 26) Objectives from Grade 3, 2012-2013 Curriculum, and the Total Number of Games, Levels, Puzzles and Lives in each Objective. NOTE: Numbers of Lives and Puzzles are Unevenly Distributed across Game Levels.

Objectives	Equal Groups	Multipli- cation Concept	Place Value Concept	Place Value Bundles	Ordering & Compar- ing Whole Number	Addition & Subtrac- tion with Regrouping
Total # of Games	7	6	5	7	11	8
Total # of Levels	31	20	15	29	38	31
Total # of Puzzles	221	138	106	202	267	186
Total # of Lives	62	40	30	58	76	93

generation technique based on the 957 ST Math puzzles played by 7,394 third grade students in our data sample. The proportions for each type are 38% randomly generated, 35.7% randomly selected, and 26.2% randomly ordered. The 364 randomly generated levels use rules to assign random attributes to create different puzzles, such as assigning 1-9 to each digit to create two digit numbers. We considered any level that had more than 20 different puzzles to be a randomly generated level. In the 342 randomly selected levels, students answer puzzles that are randomly selected from a pre-made puzzle pool. For the third-grade curriculum, we estimated that most of these levels have slightly more puzzles to choose from than the number of puzzles required to pass the level, but less than 20. On average, the pre-made puzzle pools contains 4.9 more puzzles than what are required to pass the level. However, the most frequent number of extra puzzles occurring in a pool is 1. In other words, the mode of the difference in the number of puzzles to select from and the number of puzzles needed to pass the level is just one. Thirty-three percent (33%) of the levels contain just one or two more puzzles beyond the number needed in the pre-made pool. In the 251 randomly ordered levels, all students answer the same pre-designed puzzles in a random order. Note that for a small number of levels, some puzzles follow a different template than the rest. For example, the first two puzzles of a level can be randomly ordered and the rest of the level can be randomly generated. In general, when data are analyzed in this dissertation, it is assumed that there is considerable similarity in the content that students see at each level.

In general, in the majority of levels, students play a randomized set of puzzles. Therefore, when data are analyzed in this dissertation, it is important to note that students will not be playing exactly the same puzzles, or in the same orders. ST Math levels were specifically designed to be testing the same skill, by math educators. This means that the puzzle templates should have been designed appropriately, so that students experience the same difficulty in all versions of a randomly generated puzzle. For example, we assume that additions with randomly-generated digits would be designed to have the same carry operations needed across all the generated examples.

ST Math provides functionality for the teacher to manipulate at the objective level. For example, teachers can re-ordering objectives and assign objectives as homework for the whole class. ST Math also provides feedback for teachers during and after student gameplay. During gameplay, ST Math displays information on an individual student's screen, including the remaining lives at the current level, the number of attempts conducted at the current level, a warning hurdle for too many attempts, and whether students click the 'raise hand' button. After gameplay, ST Math shows reports on a student's progress (number of levels completed in the curriculum), current objectives, pre- and post-test scores, and some other information.

ST Math provides unique research opportunities, as it is one of the few games integrated into classrooms at scale. This gives us a pioneering view on how teachers use educational games in classrooms, and informs the future design of such games. This also enables us to investigate gameplay behaviors that arise with abundant time and flexibility, such as students' optional replay of passed levels, and teachers' re-ordering of game's objectives. Moreover, ST Math is designed to be played along with the curriculum for the entire school year. Such data enable us to connect gameplay performance with the progress and transfer of learning. For example, we can see how the performance on earlier math objectives affect the later ones, and how students transfer the same math skills across different games. Lastly, ST Math allows us to investigate the practices of using educational games in classroom scenarios, and suggest game design that benefits such practical use.

Meanwhile, the data-driven insights and analytic methods derived from research on ST Math can be potentially generalized to other games and e-learning platforms. For example, ST Math's drill-and-practice mini-game mechanism is widely applied in math educational games (see Appendix A). These games will benefit from ST Math's data-driven methods and design insights, such as the application of learning curves to pinpoint game design flaws.

1.1.1 Data and Population

In the first stage of this project, we analyzed the state, student, and ST Math level data from 4,827 3rd-5th graders from California, during the school year 2012-2013. The students came from 17 schools and 221 classrooms, with their demographics information displayed in Table 1.2. These students created a total of 2,524,681 level attempts, with statistics regarding their specific gameplay behaviors described in Chapter 3.

The state-level data include student demographics information and their test scores on the state yearly standardized math test. The student-level data include student grades (year in school), curriculum (pre-designed sequences of math objectives to teach), classroom IDs, school IDs, and in-game pre- and post-tests of each math objective. Next, the level-level ST Math data record gameplay as streams of level-attempts. Each observation contains a timestamp, a student ID, the level played

Table 1.2 Populations' Demographics Information.

	Grade3	Grade4	Grade5
#Students	1567	1528	1732
Male	50.6% na:2.9%	50.1% na:2.0%	52.2% na:3.5%
Eligible for Reduced Lunch	80.7% na:2.9%	77.8% na:2.1%	81.4% na:3.2%
Hispanic or Latino	84.7% na:2.8%	82.3% na:1.9%	83.5% na:3.1%
English Language Learner	66.2% na:2.9%	56.1% na:2.1%	53.0% na:3.2%
With Listed Disability	10.9% na:2.1%	11.5% na:1.7%	11.9% na:2.8%

(with the game and objective it belongs to), and the number of puzzles correctly answered in this level attempt (before passing the level or exhausting all lives). The percentage of puzzles completed in a level attempt is defined as performance of this level attempt.

At the second stage of this project, we were provided ST Math puzzle data from 7,394 3rd grade students in a southern U.S. region. The puzzle data record gameplay at the finest granularity—students' answer to a specific puzzle. Each observation contains a timestamp, a student ID, the puzzle played (with the level, game, and objective it belongs to), the student's answer, the correct answer, and time took for the student to answer. Unfortunately, the puzzle-level data were not linked to demographics data and objective level pre- and post-tests, which limited our ability to conduct fine-grained analyses at puzzle level factoring in the characteristics of our student population, such as their math state test scores prior to playing ST Math.

Lastly, we were provided with opportunities to observe the use of ST Math in authentic classrooms. We selected eight teachers from a district in Southern U.S. teaching 3rd-4th graders. We collected field observations notes and interviews approved by IRB as described in Chapter 3.

1.2 Research Questions

This work has three overarching research questions listed below. All studies contributed to answering the overarching question 1. Studies 2-3 focused on addressing question 2, and study 4 focused on addressing question 3.

1. How can we derive practical insights on gameplay and learning, with a focus on gameplay and outcomes that arise from use over time?

2. What data-driven methods can we design to evaluate game content against educational outcomes, and to inform designs that maximize students' learning?
3. How can we understand the use of ST Math in classrooms, to inform game designs that could benefit practitioners and learners?

CHAPTER

2

LITERATURE REVIEW

This chapter is organized as follows: the first section discusses the state of art in the design and assessment of math games—the domain subject of my thesis. The second section describes the research fields Educational Data Mining and Learning Analytics, which inspired our analytic methods. The third section describes serious game analytics—the specific domain that uses data-driven analytics to improve learning in games. The last section discusses literature on the classroom use of educational games—the necessary practical considerations to inform game design.

2.1 Educational Games for Math

The last few decades have seen a surge in research and development of math educational games. In the 1990s, the majority of educational games were designed to teach social science [Ran92]. In the few studies evaluating math games, the majority found math games led to better educational outcomes compared to classroom instruction [Ran92]. Later, in the 2000s, math games took inspiration from successful commercial game styles and their platforms, such as pair-wise competition [Lee04; Con02], SimCity [Pol10], and the Nintendo platform [Ros03] to attract young learners. Meanwhile, educators placed emphasis on improving game design to support learning [Don07]. Nowadays, math educational games are increasingly popular. This section describes the contemporary trends and challenges for math educational games from the game design and assessment perspectives, and discusses why my work on ST Math can be generalized to benefit game-based learning in math.

Appendix I shows the math games from searching academic literature in the recent decades. These works came from Google Scholar, ERIC, and the NCSU Libraries' Databases, through searching keywords such as: math, educational games, serious games, and computer games. These works do not include commercial or individually-designed math games that have not been researched or published about their designs and assessments.

2.1.1 The Design

Modern math games exhibit three trends: the increasing diversity in game mechanisms to support other skills alongside math learning; the emergence of mini-games targeting a wide range of math skills; and the alignment between educational games and academic standards.

With regard to the first trend, game designers have realized educational games can provide cognitive and affective benefits beyond their targeted math skills. This has resulted in an increasing diversity of game mechanisms and features that target skills beyond math learning. For example, *Monkey Tales* [Der16] incorporates the popular drill-and-practice mechanism with increasing time pressure. The designers assumed that time pressure would enhance students' mental arithmetic skills and working memory [Der16]. Other popular features include collaboration [MM04; Bai12a; Sto11; Vru15] and visual-spatial skills [Rut14; Ven13; Der16; Kos18].

The next trend is the emergence of mini-games targeting a wide range of math skills [Ke06; Rut14; Str14; Der16; Bak15; McL17; Hie17]. Instead of practicing a specific game skill, these mini-game groups practice a set of math skills, usually from the same domain or from a set of math concepts. These mini games have many advantages, including allowing students to practice math skills under a rich variety of presentations and problem-solving scenarios. Other advantages of mini games include flexible time duration and low technology requirements [Jon09]. These advantages make it easier for educational games to be integrated into classroom activities.

The third trend is the alignment between educational games and academic standards. These games contain a large amount of educational material and are designed to be played alongside the school's curriculum over the duration of the semester. Such curriculum-aligned games have been recognized as a key market demand [Ric13a; TV14], influenced by the federal No Child Left Behind (NCLB) Act [Ric13a]. One reason is that educational games have been found to help weaker students catch up [Bai12a; McL17; Mas17; Ku14], and improve students' attitude towards math learning [McL17; Cas14; Ric13b; Bai12a; KG07; Ke06; Ke08; Ku14]. Consequently, more games are designed as supplementary curriculum resources to help schools meet standardized assessments. Examples of these games are *Astra Eagle* [Ke06], *MotionMath* [Ric13b; TJ17], *VmathLive* [Kin11], *Knowledge Battle* [Hie17], *My-Pet-My-Quest* [Che12; LIA12], and *ST Math* [Rut14].

ST Math is representative of these contemporary trends in math educational games. *ST Math* is implemented as a supplemental program to a school's existing mathematics curriculum. Each math

objective contains a set of mini-games, representing different problem-solving scenarios under a drill-and-practice mechanism. ST Math does not contain written instructions so it requires both math skills and spatial-temporal reasoning skills to solve its puzzles [Rut14]. Therefore, studying the ST Math game will provide a pioneering view on the design of future games and impact game-based learning for future generations.

2.1.2 The Assessment

Decades of research have found that math educational games are effective at improving learning outcomes as compared to traditional teaching methods alone, such as [Ran92; Ke09; BH13; Ros03; CZ04; Ke08; KG07; Cas14; Sch14; Bai12a; Ku14; Bak15; Hie17; McL17]. Many math games have led to other positive outcomes, such as improved motivation and attitudes towards math learning [McL17; Cas14; Ric13b; Bai12a; KG07; Ku14; Ke06; Ke08]. Increasingly, researchers have conducted longitudinal empirical studies to prove the effectiveness of math games as compared to traditional classroom instruction, with improved learning outcomes found in most [Ros03; Rut10; Rut14; Sch14; Bai12a; Bak15].

However, as shown in Appendix A, the majority of math game assessments were limited to pre- and post-tests and questionnaires. Some studies (17 out of 50) generated game usage and performance statistics from in-game logs to assist analysis. However, few studies (8 out of 50) applied learning analytics to delve deep into user performance or behavior patterns, or related them to math performance or math game design [Bak07; Kli11; Ven13; LIA12; Dav13; Mar15; MN17; Kou17]. Moreover, most studies lasted only a few sessions or a few weeks long. Among the 14 studies that ran over three months, six relied on pre- and post-tests before and after gameplay, which ignored the much longer process of gameplay and learning in between tests [Keb10; Bai12a; Rut10; Rut14; Sch16; Kin11]. Among the rest of the eight studies that collected data during gameplay: three applied classroom observations and user diaries to gather qualitative data [Der16; Bes17; Ros03]; five looked into games' data logs, in which four studies (three of them belongs to the same Maths Garden game) went beyond reporting basic usage statics and applied game learning analytics [Kli11; Ven13; MN17; LIA12].

New trends in math educational games have created new demands for game assessments where traditional pre- and post-tests methods fall short. With the increasing diversity of game designs, we need to better understand how a game's design influences learning outcomes [Ke09; Boy16; Mp16]. As educational games grow larger in size, scale, and intended time of play, traditional methods become costly and inefficient. For example, ST Math contains hundreds of mini-games for each grade's curriculum. Traditional methods would involve selecting students who played each of these games, setting up multiple experiments, and analyzing large amount of user study artifacts from each game. Instead, we need methods that quickly pinpoint game design flaws and take advantage of the

large amount of students already playing ST Math. We need data-driven methods that penetrate the black box between pre- and post-tests in order to better inform game design to achieve educational goals.

Moreover, penetrating this black box will significantly benefit the math domain. For example, researchers have increasingly shown that math learning is inherently hierarchical [Dun07; RS11]. Children tend to follow a certain trajectory as early skills build upon each other to form a more complex understanding of mathematics [CS14]. Data collected during the learning process would help us understand how such math learning occurs in the game context, and how to design games that support such learning trajectories [Con17b; Con17a]. Another example is that certain math skills, such as fractions, are difficult for 3rd graders, which calls for research on how to teach them better [MH08; RJ09; Han17; Jor17]. Educational games, especially those applied at large scale such as ST Math, collect rich data that would yield valuable insights on how students develop understanding of certain math skills. These insights can not only inform game design, but can benefit the broader field of math education.

However, data-driven methods alone are not enough. Previous work [Ke08; Lin11; Der16; TJ17; Bes17; LJ17; Nan18] has shown that qualitative methods compliment game learning analytics by providing rich, contextual information on gameplay that can lead to valuable game design insights. For example, Lindstrom et al. [Lin11] conducted a nine-week-long field observation of students playing a place-value math game. By observing students' gameplay, the authors identified possible mismatches between pedagogical principles intended by the designer and the actual use by students. Because ST Math is a curriculum-integrated game commonly used in school, applying field studies and qualitative methods to understand the use of the game in authentic classroom scenarios is of crucial importance. Qualitative methods would help interpret data-driven insights, which will guard against the spurious results that could arise from the sheer amount of data with no empirical meaning or practical impacts. Additionally, qualitative methods empower teachers to share their perspectives, which can lead to game design principles that could empower practitioners.

2.2 Educational Data Mining and Learning Analytics

Educational Data Mining [RV13] (EDM) and Learning Analytics [SB12] (LAK) are new fields with a common goal to improve and understand learning through data. EDM has greater focus on automatic discovery, and LAK puts more emphasis on informing and empowering instructors and learners [SB12; BI14].

My work in ST Math has a primary goal of deriving data-driven methods and insights to inform game design. On one hand, we need to provide interpretable insights and actionable recommendations for game designers and teachers. On the other hand, with the large number of game content in ST Math, we need to data-driven methods to do so in an efficient and scalable fashion. To achieve

these, we need insights from both EDM and LAK communities. The following subsections describe exemplary work that relates to my application.

2.2.1 Mining the Relationships in Educational Content

There has been a growing interest in using data-driven approaches to inform curricular design, especially the sequencing of educational content.

At the skill granularity, researchers have informed educational content sequencing through mining prerequisites between skills. Chen et al. [Che16] applied Structural Expectation Maximization to learn the optimal Bayesian network structures that represent skill prerequisites using students' performance from textbook exercise for an English Proficiency test. Chen et al. [Che15] applied Association Rule Learning on the same educational context, which assumed that one skill is a prerequisite for another if the probabilities of mastering skill A given that B is mastered, or not mastering skill A given that B is not mastered, were both above a certain threshold. Brunskill [Bru11] used Expectation Maximization with the conditional probabilities of observing certain performances on test questions, which are designed to map each question to a single skill, to discover prerequisite relationships between linear inequality math skills. This mapping process also faces the risk of expert blind-spots—pedagogical organization based on “the structure of the domain rather than the learning needs of novices” [Nat01]. Moreover, these methods require a significant amount of effort to map educational content to specific skills, which can be extremely costly in large-scale games such as ST Math.

At the problem set granularity, Doroudi et al. [Dor16] assigned each student's activity trajectory to violations of sequencing constraints, and compared how different sequencing constraints affected performance. However, it is difficult to design similar constraints for curricula with longer time spans and more varied content. Desmarais et al. [Des06] designed a partial order knowledge structure algorithm (POKS) that applies conditional independence tests to search for links between question items. This method focuses on discovering hierarchies between sets of pre-ordered test questions, whereas in ST Math, students completed objectives in a variety of orders over time. Our focus is on understanding the consequences of specific orderings to derive actionable feedback. Vuong et al. [Vuo11] applied binominal tests in all possible pair-wise relationships among curriculum units in an intelligent tutor. The binominal test compared the performance on unit A between students who mastered and had not mastered unit B, and deemed unit B as a prerequisite of unit A when a statistical significance was found in the test. Although this work created a heatmap describing the existence and strength of relationships, it lacked analysis of the meaning and implications of these relationships. Methods at this granularity are applicable to ST Math data and could help us investigate relationships between math objectives as groups of math puzzles under the same concept. However, we must address the challenges in the serious game domain and ensure our

results are interpretable and can derive actionable insights for game design and usage.

At the course granularity, Ochoa, Yang et al. and Moretti [Och16; Yan15; Mor14] applied data-driven methods on high-level data such as enrollment, web-mined course descriptions, and online ratings to inform the selection and sequencing of college courses. For example, Yang et al. [Yan15] projected courses to concepts (skills) extracted from online course descriptions, used pre-existing course prerequisites to optimize the relationship between concepts and predicted course prerequisites of another university. However, these methods are not applicable to ST Math. ST Math does not have enough verbal descriptions on games' math concepts to apply these methods, as it is designed to be a language-light environment that teaches math through visual-spatial reasoning. At the system granularity, Pechenizkiy et al. and Pechenizkiy and Toledo [PT14] designed conceptual frameworks for the cyclic process of curriculum mining and designing.

However, little work has applied such data-driven methods in serious games. In Chapter 4, I present a method that can inform the sequencing of ST Math objectives at the problem-set granularity.

2.2.2 Constructing Learning Curves

Simply finding correlations or predictive relationships among games is not enough to inform content design within games, because within an objective, games are played following a defined order in ST Math. However, games represent various problem-solving scenarios under the same math topic, which contain levels of progressing difficulties. Investigating how levels and games maps to the fine-grained skill can yield insights on how students transfer and improve performance across problem-solving scenarios. Such insights can pinpoint where game design can better support student learning.

One common approach to mapping items to skills is constructing Q-matrices or skill matrices [Bar05; DN13; Lan14; Cas17]. In a Q-matrix, each row represents a problem item and each column represents a skill. A number one in row i column j means that item i contains skill j . The product of the Q-matrix and students' 'knowledge state' (e.g., a matrix representing if a student i knows a skill j) results in students' predicted responses on problem items. To derive Q matrices, previous researchers have used different Matrix Factorization methods [DN13; Cas17], hill-climbing machine learning algorithms [Bar05], and maximum likelihood estimation [Lan14] to estimate both the Q-matrix and students' knowledge state simultaneously from student responses. However, one constraint of such methods is that they assume that the student 'knowledge state' is set prior to answering problem items and will not change during problem answering. Many methods related to item-response theory make this same assumption when discovering relationships between items (e.g., [Des06; Vuo11; Che15; Che16]). In ST Math, we need to map game content to skills through modeling the dynamic growth of a 'knowledge state' through its drill-and-practice mechanism.

Moreover, many derived Q-matrices aim to increase the accuracy of predicting student responses [Lan14; DN13], rather than producing interpretable relationships. To inform the content design of ST Math, we need methods that yield interpretable results.

Learning curve analysis is an interpretable method that models knowledge growth through connecting content to skill. Learning Curves are derived from the cognitive theory of Newell and Rosebloom [NR81]. This theory assumes that with more practice, a students' accuracy at answering a question improve following a logarithmic curve. In other words, a good learning curve shows that students' accuracy increases, but the amount of this increase gets smaller over time. After enough practice, the increase will be negligible, and students can be considered to have reached their best performance. Learning curves are great choices for modeling data in games with a drill-and-practice mechanism, as students practice the same skills consecutively.

Several learning models have been applied to fit learning curves. The Additive Factors Model (AFM) [Cen06] is a logistic regression model that assumes the probability of correctly answering a question depends on individual students' parameters, the skill difficulties, and the number of previous practice opportunities on the skills. For questions containing multiple skills, the difficulty and practice attempts of these skills are summed together. This assumes that a student can correctly answer a question without knowing all of the skills involved, provided the summation of known skills passes a certain threshold. The Conjunctive Factor Model (CFM) [Cen08] is similar to AFM, but assumes that the difficulties and practice of skills are multiplied together for questions with multiple skills. This means that a student can never answer a question correctly unless they know all the required skills. Another method is Performance Factor Analysis (PFA) [Pav15], which is similar to AFM with an additional assumption that success and failure have different impacts on learning. Because we don't currently have evidence that this additional assumption of PFA holds in ST Math, we focused on applying AFM and CFM in this paper.

Despite the wide application of empirical Learning Curves in intelligent tutors and other e-learning platforms, there has been little application in serious games. As an exception, Harpstead and Alevan [HA15] applied the AFM model in a physics game. Through examining learning curves, they identified an unforeseen shortcut strategy with which students could pass the game without sufficiently mastering its underlying math concepts. Similarly, Baker et al. [Bak07] fit learning curves in a action-based math game to model gains in speed and accuracy over time. They found that modeling accuracy helped to identify skills that needed extra support and scaffolding. However, modeling speed (time taken to answer a question) was not as helpful, because it was hard to separate the increases in speed due to increased math fluency from those that were due to the increased familiarity with the gameplay. Lomas et al. [Lom13] applied learning curves to a game locating numbers on a number line. They found when students were allowed to pass the game with less accurate estimations, the learning rate was lower.

In these cases, estimating learning curves revealed areas for modification and improvement

within serious games. However, in all previous work, the games studied represented a single problem-solving scenario. In contrast, games such as ST Math practice math skills through various problem-solving scenarios. Applying learning curves to ST Math can yield not only game design insights, but can help understand how students transfer knowledge across different problem-solving scenarios.

2.3 Game Learning Analytics

Serious games are games designed for a primary purpose other than entertainment [Loh15]. Similarly, Game Learning Analytics [Fre16] (also called Serious Game Analytics [Loh15]) grounds game design in students' learning and performance on targeted skills, whereas traditional game analytics primarily focuses on player enjoyment [Loh15]. Game learning analytics connects data during gameplay, such as user actions, choices, and performances, with assessments outside gameplay, such as learning gain and psychological responses. This approach penetrates the black box of what happened in between the traditional pre- and post-tests and converts user-generated data into actionable game-design insights. Game learning analytics also shares similar and potentially transferable methods with Learning Analytics [SB12] and Educational Data Mining [RV13]. However, Game learning analytics is a more specific domain, whose nature of play and learning can be inherently different from other e-learning environments. We need more research to verify the applicability of these methods across domains, and to adapt these methods to solve contemporary challenges in educational games.

Game learning analytics connects game design to gameplay behaviors and learning. This provides feedback for game designers to better achieve their educational goals. In another work, Harpstead et al. [Har14] clustered data from 174 students to investigate the alignment between the game's feedback (how it reacts to the student gameplay artifacts), and the game's educational goals (physics principles of stability). This work suggested redesigns for more consistent and immediate feedback. Hicks et al. [Hic16] applied survival analysis and interaction networks [Eag12] to evaluate the gameplay from 433 students on 30 levels of a physics game. This work pinpointed problem spots where students frequently dropped out due to a rough progression of levels' difficulties. ST Math contains dozens of math objectives and hundreds of mini games for each grade's curriculum. Thus, applying game learning analytics would help pinpoint problematic content efficiently and yield valuable feedback to game designers to improve gameplay and learning.

Game learning analytics can help us better understand learning and provide educational insights across domains. For example, Martin et al. [Mar15] investigated fraction learning through a laser-splitting game. This work visualized and clustered game board states and found that students improved in fraction learning through splitting, especially students who broadly explored splitting in the game's mathematical space. Snow et al. [Sno15] measured the consistency of students' choices in a game. This study showed that students with more controlled and determin-

istic patterns in their gameplay choices demonstrated higher target skill acquisition. Bauer et al. [Bau17] analyzed problem-solving behaviors through visualizing and quantifying solution spaces in a science-discovery game. They found high-performing problem solvers explored more hypotheses, explored more broadly in solution spaces, and engaged less in greedy optimization for reaching local optima. Similarly, Kang et al. [Kan17a] analyzed problem-solving through applying sequential data mining in a problem-solving simulation game. This work identified stages of problem-solving, and that high-performing students used in-game tools more strategically to break complex problems into sub-goals. These examples show that game learning analytics could contribute to our general understanding of learning. Because ST Math is a curriculum-integrated game, applying game learning analytics on ST Math would yield valuable insights that improve the learning and teaching of standardized math content.

Additionally, game learning analytics has often been combined with Machine Learning to track and support learning in real time. One of the earliest works by Conati and Zhao [CZ04], and a more recent work by Davoodi and Conati [DC13] used dynamic Bayesian networks to connect students' gameplay behaviors with the probabilities of mastering target skills. Through real-time knowledge tracing, this work provided customized in-game hints. Some resulting hints were found to be significantly correlated with learning gain. Min et al. [Min17] applied long short-term memory neural networks to a science simulation game. This work predicted students' competency levels through students' in-game actions and external pre-learning measurements. Rowe et al. [Row17] used data-mined learning strategy detectors and automated coding of interaction networks to measure implicit learning in a physics game, showing that in-game measures of implicit learning significantly improved the prediction of post-test scores. However, many of these methods were designed for specific game environments, and focus more on automation than interpretability. ST Math data contain much simpler gameplay interactions, with my goal being understanding of game use and providing actionable feedback.

2.4 Educational Games in the Classroom

One primary goal of my thesis is to suggest practical, actionable feedback to support the use of ST Math in authentic classrooms. To do so, it is necessary to understand the classroom uses of ST Math and consider ST Math from the practitioner's perspective.

Our approach is informed by prior research that argues that teachers should conduct a wide range of activities before, during, and after gameplay. For example, through their review of 35 studies of digital and non-digital games in classrooms, Kangas et al. concluded that there are five phases of teacher usage that should be intentionally supported [Kan17b]. These phases are: 1) *planning*: creating a pedagogical frame for game-based learning such as situating game in class or organizing student groups; 2) *orientation*: introducing the game and background concepts before

play; 3) *playing*: tutoring students' gameplay; 4) *elaborating*: leading discussion and reflection after gameplay; and 5) *reflecting*: teacher's self-reflection process to develop teaching practices.

A limited number of studies have investigated contextual practices surrounding the use of specific educational digital games in classrooms. In a case study of one teacher, Watson et al. identified three key teacher strategies: learning by playing the game; setting student goals for gameplay and encouraging reflection; and identifying teachable moments during gameplay [Wat11]. From three teacher case-studies, Eastwood and Sadler found that teachers viewed and used curriculum-integrated materials differently, demonstrating the importance of flexible and adaptive design [ES13]. Nanavati et al. found that games caused some shifts of authority from teachers to students, and identified disconnects between how games were intended to be used and how they were used in context [Nan18]. Callaghan et al. interviewed and surveyed ST Math teachers and reported that they needed most help with 1) assisting struggling students 2) being aware of the game's purpose and the underlying mathematics in later game levels, and 3) better aligning games with curriculum [Cal18].

Controlled experiments have suggested that digital games are more effective when combined with structured classroom activities, such as debriefing and bridging. In a longitudinal study, Bakker et al. found that playing at home with debriefing at school was the most effective, followed by integrating games into lessons at school, as compared to playing only at home and having no game at all [Bak15]. Similarly, Rowe et al. found students learned better from a game when teacher bridged game-based learning to classroom learning using game examples and discussion [Row17]. In a study of games and student choice in classrooms, Barendregt and Bekker found that having teachers provide students with choices from a small sample of activities promoted higher student interest than students' free choice, or teachers rigidly defining activities [BB11]. However, all of these studies of digital games focused on evaluating pre-designed conditions instead of understanding authentic classroom and teacher practices.

These studies have shown that teachers can use varied practices used to make game-based learning work. In Chapter 6, we build on and extend this previous work to explore the use of a single game (ST Math) across several contexts. We provide rich information about the teachers and their classrooms through an integrated grounded theory approach, combining field observations with semi-structured interviews, discussing our findings from a game-design perspective.

2.5 Summary

To summarize, prior literature has shown that:

- ST Math is representative of the contemporary trends in math educational games. Studying ST Math can yield pioneering insights for the design and use of math games for learning.
- The majority of assessments of math games are limited to pre- and post-tests, which fall

short as games grow in size and in their intended time of usage. Applying game learning analytics can penetrate the black box between pre- and post-tests and significantly improve our understanding of game design and learning, which will be extremely beneficial to the math domain.

- Advances in Educational Data Mining and Learning Analytics have opened opportunities to take game learning analytics to a higher level. Designing such data-driven methods would inform the design and use of educational games to maximize learning outcomes in an efficient fashion.
- Teachers face difficulties integrating educational games effectively with classroom teaching. Combining data-driven insights and qualitative research could inform game design to maximize the practical impact of educational games in classroom settings.

The following studies are organized as follows. Study 1 investigates student's replay behavior; it is also the first study aim to understand ST Math data and environment. Study 2 and 3 design data-driven methods to generate both game design and educational insights. Study 4 applies qualitative approach to understand the empirical context of using ST Math, which helps interpreting our data-driven insights in practical scenarios. These combined studies derive practical, actionable game design insights from studying ST Math with both quantitative and qualitative approaches.

CHAPTER

3

STUDY 1: IS STUDENT-INITIATED GAME REPLAY WORTH THE EFFORT?

Liu, Z., Cody, C., Barnes, T., Lynch, C., and Rutherford, T. (2017). The antecedents of and associations with elective replay in an educational game: is replay worth it. In Proceedings of the 10th International Conference on Educational Data Mining (EDM). Full Paper. [Liu17]

Abstract. Replayability has long been touted as a benefit of educational games. However, little research has measured its impact on learning, or investigated when students choose to replay prior content. In this study, we analyzed data on a sample of 4,827 3rd-5th graders from ST Math, a game-based educational platform integrated into classroom instruction in over 3,000 classrooms across the U.S. We identified features that describe elective replays relative to prior gameplay performance, and associated elective replays with in-game accuracy, confidence, and general math ability assessments outside of the games. We found some elective replay patterns were associated with learning, whereas others indicated that students were struggling in the current educational content. We suggest, therefore, that educational games should use elective replay behaviors to target interventions according to when and whether replay is helpful for learning.

3.1 Introduction

This work analyzed gameplay logs from a series of math games within the year-long supplemental digital mathematics curriculum Spatial Temporal (ST) Math. We analyzed gameplay data from 4,827 3rd-5th graders throughout the 2012-2013 school year. Our data contained 37,452 logged elective replays, accounting for 1.48% of the logged play. We analyzed gameplay and elective replay features in association with students' demographic information, in-game math objective tests, and the state standardized math test. We sought to answer three research questions: Q1: What are the characteristics of students who engage in elective replay, Q2: What gets replayed, and under what circumstances? And Q3: Is elective replay associated with improvements in students' accuracy on math objectives, confidence, and general math ability?

3.2 Background

“Replayability is an important component of successful games.” [Pre05] In most games, there are two types of plays: play and replay to pass a level (pass attempts) and replay after passing a level (elective replay). In this study, we investigate the latter. Elective replay (ER) is particularly interesting because the motivations behind a student's decision to replay and the impact of those replays are relatively unknown. We explore potential associations between elective replay and student characteristics and performance in the domain of educational games.

Replayability has been touted as a benefit of educational games [Gee07]. Replayability encourages players to engage in repeated judgement-behavior-feedback loops, where users make decisions based on the situation and/or feedback, act on those decisions, and receive feedback based on their actions [Tho04]. In the RETAIN model designed by Gunter et al. [Gun08] to evaluate educational games, replayability is a criteria for naturalization — an important component in helping students make their knowledge automatic, reducing the cognitive load of low-level details to allow for higher order thinking. In the RETAIN model, “replay is encouraged to assist in retention and to remediate shortcomings.” [Gun08] Meaningful elective replay is often encouraged by game features such as score leaderboards, which inspire students to replay for higher scores [Boy11]. Because higher scores typically require a deeper understanding of the educational content in a well-designed game, encouraging elective replay may promote mastery. Games with replay also allow the student to be exposed to more material and give them more freedom to control their learning. Studies have shown that giving students control over their learning process can increase motivation, engagement, and performance [Cal05; CL96].

Despite the believed benefits of replayability [Gee07; Tho04; Gun08; Boy11], few studies have investigated the educational impact of elective replay. Boyce et al. [Boy11] evaluated the effects of game elements that were designed to motivate gameplay and elective replay. These included a

leaderboard that shows each student's rank based upon their score, a tool for creating custom puzzles, and a social system for messaging among players. The experimental design required students to play the game in one session, and to replay the game as more features were added in the subsequent sessions. The study found a sharp increase in test scores as these features were added to the game. The authors concluded that features designed to increase replayability can increase learning gains. However, this result may be due to increased time on task as the same group replayed the base game with new features. In another study, Clark et al. [Cla11] analyzed logged student-initiated elective replay in a digital game. They found that frequency of elective replay did not correlate with learning gains, prior gaming habits/experience, or how much students liked the game. They also found that, although there was no statistically significant difference between the male and female students, males replayed more than the females. This may have been responsible for their slightly higher, although not statistically significant, "best level scores"—the highest score received on each level.

However, more research is needed to understand the potential educational impact of replay in educational games, particularly elective replays initiated solely by the players. One reason for the lack of such research is that educational game studies are often comparatively brief, so replayability is often minimally assessed with post-game questionnaires asking about students' intention for future play [Pla13; Bur15]. Consequently, there is a need to investigate elective replay with actual logged actions in a game setting where students have sufficient time and freedom to replay.

3.3 Methods

3.3.1 Data & Features

MIND Research Institute (MIND), the developers of ST-Math, collected and provided to the researchers gameplay data from 4,827 3rd-5th graders during the school year 2012-2013. These students came from 17 schools and 221 classrooms. Their demographics information is displayed in Table 1.2.

This gameplay data includes pre- and post-tests for each objective and the number of level attempts. For each pre- and post-test, ST Math logged students' accuracy and self-reported confidence level (1 for 'high' and 0 for 'low') for each question. For each play at a level, ST Math logged the student's ID, timestamp, and the number of puzzles completed. From these data, we identified ER as plays made after a student initially passed the level. We found ERs in 89.6% of all objectives in ST Math, accounting for 1.48% of all level attempts. Among 4,827 students, 59.85% ERed at least one level, with an average of 7.84 levels (SD=12.99, 95% CI [7.37, 8.32]) across 3.06 average objectives replayed per student.

We created features at three different levels of granularity (from finest to largest): level, objective, and student. For the level granularity, we treated each unique student-level combination as an

observation. We calculated the features by averaging all gameplay for a specific student at a specific level. For objective granularity, each unique student-objective combination was treated as a single observation. Features were created by averaging across all levels played by a specific student within a single objective. The objective granularity also included the objective pre- and post-test accuracy and confidence. For the student granularity, we treated each student as a single observation. We calculated the features by averaging across all objectives played by a student over the entire year. The student granularity also included student demographic data and state standardized math test scores. These granularities ensured that our analysis did not favor units with the majority of data logs. Each student was considered equally in our analysis, regardless of how many objectives they played. Our data contained 4,827 students and 2,524,681 plays, which yielded 1,462,660 student-level observations, and 74,985 student-objective observations.

Table 3.1 shows five example plays of “Division-Level3,” including four pass attempts and one ER of this level, interspersed with ERs from other levels. We consider consecutive ERs as an ER Session, as these ERs are circumstanced on the same pass attempts.

Table 3.1 Example of Elective Replay and Pass Attempts.

Play	Objective-Level	Passed?	Play Type
1	Division- Level3	No	Pass Attempt
2	Division- Level3	No	Pass Attempt
3	Division-Level1	Yes	ER (ER Session1)
4	Division- Level3	No	Pass Attempt
5	Division-Level1	Yes	ER (ER Session2)
6	Division- Level3	Yes	Pass Attempt
7	Division- Level3	Yes	ER (ER Session3)
8	Subtraction-Level1	No	ER (ER Session3)

3.3.2 Pass Attempt Features

We defined performance to be the percentage of puzzles a student completed before losing all lives on the level. Pass attempts are plays prior to ER, where we assumed students play with the intention of passing the level. Pass attempt features included: performance when a student first attempted a level (1st pass attempt performance), number of attempts taken to pass a level (# pass attempts), and average performance of all pass attempts (average pass attempt performance). At the student granularity, students took an average of 1.91 (sd=0.89) attempts to pass each level, with average performance of 0.80 (sd=0.10) on the first pass attempt, and 0.87 (sd=0.07) on all pass attempts (indicating overall improved performance on later attempts).

3.3.3 Elective Replay Features

Table 3.2 shows ER features that describe ER from three angles: (I) the frequencies of ER, (II) the performance of ER, and (III) the circumstances of ER in terms of the ER's prior plays. To summarize, the majority of ERs had higher performance than their levels' first attempt, and resulted in another pass of their levels. Levels that were ERed had similar performance compared to levels that weren't ERed, but levels that were followed (54.65%) or interrupted (54.35%) by ER had much lower performance than those that weren't followed or interrupted by ER. Among all ERs, most ERs' immediately prior pass attempts were from different levels or objectives. There were few instances (9.80%) where students passed a level and immediately ERed it following the pass.

3.3.4 Student Grouping from ER Features

We created student groups to encapsulate the circumstances under which ER occurred, based on students' majority ER and ER sessions. Based on prior literature, we hypothesized that ER is a habitual behavior that arises from individual needs, such as gaining higher scores [Bar96], avoiding progress on the current task [Mos02], or taking a mental break from negative emotions [Sab13]. Thus, grouping students based upon the circumstances of replay based on their majority behaviors provides high level profiles to investigate characteristics of students who engaged in ER and benefited from ER.

We characterized ER by the timing relative to the student's current learning objectives and gameplay. The first grouping describes whether the majority ER sessions started before (Group B) or after (Group A) passing the previous attempted level (current learning objective). If there is a tie between the two types of replay session, the student belongs to neither group. For example, Table 3.1 describes a group B student, who has two replay sessions before passing "Division-level3," and one replay session after passing this level but before moving on to the next level.

The second grouping describes whether an ER followed plays on the same level (SL), a different level under the same objective (DLSO), or a different objective (DO). For our example in Table 3.1, the student's pass attempts on "Division-Level3" was interrupted twice, by replays on "Division-level1" (DLSO). After passing "Division-level3," the student replayed the same level (SL) once during the seventh play, and a different objective "Subtraction-level1" (DO) once during the eighth play. This Group B student had two DLSO replays: one SL, and one DO. Thus, this student also belongs to Group DLSO, because the two groupings are independent of each other.

Table 3.2 Elective Replay (ER) Features and Their Descriptive Statistics Among Students who Electively Replayed, Collapsed to the Student Granularity.

ER Features	Descriptive Stats
I. Frequencies of ER	
% ER out of all plays	M=2.40%, SD=4.26%
% Objectives that have been electively replayed	M=22.94%, SD=20.89%
% Objectives whose pass attempts were interrupted/followed by ER	M=19.48%, SD=17.57%
II. Performance of ER	
Performance of ER	M=0.71, SD=0.28
% ERs performed better than the level's first attempt	M=71.96%, SD=31.44%
% ERs that result in another pass of the level	M=60.36%, SD=35.51%
III. Circumstances of ER	
The Replayed Level e.g., "Division-lvl1," "Division-lvl3," and "Subtraction-lvl1" in Table 3.1	
Pass Attempts Features	M=0.79, 1.98, 0.87 for 1st performance, #pass attempts, and avg performance
The Immediately-Prior play of the ER e.g., Play 2 is the immediately-prior play of play 3 in Table 3.1	
Performance on the immediately-prior play	M=0.63, SD=0.29
% ERs whose immediately-prior plays is also an ER	M=0.31, SD=0.28
% ER whose immediately-prior pass attempt is on the same level	M=9.80%, SD=23.84%
% on a different level in the same objective	M=40.75%, SD=39.09%
% on a different objective	M=49.44%, SD=40.76%
The Immediately-Prior Pass Attempts followed or interrupted by ER and ER Session e.g., "Division-lvl3" for all ER Sessions in Table 3.1	
Pass Attempts Features	M=0.51, 3.62, 0.55 for 1st performance, #pass attempts, and avg performance
% ER sessions whose prior pass attempt passed the level	M=45.65%, SD=40.69%

Note. statistics are reported at the student granularity, which are calculated through averaging across all objectives played by a student, and then averaged across all students who electively replayed. This means each student contributes equally to the average, regardless of how many objectives s/he played.

3.4 Results & Discussion

3.4.1 Who Engaged in Elective Replay?

We first investigated the demographic characteristics of students who engaged in elective replay. We found that males did so more often than females (male: 63.2%, female: 57.0%, $\chi^2(1, N=4827) = 17.99$, $p < .001$). We also found that English Language Learners (ELL) did so more often than their non-ELL

peers (ELL: 62.3%, non-ELL: 57.1%, $c2(1, N=4827) = 12.69, p < .001$), as did students with reported disabilities (disability: 68.7%, non disability: 59.1%, $c2(1, N = 4827) = 18.17, p < .001$). There were no statistically significant differences in the frequencies of ER based on race when operationalized as Hispanic/non Hispanic, or based on free/reduced lunch eligibility. The frequency of ER was not found to be correlated with other out-of-game student factors, such as state standardized math test scores.

The frequency of ER was also not correlated with in-game pre-test accuracy and confidence at the objective granularity. Next, we investigated the gameplay characteristics of students who electively replayed. We first separated students into groups based on their replay patterns. The first 5 columns of Table 3.3 shows the results of Mann-Whitney U tests with Benjamini-Hochberg correction to compare each group in-game performance to the students who never electively replayed any levels (the Base group). The last column compares the averaged ER performance of each group to the rest of students who electively replayed.

Table 3.3 Mann-Whitney U Tests Comparing Gameplay Characteristics Between Student Groups of Different Elective Replay Patterns.

Group (# students)	Pre-test Accuracy	Pre-test Confidence	Avg Pass Attempts' Performance	Avg 1st Attempt Performance	#Pass Attempts	ER Performance
Base: No ER (N=1938)	M=0.61 SD=0.17	M=0.75 SD=0.23	M=0.88 SD=0.08	M=0.81 SD=0.11	M=1.82 SD=0.84	NA
ER (N=2889)	*M=0.57 SD=0.17	M=0.74 SD=0.24	*M=0.87 SD=0.07	*M=0.80 SD=0.10	*M=1.92 SD=0.78	M=0.72 SD=0.29
Group A (N=1114)	M=0.62 SD=0.16	M=0.77 SD=0.22	*M=0.90 SD=0.05	*M=0.84 SD=0.08	*M=1.62 SD=0.52	*M=0.77 SD=0.27
Group B (N=1464)	*M=0.52 SD=0.17	*M=0.72 SD=0.25	*M=0.84 SD=0.07	*M=0.75 SD=0.09	*M=2.28 SD=1.09	*M=0.67 SD=0.29
Group SL (N=173)	M=0.61 SD=0.17	M=0.75 SD=0.23	M=0.88 SD=0.07	M=0.81 SD=0.09	M=1.82 SD=0.81	*M=0.84 SD=0.29
Group DLSO (N=983)	*M=0.54 SD=0.18	M=0.73 SD=0.24	*M=0.84 SD=0.08	*M=0.76 SD=0.10	*M=2.27 SD=1.16	*M=0.67 SD=0.32
Group DO (N=1399)	*M=0.58 SD=0.16	M=0.75 SD=0.23	M=0.88 SD=0.06	M=0.81 SD=0.08	M=1.80 SD=0.71	M=0.73 SD=0.26

Note. 1) Green and red indicate statistical significance levels higher and lower than the base class, with $*p < .001, +p < .01$ 2) Group A, B: most ER sessions happened before (B), after (A) passing the prior non-replay level. Group SL, DLSO, DO: most ER followed pass attempts on the same level(SL), different level in same objective(DLSO), or different objective (DO)

Compared to the base group, students for whom most replays happened before passing the prior non-replay level (Group B) and students for whom most replays followed a different level on the same objective (Group DLSO) started with significantly lower pre-test scores and did worse in gameplay, as measured by the three pass attempt features described in section 3.3.2. For example, students in Group B started with lower accuracy and confidence at pre-test, took an average 0.5 more attempts to pass a level, and had lower performance on the 1st pass attempt and all pass attempts (including the 1st). It seems that Group B students who replayed earlier levels before passing the current one had less prior knowledge and struggled more in the game. By contrast, students in Group A, for whom most replay happened after passing the current level, did slightly better in gameplay compared to students who never electively replayed (the Base group). Because these students started with pre-test scores that were not statistically significantly different from the base group, their replay patterns are associated with higher gameplay performance.

3.4.2 What Gets Replayed, and When?

Next, we studied what levels get replayed, and under what circumstances. We used a decision tree classifier which allowed us to identify which factors are most important in relation to ER. Our goal was not to find precise predictive models, but to augment our understanding of performance and its relationship to ER. We used R's rpart package with parameters `minsplit=5%` and `cp=0.02` to build trees to classify levels that were replayed from levels that were not replayed, and levels whose pass attempts were interrupted or followed by replay from levels that were not interrupted or followed by replay. We randomly undersampled the majority class (levels without replay, levels not interrupted or followed by replay), so that each class represented half of the observations. We used pass attempt features at the level granularity together with pre-test results, objective, and demographic information to build our tree. We used 10-fold cross validation to assess the accuracy of the learned trees.

Table 3.4 reports the derived decision trees and the importance of the features. We found that a student's performance on a particular level influenced whether replay happened during/after the level's pass attempts. For example, a student was more likely to replay a different level under the same objective (DLSO) if they took more than two attempts to pass the current level. This result is related to the previous result in Table 3.3, showing that, at the student level, those with lower gameplay performance were more likely to replay another level under the same objective.

On the other hand, the objective to which a level belongs influences whether or not a level would be ERed. We built trees to predict if a level is replayed following the same level (same condition of the last row in Table 3.4, $N=1,776$), the same objective but a different level ($N=12,616$), or a different objective ($N=31,852$). For all three conditions, the trees only contains a single node - objective, with accuracy of 55.2%, 62.0%, and 66.9% respectively. This ER decision could have been influenced by

Table 3.4 Decision Trees to Predict Levels whose Pass Attempts were Interrupted or Followed by Elective Replay.

Condition: interrupted or followed by	Trees
ER from a different level in the same objective (N=8,094)	77.8% accuracy #pass attempts < 2.5, No #pass attempts ≥ 2.5, Yes
ER from a different objective (N=12,506)	78.7% accuracy 1st attempt performance ≥ 0.94 —objective group A, No —objective group B, Yes 1st attempt performance < 0.94 —objective group A —# pass attempts < 6.5, No —# pass attempts ≥ 6.5, Yes —objective group B, Yes
ER on the same level (N=1,766)	55.2% accuracy objective group A, No objective group B, Yes

Note. Trees are presented in text format. For example, the first tree shows that if a student passed a level with less than 2.5 pass attempts, the tree predicts this student will not replay another level during/after this level.

either the content or timing of the objectives. In our tree node, we noticed that many objectives with a higher chance of ER occurred earlier in the curriculum, this could be because students had more time in which these objectives were available for ER. Our tree model also had only 55.2% accuracy when predicting whether a level would be ERed following the pass attempts of itself. One explanation is that we do not have puzzle granularity data on how many lives a student actually lost. From prior literature [Boy11] [Cla11], students may replay the same level following it pass attempts to get a better score, which means losing fewer lives (making fewer errors) at a level. As shown in Table 3.3, Group SL students who performed most of their ERs after the same level also achieved the highest ER performance.

3.4.3 Is Elective Replay Associated with Gains?

In this section we will address our second research question. As part of our analysis we considered three gain scores: accuracy gain, confidence gain, and math gain. The first two were measured by in-game pre- and post-tests. Recall that both before and after a student attempts an objective, ST Math logs the students' correctness and confidence scores on each question on the pre- and

post-tests. We averaged these scores across the pre- and post-test questions to compute the first two gain scores. These were assessed at the objective granularity. Math gain was calculated based upon the difference between the students' state standardized math test scores in years 2012 and 2013. This was assessed at the student granularity.

Table 3.5 % of Data Observations with Gains, No Gains, and Percentage Dropped for the Three Gain Types.

Gain Types	ER?	Gained	Dropped	No Gain
Accuracy (N=75,083)	ER	48.10%	8.60%	37.90%
	No ER	43.70%	6.10%	36.60%
Confidence (N=75,083)	ER	28.30%	42.60%	23.70%
	No ER	26.40%	37.40%	22.70%
Math Test (N=4,827)	ER	41.60%	0.40%	46.90%
	No ER	40.80%	0.50%	45.70%

Note. 1) Observations in the 'Dropped' column (pre- and post-tests were both 0 or 1) were excluded from analysis. 2) Accuracy and Confidence Gains were measured at objective granularity, Math gain was measured at student granularity. 3) ER and no ER were collapsed across level.

11.8% of the students were excluded from the math gain analysis due to missing state math test records. These excluded students performed statistically significantly worse in the game as measured by the three pass attempt features; this implies that we excluded weaker students. 8.5% of the objective observations were excluded from the accuracy and confidence gain analysis due to missing pre- or post-tests. These excluded observations were not statistically significantly different from the rest as measured by pass attempt features. The accuracy and confidence gains were significantly correlated ($r=0.37$, $p<0.001$), but these two gains were not strongly correlated with math gain scores at the student granularity ($r<0.1$, $p<0.001$). Table 3.5 reports the percentage of data points that gained, dropped (mainly for avoiding ceiling effect in this data), and did not gain for each type of gain based on the Marx and Cummings Normalization method [MC07].

We first constructed decision trees to partition our data to see which factors influence gains, using the method described in the prior section. No sampling was necessary because the groups had similar sizes. We used pass attempt features, ER features, pre-test results, and demographics. For student granularity, we also added the percentage of required objectives attempted by the student.

At the objective granularity, we found that pre-test accuracy and confidence were the only selected nodes that predicted accuracy (70.0% accuracy) and confidence gain (74.1% accuracy). Students with a pre-test accuracy of < 0.71 (at least 2 questions wrong out of 5-10) had a 64.7% chance of positive accuracy gain in the same objective, although the remainder of the students had

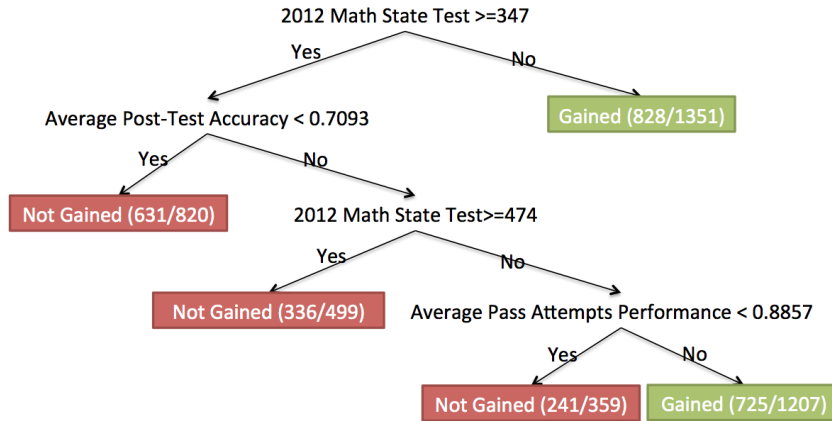


Figure 3.1 Decision Tree to Predict Whether a Student will Gain in State Standardized Math Test.

only a 25.9% chance. Students with high pre-test confidence (≤ 0.95 , indicated confidence on almost all questions) had a 62.5% chance of positive confidence gain in the same objective. It could be that these in-game tests were too easy, as 18.9% of pretests showed full scores in accuracy and 54.5% showed full scores in confidence.

Our decision tree for the student granularity is shown in Figure 3.1, with a cross-validated accuracy of 57.8%. Students who started with medium level of math abilities (2012 state test math scores < 474 , and ≥ 347) improved their scores when they performed well in ST Math (average pass attempts performance > 0.8857). This shows that the gameplay data in ST Math has predictive power for assessment outside of the game. However, for all three gain scores, the ER features were not selected for inclusion in the decision tree nor was any correlation found with the students gains.

Finally, we investigated how ER patterns relate to gains. Table 3.6 reports the result from separating students into 6 groups based on ER patterns and conducting Mann-Whitney U tests with Benjamini-Hochberg correction (as in the previous section). Moreover, although decision trees constructed from the complete dataset show that low pre-test results led to more gains, some ER pattern groups showed opposite trends. For example, Group B, who primarily ERed before passing the current level, started with lower pre-test scores, did worse in the game, and had fewer gains that were statistically significant, in all three gain measures. The same applies to Group DLSO. These two groups of students also had the lowest ER performance.

On the other hand, the Base group and Group A (who mostly ERed after passing the current level) started with pre-test accuracy and confidence scores that are not significantly different (Table 3.6), but Group A did significantly better in game, and had higher gains in accuracy and confidence, which were statistically significant. Because the mean pre-test score for the Base and A groups is approximately 0.6, these students were reasonably familiar with the objective before they began playing it. The difference in accuracy and confidence gains suggest that ER after students successfully

Table 3.6 Mann-Whitney U Tests Comparing Gains Between Student Groups of Different Elective Replay Pattern.

Group (# students)	Math (max=600)	Accuracy (max=1)	Confidence (max=1)
Base:No ER (N=1938)	M=31.5 SD=146.6	M=0.31 SD=0.25	M=0.33 SD=0.38
ER (N=2889)	M=27.3 SD=139.7	M=0.30 SD=0.25	M=0.32 SD=0.37
Group A (N=1114)	M=53.4 SD=167.9	*M=0.35 SD=0.24	+M=0.38 SD=0.36
Group B (N=1464)	+M=6.7 SD=109.0	*M=0.24 SD=0.25	*M=0.26 SD=0.37
Group SL (N=173)	M=46.2 SD=161.2	M=0.31 SD=0.28	M=0.31 SD=0.37
Group DLSO (N=983)	M=21.4 SD=123.0	*M=0.25 SD=0.26	*M=0.27 SD=0.37
Group DO (N=1399)	M=32.3 SD=150.6	M=0.32 SD=0.23	M=0.34 SD=0.36

Note. green and red indicate statistically significances higher and lower than the base class, with * $p < .001$, + $p < .01$

pass a level helped students learn, or implied better learning in the previous gameplay.

3.5 Contribution

In summary, this work adds new insights to our understanding of elective replay in educational games. Our work reveals differential associations between elective replay and performance when replay is categorized by the timing in relation to the student’s current learning objectives and gameplay. Our work suggests that low-performing students did not benefit from ER; high-performing students both chose ER at better times and their ERs were associated with benefits from either ER or previous gameplay, which supports the results of prior self-regulation research by Aleven et al [Ale03]. This work presents prospects for both examining more detailed characteristics of replay and utilizing experimental manipulations.

From the application perspective, as expected from this complex environment, our effect-sizes are too small to claim ER itself as a powerful intervention for learning. Instead, our findings suggest the potential of using ER patterns to identify weaker students and their struggling moments for intervention. For example, students with Group B ER patterns started weaker, did poorly in the game, and had lower gains in learning, confidence, and math state test scores. It may be the case that Group

B ER (before passing a level) is a signal that students are struggling in current content and are in need of a mental break [Sab13; Tal17] or help. If this is the case, it would be beneficial upon detecting these ER patterns for ST Math to alert teachers or to provide interventions, such as suggesting the student to take a break [Sab13] or providing supplemental resources to further explain the math concepts from the pass attempts interrupted by ER. Our results also suggest avenues for experimental studies that designs a more effective ER experience, such as preventing work-avoidance in ER. For example, changing the number of lives students have at each replay, or constraining the problems offered each time they are replayed to be isomorphic but not identical.

This work has several limitations. First, the in-game pre- and post-tests may be too easy for students, as 18.9% of pretests achieved a full score in accuracy, and 54.5% showed a full score in confidence. The high percentage of students with non-positive learning and accuracy gain could also be caused by students' slipping or guessing in multiple-choice questions (e.g., 1 incorrect answer reduces accuracy by 14%-20%). The accuracy of the pre- and post-test questions for assessing knowledge might be improved by using short answer questions. The second limitation is that we did not have puzzle granularity data on the types of errors they made. Third, the grouping of students based on the majority of elective replay assumes that elective replay is a habitual and consistent behavior. Future research should investigate other groupings, as well as examining whether there were changes in how students used replay, and what caused the changes. Fourth, future work may also include creating quantified features to compare the content and game features across objectives so we may better understand how the game's content influence students' decision to engage in elective replay.

CHAPTER

4

STUDY 2: INFORM CURRICULAR SEQUENCING THROUGH MINING PREDICTIVE RELATIONSHIPS BETWEEN MATH CONTENTS

Peddycord-Liu, Z., Cody, C., Kessler, S., Barnes, T., Lynch, C. F., and Rutherford, T. (2017, October). Using Serious Game Analytics to Inform Digital Curricular Sequencing: What Math Objective Should Students Play Next?. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play (pp. 195-204). ACM. [PL17]

Abstract. In this study, we applied game learning analytics to inform digital curricular sequencing in a longitude, curriculum-integrated math game, ST Math. When integrating serious games into classrooms, teachers may have the flexibility to change the order of math objectives for student groups to play. However, it is unclear how teacher decisions, as well as the sequencing of the original curricular order affect students. Moreover, few researchers have applied data-driven methods to inform content ordering in educational games, where the nature of educational content and student behaviors are different from many e-learning platforms. In this study, we present a novel method that suggests curricular sequencing based on the prediction relationship between math objectives.

Our results include specific design recommendations for ST Math, and general data-driven insights for digital curricular design, such as the pacing of objectives and the ordering of math concepts. Our method can potentially be applied to data from a wide range of games and digital learning platforms, enabling developers to better understand how to sequence educational content.

4.1 Introduction

This study aims to apply game learning analytics to help one specific kind of decision when integrating games into the classroom: what math objective should students play next? These analyses are situated on a year-long supplemental digital mathematics curriculum Spatial Temporal (ST) Math. Although game designers proposed pre-designed math objective sequences, teachers were given the flexibility of reordering and skipping math objectives for their classrooms.

We designed a novel method to understand curricular sequencing by modeling the potential predictive relationships between objectives. We demonstrated this method on gameplay logs from 1,565 3rd graders from 17 schools and 78 classrooms during the 2012-2013 school year. Students completed different sets of objectives and often did so in unique orders depending upon their classroom structure and individual needs. Based on our new data-driven predictive models, we determined specific recommendations for curricular improvements, including the appropriate location of objective groups, the content that could be used to prepare for objectives, the pacing and the ordering of objectives, and re-design of game puzzles. Our method can potentially be applied to data from other serious games and e-learning environments that are intended to be used over a long timespan.

4.2 Background

Educational research has shown that mathematics is inherently hierarchical—early mathematics skills are needed for later mathematics [Dun07; RS11]. Due to this hierarchical nature, children tend to follow a certain trajectory as early skills build upon each other to form a more complex understanding of mathematics [CS14]. Thus, in serious games that aim to help children develop complex interdependent skills of this type over a long timespan, math objective sequencing is an important design consideration to evaluate. The sequencing of math objectives is also an important factor in the use of ST Math. From previous literature, 42% of 863 surveyed teachers re-ordered math objectives in ST Math[Cal18], and the reordering actions positively associated with students math learning gain in state standardized math tests.

However, it is unclear whether teachers are aware of the impact of objective selection and sequencing. Galant [Gal13] asked 46 third grade teachers to justify the selection and sequencing of two mathematical tasks. Their responses reflected variance in the strength and weakness of

subject-matter knowledge and pedagogical content knowledge across teachers, with several teachers showing a lack of understanding of progression and development of math concepts. A recent study by Molin [Mol17] suggested that teachers not only have a limited time to prepare and play a game for game-based learning, but they are not usually engaged with game designers in making decisions on content or ordering. Data-driven methods can help discover the connection between math objectives in ST Math and inform its digital curricular sequencing.

4.3 Method

4.3.1 Participants & Data

MIND Research Institute, the developer of ST Math, collected and provided to the researchers gameplay data from 1,565 3rd grade students during the 2012-2013 school year. These students came from 17 schools and 78 classrooms. Among these students, 50.7% were male (2.9% NA), 80.7% were eligible for free/reduced lunch program (2.8% NA), 84.8% were Hispanic or Latino (2.8% NA), and 66.2% were English Language Learners (ELLs) (2.8% NA), and 10.9% were noted as having a disability (2.1% NA). For each attempt at a level (level attempt), ST Math logged the student’s ID, timestamp, and the number of puzzles completed.

The 3rd grade’s pre-designed curriculum contains 27 default objectives in a particular order, as shown in Table 4.1. Teacher activities are the root cause of the variance in the order of objectives played by students. A teacher can separate his/her class into groups, and reorder objectives in the curriculum for each student group. For example, if some students missed the first class, a teacher can put them in a separate group, and assign what the rest of the class are currently playing first, and the missed objectives later. A teacher can also allow certain students to ‘skip’ an objective, which means playing the next objective before completing the current objective. Meanwhile, students can revisit skipped (allowed by teacher) objectives at any time. This causes variance in the order of objectives played by individual student.

Table 4.1 A Description of Objectives in the Curriculum and Objectives with Predictive Power Over Each From Regression Analyses.

Objective Name	Descriptions
SO1E-Equal Group	Figure out N, X, Y for scenarios that split N objects or building block into X groups of Y items

Table 4.1 (continued).

SO1F-Multiplication Concept	Complete a multiplication formula with given objects; Represent a multiplication formula as addition using item, number line, addition formula; Calculate result of simple multiplication with help of building block.
SO1O-Place Value Concept	Represent three digit numbers as flower petals in group of 1,10,100.
SO1P-Place Value Bundles	Represent three digit numbers as flower petals in group of 1,10,100; Addition of numbers to form building blocks in unit of 10s and Hundreds.
SO2-Order & Compare Whole Number	Locate number on number line; Compare number with/out help of items and number line; Round numbers up and down on number line.
SO1Y-Add & Subtraction w Regrouping	Add and subtract three-digit numbers represented as flower petals, as numbers with flower petals, and as numbers with flower petal animations; Regroup flower petals in calculation.
SO28-Division Concept	Divide countable objects into X groups equally.
SO29-Division	Calculate simple division; Present answer through objects, building blocks, and number line.
SO5-Lines & Angles	Match types of angles with English name; Identify parallel and perpendicular lines from given lines.
SO2I-Shapes	Match 2D shape objects with English name; Identify edge numbers for shapes, objects, and shape's English name; Draw shapes with pins and a rubber band; find shape pairs with same number of edges.
SO2J-Shape Attributes	Identify number of edges, vertices, faces in 3D geometry.
SO8-Algebraic Expressions & Equations	Represent equation with objects; separate numbers and units objects into equal groups; Identify missing operations from equation.
SO9-Functional Relationship	Find X when split N into X groups of Y objects; Fill in missing number of linear pattern (e.g., 2,4,?,8) with/out help of countable objectives.
SO2S-Fraction Concepts	Split bar into equal areas; Adding parts of pie chart; Represent fraction with bar or pie chart; And represent pie chart as fraction.
SOB-Fraction Addition & Subtraction	Represent fraction's denominator and numerator as parts of robot, slices of pie chart; Add and subtract with these countable objects; Locate calculation result on number line with unit 1.

Table 4.1 (continued).

SOC-Fraction Addition & Subtraction L1	Add and subtract fractions without countable objects; calculate on fractions greater than 1.
SOD-Money & Decimals	Recognize money unit and coins; Compare money amount; Add, subtract, and find combinations of paper money and coins equal to certain amount.
SOE-Measurement	Compare weight of items using balance and scale; measure with ruler; Translate between hours and minutes, and between inch and foot, and specify the answer on number line.
SO3C-Concepts of Area and Perimeter	Calculate perimeter and area of 2D shapes presented with/out square grid or fill in calculation formula; Identify 2D shapes for given perimeter and area; Select grids based on difference of 2D shape area and given number.
SO3D-Volume & Weight	Order objects' weight using balance and scale; Indicate volume of a given 3D object in cubic.
SOG-Addition & Subtraction to 10,000	Calculate multi-digit addition and subtraction; Represent answer on number line or unit block; Fill in missing digit given the result of a multi-digit addition and subtraction.
SO3M-Number Patterns	Identify or fill in linear pattern (e.g., 10, 20, 30, 40) with help of building block; Use data-table representation for addition, subtraction, multiplication test.
SOI-Multiplication of Multiple Digit	Use building block (grid) to represent multiplication; Calculate multi-digit multiplication with 2 dimensional representation of building block.
SOJ-Fraction & Decimal Equivalence	Add and subtract fraction&decimal, present result in 10-by-10 grid; Locate fraction&decimal on number line of unit 1; Represent decimal as formula of fraction addition; Use coins to present dollars in decimal number.
SOK-Outcomes	Distinguish whether countable object is "impossible", "unlikely", "likely" or "certain" to be picked from a jar and from lucky wheel.
SOL-Using Data & Graph	Present numbers of objects in bar chart and pie chart, and read number from the charts.

For the purpose of this analyses, we filtered out three types of level attempts: students' elective

replay on previously passed levels (1.69%), level attempts from default objectives where all students completed less than 25% of the objectives' levels (0.54%), level attempts from 15 optional objectives provided by ST Math, and the last default objective (3.12%). We filtered out optional objectives because very few students attempted them. We filtered out the last default objective because it was a “challenge” level with spatial puzzle games that were not designed to teach math concepts. After filtering, students attempted (completed 25% levels) an average of 15.82 default objectives (sd=8.07, 95% CI: [15.42, 16.22]), with 371 students from 51 classes completing at least 25% of levels from each of the 26 default objectives. All of the remaining data (94.65%) from 1,565 students were included in the analyses.

We first extracted objective sequences as the order of objectives that students played throughout the year. Each objective occurred only once in the sequence, representing where most of the students' effort (level attempts) occurred. For example, if a student attempted the first level of objective A, and was ‘skipped’ by the teacher to objective B, completed all levels in objective B, and went back to complete the rest of the levels in objective A, then objective B is ordered before objective A because most of the gameplay happened before objective A. Note that without teacher-allowed ‘skipping,’ a student cannot play the next objective until they have completed the latest objective's post-test. Thus, except for the rare skipping case, a student cannot complete two objective simultaneously.

Table 4.2 Statistics of the Sequence of Objectives 3rd Grade Students Played in ST Math.

N	#Students Played \geq N Objs	#Diff. Orders Students Played the First N Objs	#Students in the Top 5 Most Frequent Order
5	1414	107	550,214,89,69,37
10	1104	194	322,131,64,54,28
15	876	205	263, 74, 48, 39,22
20	595	156	186,47,41,24,20
26	371	84	129,34,30,20,19

Table 4.2 reports the statistics from objective sequences. Our data have over a hundred different objective sequences, with a couple of objective sequences occurred in high frequencies. This was largely due to the nature of sequencing in ST Math—players followed a pre-designed curriculum order, with variance introduced by individual activities.

4.3.2 Mining Predictive Relationship

Next, we sought to understand the relations between math objectives, and to provide data-driven insight on the role and ordering of objectives. The first step was to select a single feature to measure student learning from an objective. We used percentage of levels completed within an objective, multiplied by the objective performance. We defined objective performance to be the average performance across the given objective's levels, and a level's performance as the average level attempt performance prior to and including the first attempt that passed the level. A level attempt performance is the percentage of puzzles a student completed before losing all lives on a single attempt of a level. The number ranges from two to four lives depending upon the number of questions in the level and their difficulty. Prior research has shown that this measurement can be used to predict normalized gains on the students' 2012 and 2013 state math test scores [Liu17]. These tests provide standardized measurements for students' math ability outside this game environment. We chose to not use the pre- and post-tests in ST Math in this analysis due to their lack of precision. The pre- and post-tests contain 5-7 multiple choice questions, so a lucky guess and a mistake could cause a 14-20% change in score. The pre- and post-tests also have missing data (11.8% in pre- or post-test) and ceiling effects (19.0% achieved full pre-test scores).

In the second step, it was necessary to focus on the most influential objectives. To do so, for each objective (we refer to as goal objective), we conducted linear regression analyses to find which prior objectives were most predictive of its objective performance. To avoid over-fitting, the first step was to filter the objectives included in the regression. To do so, we applied Spearman Correlation for each goal objective, against all the objectives that happened before it. To ensure there were enough observations, we applied a power test, and filtered out objective pairs with fewer than 30 observations. This step left us 430 correlations for 26 goal objectives. Given the fact that all objective performances were influenced by a student's general math ability, our correlation table had an average value of 0.39. Thus, we decided to apply linear regression with only those objectives with a correlation of at least 0.39. This resulted in considering only 189 of the 430 possible objectives-goal pairs.

In the third step, we applied step-wise mixed effect linear regression for each goal objective, as:

$$P_g = \beta_0 + \beta_1 M + \beta_2 I_g + C \sim N(0, \sigma^2) + \sum_{i=1}^k \beta_i P_{g_i} \quad (4.1)$$

Each regression included only the students who attempted the goal objective. The dependent variable P_g is the percentage of levels completed, multiplied by the averaged objective performance of the goal objective g . Betas are coefficients for independent variables, derived from the regression. M is the state math test in 2012, which measured students' general math ability prior to playing ST Math. It was also necessary to include classroom C as a random effect to account for students'

being taught by different teachers, especially given the importance of teachers in deciding upon the objective ordering of their classes and groups in classes. The position of objective j in each student's objective sequence is denoted by I_g . We assumed there is a linear relationship between how many objectives students played in ST Math, and the ability to perform well in the same game environment. $P_{g_1} - P_{g_k}$ describe the k objectives found from the previous step: happened before goal objective g in a sufficient number of sequences (at least 30 and passed the power test for correlation), correlated with g ($r > 0.39$), and were independent of each other ($VIF < 1.5$). Based on our outcome measurement of learning, P_{g_i} is 0 if the student did not do an objective i before the goal objective. We also used a step-wise process to automatically add or remove variables based on AIC (Akaike information criterion). This process helped select variables that had the most predictive power for goal objectives, under the influence of student's general math ability, classrooms, and the sequential position of the goal objective. This step also avoided over-fitting through cross-validation.

After applying the Benjamini-Hochberg Correction for multiple statistical tests, a total of 96 predictor objectives were found to be statistically significant predictors for the 26 goal objectives. The coefficients had an average standardized beta of 0.12, with 92 coefficients being positive.

4.4 Results & Discussion

This section presents the results from two angles: categorization of objectives, and categorization on objective pairs. These categorizations were interpreted in relation to our research goal—what data-driven insight we can provide to inform the design of digital curricular sequencing.

4.4.1 Categorization of Objectives

Figure 4.1 is a plot of the predictive power and predictability of objectives, categorized by k-means clustering. Objectives in Group Dark Blue predicted a high number of other objectives. These objectives also occurred early in the pre-designed curriculum, which suggests that they captured basic skills needed in the later objectives. Objectives in this group could be used as early indicators of later performance. For example, if a student skipped or performed poorly on these basic objectives, the teacher may consider intervening to prevent poor performance in later objectives.

Objectives in Group Red were predicted by a high number of other objectives. This implies that playing objectives in this group required math skills learned from many other objectives. For example, games in SOD: "Money and Decimal" required students to compare, add, and subtract money in the form of both paper money and coins. Performing these tasks required a variety of math skills including unit transformation, calculation, number comparison, and a preliminary understanding of fractions and decimals, because coins represent fractions of a dollar. Additionally, three objectives in this group (SO3M: "Number Patterns," SOD: "Money and Decimals," and SOB: "Fraction Addition &

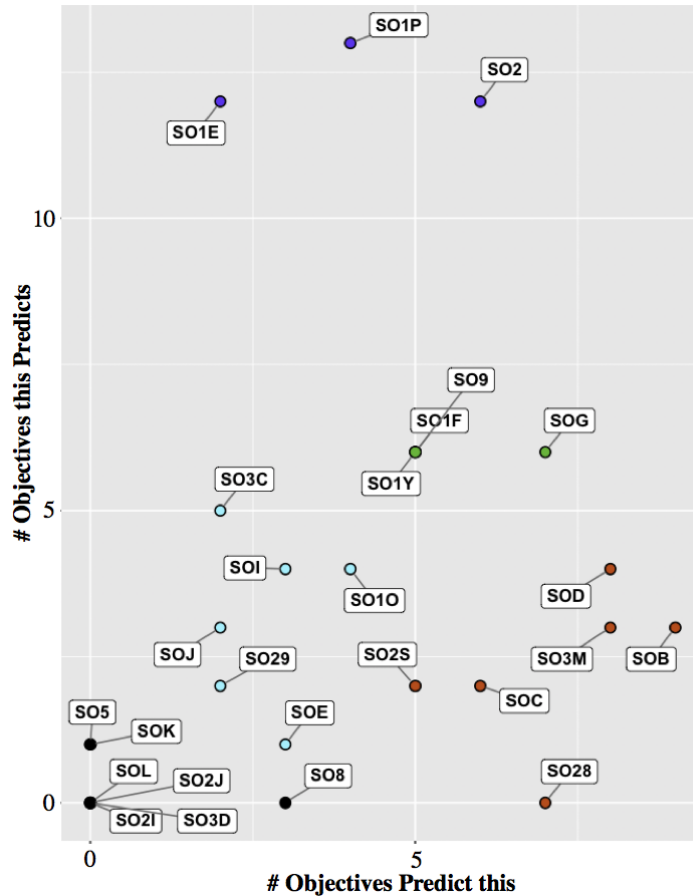


Figure 4.1 Objectives, the # of Objectives They Predict, and the # of Objectives that Predict Them.

Subtraction”) were among the top five most difficult, in that students took the most attempts to pass levels within these objectives. This indicates that foundational skills may be especially important for these objectives, and struggle within them may indicate to teachers that students lack prerequisite knowledge.

Objectives in the Group Green both predicted a high number of objectives and were themselves predicted by a high number. This implies that the math skills covered in these objectives are more advanced (depend on many prerequisites), and useful (as other objectives depend on them). For example, both SO1Y and SOG involved multi-digit addition and subtraction tasks. These tasks required a solid understanding of numbers and units, and strong calculation skills. Playing puzzles in these objectives, in return, enhanced understanding on numbers and provided practice in the calculation skills needed for many other objectives. This indicates that the Green Group objectives were difficult, but should be completed before moving on because they could affect later performance. In other words, if some students weren’t doing well on these objectives, we would

not recommend that teachers skip these students forward. Instead, the teachers should help the students learn this material rather than allowing them to advance in the game.

The majority of objectives in Group Black (SOL, SO5, SO2I, SO2J, SO3D, SOK) were not predicted by any objectives, and were also not predictive of any objectives. This implies that these objectives practiced math skills that were isolated from other objectives. For example, SO5, SO2I, SO2J involved tasks such as identifying shapes, faces, edges and types of angles. This suggests that Group Black objectives can be moved around more flexibly in the curriculum, and may be used at the beginning of the curriculum to familiarize students with the ST Math system.

4.4.2 Categorization of Objective Pairs

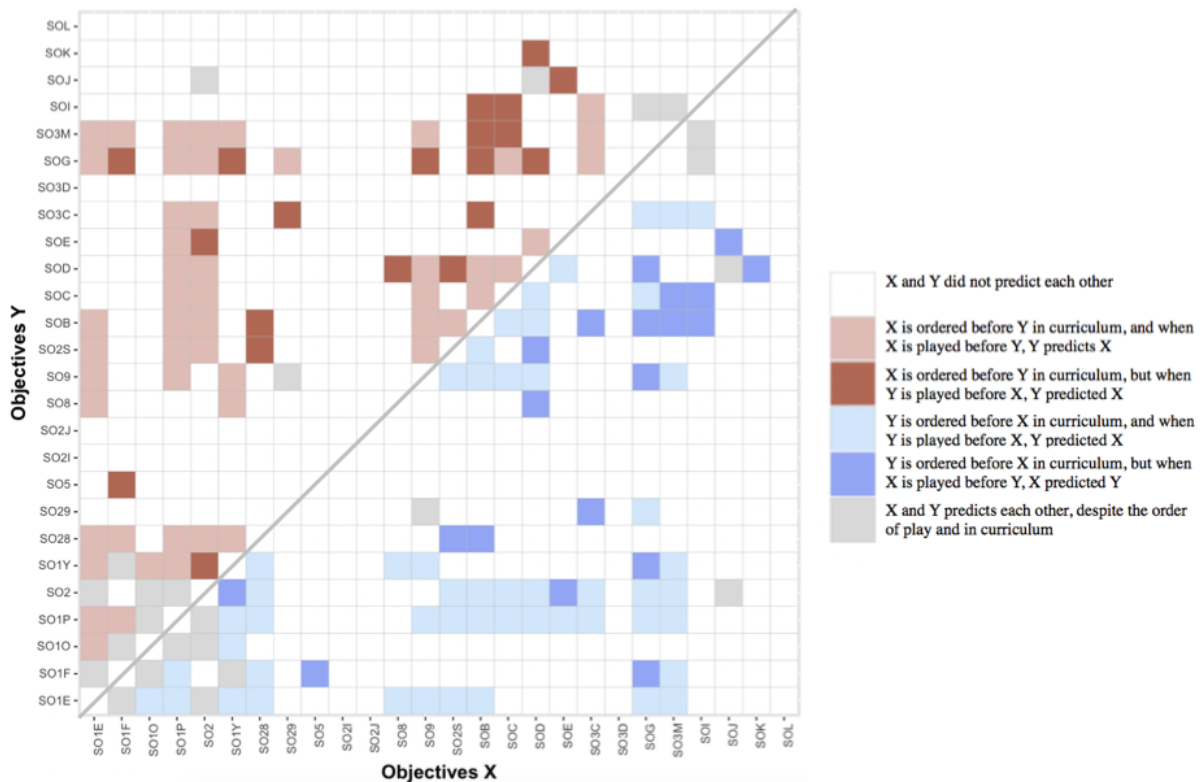


Figure 4.2 Pairwise Predictive Relationship between Objectives.

Figure 4.2 presents the prediction relationship between objective pairs, symmetrically along the X=Y line to show complete information of an objective when reading from a row or column. A grey color shows a symmetrical relationship: when X is attempted first, X predicts Y; when Y is attempted

first, Y predicts X. A blue or red color shows an asymmetrical relationship: one objective predicts another if attempted first, but not vice versa.

Within asymmetrical relationships, a light blue or red shows a curriculum-order asymmetrical relationship: the objective attempted first and which predicts the other, is also ordered earlier in the curriculum. This relationship may exist because either the later-ordered objective was not chosen as a predictor of the earlier-ordered objective, or because we simply do not have enough data (failed power-test) to tell so. For either reason, this relationship suggests keeping the current ordering of the objective pairs in the curriculum.

On the other hand, a dark blue or red shows a non curriculum-order asymmetrical relationship: the objective attempted first and which predicts the other is ordered later in the curriculum. This relationship implies that students who completed more levels and performed better in the later-ordered objective before attempting the earlier-ordered objective performed better in the earlier-ordered objective. These relationships are the most interesting as they may suggest potential changes to the existing curriculum. Below sections categorize the 20 non curriculum-order asymmetrical relationships we found based on their implications for curriculum design.

4.4.2.1 Suggesting more Practice on Certain Skills (8/20)

This group includes: SOG: “Addition & Subtraction to 10,000,” predicted SO1E, SO9, SOB, and SOD; SOI: “Multiplication of Multiple Digit” predicted SOB and SOC; SO1Y: “Addition & Subtraction with Regrouping” predicted SO2; and SOJ: “Fraction & Decimal Equivalence” predicted SOE: “Measurement.” One explanation is that these predictor objectives aimed to intensively practice a specific skill, such as multi-digit addition. These skills were needed in the earlier objectives they predicted. For example, SOJ: “Fraction & Decimal Equivalence” predicted SOE: “Measurement,” where the predictor SOJ practiced the understanding and transformation between fractions and decimals. This skill was needed in SOE, where many games required students to place certain minutes on a number line of unit hours, or certain inches in a number line of unit feet. A solid understanding of fractions and decimals gained from SOJ may help transforming minutes and inches as fractions of hours and feet. However, the relationship is asymmetrical because the predicted objectives were not designed for enhancing the specific calculation skills in the predictor objectives. Thus, they were not the major influencers of the performance and completion of their predictors.

Although relationships in this group do not necessarily suggest that these calculation-based objectives should be moved earlier in the curriculum, these relationships imply that the performance and completion of many objectives could be improved if there were more practice on certain calculation skills beforehand.

4.4.2.2 Suggesting more Applicable Examples (4/20)

This group includes: SOD: “Money & Decimals,” predicted SO8, SO2S; and SO3C: “Concepts of Area and Perimeter,” predicted SO29, SOB. One explanation is that tasks in SOD and SO3C contained examples and practical scenarios for the more conceptually challenging objectives they predicted. For example, SOD practiced the recognition, comparison, and calculation of money units such as dimes, quarters, and dollars. Because coins are fractions of a dollar, these tasks provided examples for SO2S: “Fraction Concepts.” Another example is SO3C practiced calculating area and perimeter for a given square grid or highlighting areas in a square grid for a given area and perimeter. These tasks were practical scenarios to practice addition, multiplication, and division at small-scale, with the help of countable squares. Thus, SO3C predicted SO29 and SOB. Relationships in this group suggest that performance and completion for some conceptually difficult objectives can be improved with more applicable, simplistic examples beforehand [Bok15].

One interesting result is that although SOD: “Money & Decimals” predicted SO2S: “Fraction Concepts,” the addition and subtraction of fractions predicted SOD (curriculum-order symmetrical relationship). This may be because the later levels in SOD required more complicated calculations. For example, some tasks asked students to add dimes and quarters to a double-digit amount in dollars, which could be helped by practice on fraction addition, and on the transformation between fractions and decimals (as SOD and SOJ predicted each other). These levels contained much harder tasks than other levels in SOD, as we found that these four levels in SOD cost students an average of 5.14 attempts to pass, whereas the other 12 SOD levels cost an average of 1.71 attempts. Thus, there may be a benefit in splitting SOD. An example would be moving the earlier levels for recognizing and comparing coins before fractions so these levels could serve as practical scenarios for using fractions, and moving the harder levels after or as levels within fraction calculation objectives.

4.4.2.3 Suggesting Reordering of Teaching Math Concepts (2/20)

SO2S: “Fraction Concept” and SOB: “Fraction Addition and Subtraction” predicted SO28: “Division Concept.” Although it is conventionally believed that division should be taught before fractions, with prior research finding that division predicts fractions skills (e.g., [Jor13; Bai14], there may be reasons why the specific actions within these objectives of ST Math produced the opposite order. SO2S and SOB presented fractions as countable parts of a divisible object, such as slices of a pie or groups of bricks. Manipulating these objects’ parts may help students develop a better understanding of division through partitioning items into groups in a familiar way with smaller numbers [MS05]. This relationship is asymmetrical because games in the fraction objectives generally involved the addition and subtraction of divisible objects, which were too simple to involve division calculations.

4.4.2.4 Suggesting a Change in Game Design (2/20)

SOG: “Addition and Subtraction within 10,000” predicted SO1Y: “Addition and Subtraction with Regrouping.” It could be that if a student practiced SOG and became more proficient with calculations, s/he may have more working memory to perform the extra counting and regrouping tasks in SO1Y [Imb07]. In SO1Y, digits were presented as petals in groups of one, 10, and 100. In the majority of tasks, students needed to count and regroup petals during calculation, with some counting above 10. In certain tasks, students also needed to present their calculation results in the format of petals. On the other hand, SOG practiced calculation with more traditional presentations, such as the numbers themselves, number lines, or building blocks marked with one, 10, or 100. SOG also did not require students to regroup objectives during calculation or to count over 10.

This relationship suggests that counting and regrouping petals during the calculation process may be ineffective and distracting. This is evidenced by the pre- and post-test scores. Recall that the pre- and post-tests in ST Math consist of parallel questions (same problem type with different numbers). In both SO1Y and SOG, The tests ask student to choose the correct answer when adding or subtracting multi-digit numbers. In SO1Y: “Addition and Subtraction with Regrouping,” the average pre- and post-test scores were 0.64 and 0.66; in SOG: “Addition and Subtraction within 10,000,” they were 0.66 and 0.73—a much higher learning gain. It could be that students experienced the split-attention effect defined by Chandler and Sweller[CS92], because students spent their limited working memory resources mentally integrating two sources of informations (e.g., number and petals). Also, this relationship is not symmetrical (SO1Y does not predict SOG), which could be because the performance and completion of SO1Y reflected more about the difficulties encountered in counting and regrouping petals, than in the calculation shared with SOG. This difficulty and ineffectiveness of regrouping and counting petals is also evidenced by statistics from SO1P (Place Value Bundles using petals). SO1P had the second lowest level attempt performance among all 26 objectives, and only a 0.03 gain from pre- to post-test.

Another relationship that suggests a change in game design is SOE: “Measurement,” which negatively predicted SO2: “Comparing Whole Numbers.” One likely explanation is that the number line was presented differently in these two objectives. In SOE, the number line was represented by a ruler, scaled for time units and for imperial and metric length units. This means that in SOE, a number line can be one centimeter, 60 minutes, or 12 inches. On the other hand, SO2 presented the number line as metric units. Because young students often do not have an accurate understanding of the number line [Gea08], introducing conflicting number lines may create misconceptions and interfere with their number line knowledge. This hypothesis is evidenced by the performance statistics in SO2. The first level of SO2, which requires students to locate an integer on a number line, took students an average of 5.5 attempts to pass, whereas the rest of the levels in SO2 (some contained similar tasks of a locating number on number line) only took students an average of 1.5

attempts to pass, with the max being 2.6 attempts. This relationship suggests that using different units on the same presentation of a number line may be confusing, and similar practices on unit conversion should be moved after students have developed a concrete understanding of the number line.

4.4.2.5 Suggesting a Change in Pacing (2/20)

SO3M: “Number Patterns” negatively correlated with objectives SOB: “Fraction Addition & Subtraction L1” and SOC: “Fraction Addition and Subtraction I and II.” It could be due to the pacing in the most common sequence, in which SO3M was played before SOB and SOC. The majority, 293 out of 360 students who played SO3M before SOB, played in a sequence “SO28 SO3M SO29 SO9 SO2S SOB SOC.” SOB and SOC required calculation in conceptually difficult topic in third grade curricular–fractions. SO3M contained difficult tasks, such as asking students to find or complete number patterns of 3-6-9-12, or performing multi-digit multi-step addition, subtraction, and multiplication using a new data-table representation. As compared to the pre-designed curriculum in Table 4.1, these three relatively difficult concepts, division, number patterns, and fractions, were spaced too close to each other. Thus, it could be that students get mentally overloaded when too many difficult concepts and different number representations are packed together. This relationship suggests that teachers should vary the pacing accordingly and take breaks between difficult concepts [WR07]. For example, teachers can allow students to play simpler objectives that offer more concrete examples or practical scenarios to prepare for the upcoming difficult concepts. Teachers can also assign objectives in the Group Black described in in previous, as they practiced concepts that are isolated from other objectives.

4.4.2.6 Unclear (2/20)

SO5: “Lines and Angles” predicted SO1F: “Multiplication Concepts.” One possible explanation is that SO1F was introduced very early in the curriculum, with few objectives beforehand to explain its variance in performance and completion. SO5 required selecting the right names for angle types and distinguishing parallel and perpendicular lines; it is a simple objective that could be captured by general skills, such as memorization, that helped predict SO1F. Another unclear relationship is that SOK: “Outcomes” predicted SOD: “Money.” SOK dealt with probability problems. It could be that the presentations in SOK closely resembled lucky wheels and lottery jars, which both relate to money, but the relationship is still unclear.

4.5 Contribution

This study presents a novel method that informs curricular sequencing in a serious game, taking advantage of the variance in the order of objectives that students played. We applied this method to ST Math gameplay data from 1,565 3rd graders. Our results include specific design recommendations for ST Math, and general data-driven insights for digital curricular design, such as the pacing of objectives and the ordering of math concepts. The method itself, and some of these data-driven insights can inform design for digital curricular design in other serious games and e-learning platforms.

From a technical perspective, the growth of educational games and other e-learning platforms enables a large amount of data collection. This in turn has facilitated the use of innovative analytical methods that reveal user behavior patterns, which go beyond traditional experimental trials. However, data analytics, although informative, can make it easier to find statistically-significant results that may not be grounded in theory, and simply arise because of the sheer volume of available data. Therefore, when analyzing any large data set such as ST Math, we want to ensure that our analyses help uncover educationally relevant results. From the education literature, researchers have identified some early skills that predict later mathematics skills. For example, the general knowledge of the meaning and magnitude of numbers, is highly predictive of later mathematics achievement, including number line representation and calculation fluency [Jor10]. These skills are in turn predictive of skills such as fractions [Bai14; Jor13]. We believe that our data-driven work on discovering predictive relationships can be leveraged to discover more learning trajectories[SK12], as well as to make practical sequencing recommendations that may not have been apparent from a purely top-down approach.

From a practical perspective, this analysis helps guide one crucial decision when integrating serious games into a classroom: what objective should students play next? Recent research found that 42% of 863 surveyed teachers who use ST Math re-ordered math objectives[Cal18], and such behaviors was positively associated with students' California Standards Test (CST) math scores [Cal18]. This motivates future work on designing a user interface to notify teachers of the likely impact of individual ordering decisions, and further support this well-used, and potentially effective feature. It is possible that there exists some optimal pathway through which students should experience a given game. Unfortunately, our method could not be used to find the optimal path in present context, as our data have too much variance both in objective selected and those completed. In future work, we plan to develop quantitative metrics to compare individual paths against better orderings and to relate those variances to performance.

Serious game designers can use our method and its data-driven insights to evaluate the sequencing and predictive relationships within their games. Many of our results are applicable beyond ST Math. For example, we found that counting objects during calculation distracted students, which

implies that game designers should avoid designing ‘extra’ interactions that split students’ attention during problem-solving. Other results may be limited to ST Math’s unique design characteristics. For example, fraction puzzles used countable parts of a divisible objects, such as divided pie charts, which helped students to understand division. Such a predictive relationship may or may not hold in games that use different representations. To assess the impact of differences between game environments, game designers can follow the same analytical procedure, and redefine performance variables and performance-associated mixed effects. Our method can potentially be applied to similar data outside of the serious games domain, where content is presented under varying sequences over a long time span.

One limitation of this work is that there are many factors outside of the system, such as varying instructional methods, that may affect learning and performance. We sought to control these factors by adding random effect distributions for individual classrooms. We also worked with domain experts to interpret the results, and examined the candidate predictive relationships by playing the games. Future work could collect qualitative data on teachers’ understanding of these data-driven insights and the effects of their applications in classrooms.

CHAPTER

5

STUDY 3: PINPOINT WHERE LEARNING & TRANSFER SUPPORT IS NEEDED USING LEARNING CURVE ANALYSES

(best student paper nominee) Peddycord-Liu, Z., Harred, R., Karamarkovich, S., Barnes, T., Lynch, C., and Rutherford, T. (2018, June). Learning Curve Analysis in a Large-Scale, Drill-and-Practice Serious Math Game: Where Is Learning Support Needed?. In International Conference on Artificial Intelligence in Education (pp. 436-449). Springer, Cham. [PL18]

Abstract. In this study, we applied data-driven methods to understand learning and derive game design insights in a large-scale, drill-and-practice game: Spatial Temporal (ST) Math. In order for serious games to thrive we must develop efficient, scalable methods to evaluate games against their educational goals. Learning models have matured in recent years and have been applied across e-learning platforms but they have not been used widely in serious games. We applied empirical learning curve analyses to ST Math under different assumptions of how knowledge components are defined in the game and map to game contents. We derived actionable game design feedback and educational insights regarding fraction learning. Our results revealed cases where students failed to transfer knowledge between math skills, content, and problem representations. This work stresses the importance of designing games that support students' comprehension of math concepts, rather

than the learning of content- and situation-specific skills to pass games.

5.1 Introduction

In this study, we applied empirical learning curve analyses. We fit and combined learning curves under different assumptions of how knowledge components are defined in the game and mapped to game contents. Our analyses aim to: 1) Understand and model learning in ST Math—a large-scale, drill-and-practice game that introduces and reinforces math skills through various problem-solving scenarios. 2) Derive actionable feedback to help game designers better design game content to support learning. 3) Provide data-driven insights on fraction learning and the knowledge transfer between problem-solving scenarios. 4) Suggest future research that analyzes and models learning in serious games.

5.2 Background

Learning curve analysis, as previously described in Chapter 2, originated from a classic cognitive theory assuming that with more practice opportunities, a students' accuracy at answering a question improves following a logarithmic curve [NR81]. Learning curve analysis is ideal for games with a drill-and-practice mechanic like ST Math, where students practice questions on the same skill again and again.

Learning curve analysis has been widely applied in intelligent tutors and other e-learning systems [Cen06; Cen08; Pav15]. However, there has been little application in educational games. As mentioned in Chapter 2, there is some work on applying learning curve analyses to educational games, that has pinpointed problematic game content, identified skills that students needed extra support to learn, and discovered unforeseen gaming strategies where students find short-cut to pass the level instead of sufficiently understanding the math concepts taught by the game [HA15; Bak07; Lom13]. Because ST Math practices math skills through various problem-solving scenarios, learning curve analysis would not only pinpoint where game design can be improved in this large-scale game, but it can also help us understand where learning support is needed to help students transfer across problem solving scenarios.

5.3 Method

5.3.1 Data

MIND Research Institute, the developer of ST Math, collected and provided the researchers with sample data from 3rd grade students who played ST Math from August 2016 to February 2017.

We focused on the “Comparing Fractions” objective. Performance on fractions has been found to predict future mathematical achievement [Sie12; Tor15; Bai12b]. Thus, investigating this objective will allow us to contribute suggestions for game design of instruction around a crucial math concept. This objective contains 26 levels across seven games; 1,007 students completed the first game, and 860 students completed the last game. ST Math recorded students’ IDs, answers, and response times for each puzzle attempt. The data also included the correct answer for each puzzle, and the level, game, and objective to which it belonged. We filtered out students’ replay of previously passed levels [Liu17] to focus solely on their attempts to pass an unlearned case. The final data contains 146,498 unique puzzle attempts.

5.3.2 Fitting Learning Curves

Our goal was to understand and model learning in ST Math, and to suggest better game design that would foster greater learning. ST Math is structured at the top level by objectives, where games representing different problem-solving scenarios and levels are puzzle sets of increasing difficulty. Puzzles are either randomly generated following a template, or randomly selected from a pre-designed puzzle pool. We first fit learning curves to the puzzles to identify the learning patterns at each level. We then combined the levels hierarchically within each game. We sought to find similarities between the levels—modeling levels as continuations within a single learning curve. Lastly, we sought to find associations between games. We used an expert-designed Q-matrix with knowledge components describing the shared math skills and problem representations across games. We fit learning curves using different assumptions to identify how these knowledge components interacted and affected students’ learning across games.

We used the AFM [Cen06] and CFM [Cen08] models to fit learning curves. We decided to model the probability that a student answered a puzzle correctly on the first attempt only. This is because students receive animated feedback after each answer, with some feedback enabling them to quickly correct a wrong answer without having to redo the entire problem. Thus, subsequent attempts may not truly reflect each student’s knowledge of the mathematics content. Next, we decided to fit a learning curve on the first N puzzles in each level, where N is the number of puzzles that must be answered correctly in order to pass the level. We chose N because students who passed the level without exhausting all of their lives (which is the majority in most levels) will not need another attempt. Attempts following N will only contain data from low-performing students who had to attempt the level multiple times to pass. Thus, we only consider students’ first N puzzles (practice opportunities) in order to fit our learning curve on the same population. Lastly, for student variables in AFM and CFM, we used students’ average performance in the prior two objectives: “Fraction Concepts” and “Fractions on a Number Line.” These two reflect students’ knowledge of fractions prior to attempting this objective.

5.4 Result & Discussion

5.4.1 Analyzing Puzzles in Levels

In ST Math, each level stresses skills of increasing difficulty under a problem-solving scenario defined by the game. The increased difficulty can be introduced by changes in math content (e.g., using larger numbers), changes in problem representation (e.g., use of math symbols instead of an visual object), or other factors. However, in each case the problem-solving scenario (e.g., finding shoes for animals) remains the same. Thus, we started by assuming one-to-one mappings between levels and knowledge components, and fit learning curves with AFM.

Table 5.1 shows the learning curve fit. We applied 10-fold cross-validation and reported the model's accuracy as the percentage of instances correctly predicted by the logistic regression models. The majority of puzzles were answered correctly by over 50% of students, therefore we include the AUROC to describe how well models can differentiate between true positive and false positive.

One assumption of the learning curve analysis is that problems that test the same skill should be ordered randomly. In ST Math, students do not play exactly the same puzzles, or in the same orders because puzzles are either randomly ordered, randomly selected from a pre-designed pool, or randomly generated. As described in Chapter 1, two thirds of the puzzles are randomly generated or randomly selected from a larger pool of problem, meeting the requirements for applying learning curve analysis. For the other third, the problems are randomly ordered. For this study, I assume that ST Math is designed so that the puzzles in each level are testing the same skill or knowledge component. This means there is considerable similarity in the content that students see within a level, and that the ordering of puzzles does not make a difference in students' performance.

More specifically, in the objective we analyzed, due to the limited choices for small fraction numbers and the high cost of designing graphical representations for different fractions, puzzles from the majority of the analyzed levels are randomly ordered or selected from a small fixed set of 2-11 different pre-designed puzzles. The italicized text in Table 5.1 describes the templates used to generate puzzles for each level. For levels that required students to pass N different puzzles in a random order, all students did the same N puzzles with $N!$ different orderings. Given that 860 students completed the game, most orderings would have been seen in the dataset except for Game1 L2 and Game2 L1 where there were $8!$ orderings. For these levels, we assume that the ordering did not make a difference. For learning curve analysis, the same assumption needs to hold for levels that required students to pass N puzzles randomly selected from a pre-made puzzle pool. Eight levels in the analyzed objective have a pre-made puzzle pool with $N+1.7$ puzzles on average. Assuming that there are 10 puzzles to choose from for a level that requires students to pass eight puzzles, then there would be 90 different puzzle sets with $8!$ orderings each. Each puzzle set would occur about 86 times, but only a few orderings could be covered for each puzzle set. Thus, we must again assume that such

Table 5.1 Learning Curve Plots and AFM Statistics.

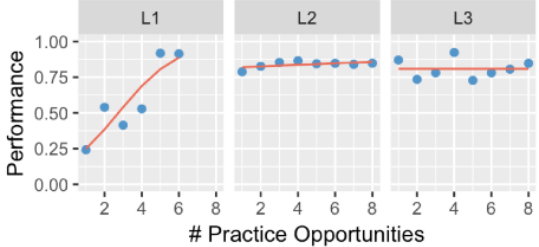
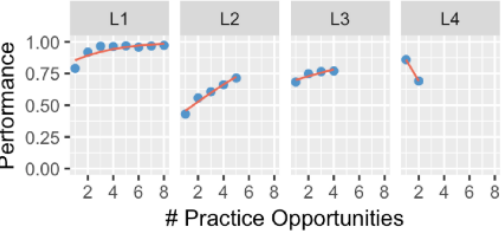
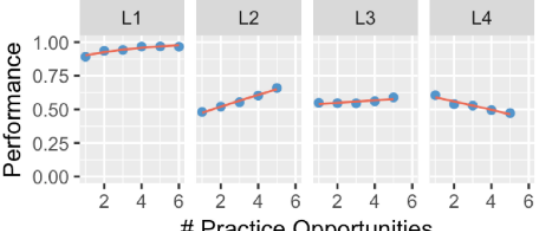
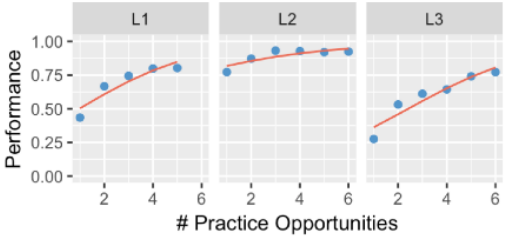
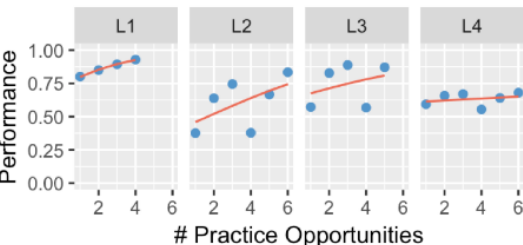
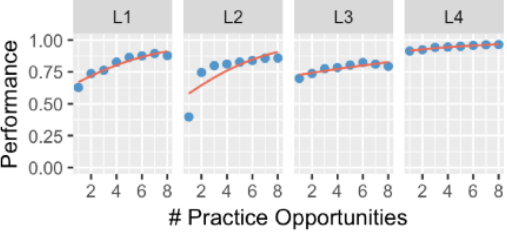
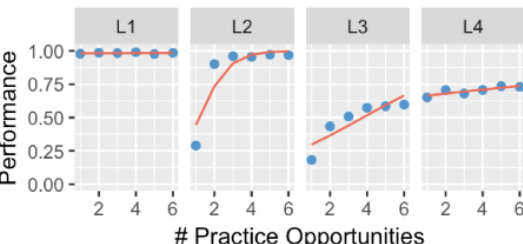
<p>The blue dots represent the percentage of students who got the first N puzzles correct on their first attempt, where \$N\$ is the number of puzzles that need to be completed correctly to pass the level. The red line represents the fitted learning curve. In the reported statistics for AFM, \$e\$ represents the easiness of each KC. A positive \$e\$ means that the level is easier (thus, higher probability of answering the question correctly on the first attempt) compared to a level with ease of learning 0. An \$l\$ represents the learning rate. The higher the \$l\$, the greater increase in accuracy with subsequent puzzle attempts. Thus, a learning rate close to 0 means that students barely improved with practice.</p> <p>The italic text describes the template used to generate puzzles in each level as explained in this section.</p>	<p>Game 1: Find Equivalent Fractions as Segments of a Horizontal Bar</p> <p><i>L1: randomly selected with ordering <1/2, 1/2, 1/3, 2/3, 1/?, 1>; L2: randomly ordered; L3: randomly selected with ordering <1/2, 1/10, 1/10, 1/10, 1/5, 1/5, 1/5, 1/5>.</i></p>  <p>AFM Accuracy: 0.806; AUROC: 0.659 L1(e, l): -2.86, 0.71; L2 (e, l): 0.08, 0.04; L3 (e, l): 0, 0</p>
<p>Game 2: Find Equivalent Fractions as Segments of a Vertical Bar</p> <p><i>L1-3: randomly ordered; L4: fixed order <1, 1/2>.</i></p>  <p>AFM Accuracy: 0.831; AUROC: 0.663 L1 (e, l): -0.02, 0.37; L2 (e, l): -2.34, 0.34; L3 (e, l): -1.12, 0.17; L4 (e, l): 0, -1.19</p>	<p>Game 3: Find Equivalent Fractions as Segments of a Vertical Bar II (a more difficult game with more details described in text)</p> <p><i>L1: randomly selected; L2-4: randomly ordered.</i></p>  <p>AFM Accuracy: 0.748; AUROC: 0.691 L1 (e, l): 2.10, 0.32; L2 (e, l): -0.53, 0.21 L3 (e, l): -0.22, 0.04; L4 (e, l): 0, -0.13</p>

Table 5.1 (continued).

<p>Game 4: Compare Fractions on a Number Line <i>L1-2: randomly generated; L3: randomly ordered.</i></p>  <p>AFM Accuracy: 0.806; AUROC: 0.700 L1 (e, l): 0.68, 0.52; L2 (e, l): 2.46, 0.31 L3 (e, l): 0, 0.48</p>	<p>Game 5: Locate & Find Equivalent Fractions on a Number Line <i>L1: randomly ordered; L2-3: randomly selected with ordering (described in later section); L4: randomly generated with ordering.</i></p>  <p>AFM Accuracy: 0.744; AUROC: 0.640 L1 (e, l): 1.10, 0.43; L2 (e, l): -0.75, 0.28 L3 (e, l): 0.36, 0.18; L4 (e, l): 0, 0.04</p>
<p>Game 6: Compare Two Fractions <i>L1-4: randomly generated.</i></p>  <p>AFM Accuracy: 0.833; AUROC: 0.572 L1 (e, l): -1.82, 0.27; L2 (e, l): -2.26, 0.32 L3 (e, l): -1.5, 0.09; L4 (e, l): 0, 0.16</p>	<p>Game 7: Order Three Fractions <i>L1-2: randomly generated; L3: randomly ordered; L4: randomly selected.</i></p>  <p>AFM Accuracy: 0.832; AUROC: 0.725 L1 (e, l): 3.54, 0.05; L2 (e, l): -1.03, 1.38 L3 (e, l): -1.78, 0.36; L4 (e, l): 0, 0.08</p>

orderings make no difference. Thirdly, for puzzles randomly generated from a template, we assume that ST Math used a randomization function that was uniformly distributed and that the ST Math designers created the generators so that the puzzles within a level are of considerable similarity. This means that in the analyzed objective, any specific number as the denominator or numerator in a fraction was no more difficult than other numbers as considered by the students. Such assumption makes sense, because the number zero is never used in any fractions and no randomly generated puzzles except for those in Game 7 L1 contain fractions bigger than one. Lastly, a few levels contain puzzles that are randomly selected or ordered following different templates within the same level. For these levels, applying learning curve analysis would pinpoint where students failed to capture the underlying concept and resulted in *disjointed learning curves* as described in the following paragraph.

We looked for four specific patterns to derive game-design feedback. A *good learning curve* displays a logarithmic pattern indicating that students increased their accuracy with practice and thus that learning appears to be well-supported. An *incomplete learning curve* is similar to a good learning curve except it does not include a flat tail. This indicates that students can still improve with more practice. Next, a *flat learning curve*, which we defined here as having a smooth curve with $e \leq 0.05$, means that students' performance did not improve substantially with practice. It could be that the level is too difficult or that the game content is not well-designed thus students did not learn with practice. It could also be that the level is too easy: students started with near perfect performance. Lastly, a *non-learning curve* does not follow a logarithmic or flat pattern. Performance in this type of learning curve increased or decreased suddenly at specific attempts. When this happens, it means that there was a change in the puzzle's template that introduced what the students perceived to be a new knowledge component. For example, a level can have half of the puzzles randomly generated with odd denominators, and the other half randomly generated with even denominators. If students failed to transfer the knowledge when denominators changed, we will see two *disjointed learning curves*.

The majority of levels (16 of 26) showed the logarithmic pattern of a learning curve, which means that the game design helped students learn (or improve their performance). A few learning curves are incomplete (e.g., Game 7 L3), suggesting that game designers should increase the number of puzzles required to pass the levels (L) as students still have room for improvement.

Four of the 26 levels had flat learning curves. Among these, Game 3 L3&4 appear too difficult and need to be re-designed to support learning. L3 presents a fraction and requires students to find two different ways of dividing a vertical bar by selecting the number of segments that equals the given fraction, as shown in Figure 5.1. However, the denominator of the given fraction is not allowed as a choice. For example, if the fraction is $3/4$, the option to divide the bar into fourths is grayed out, forcing students to divide the bar into eighths and choose $6/8$, and then $9/12$. L4 concerns a similar skill with more difficult content. A longitudinal study by Hansen et al. also found questions

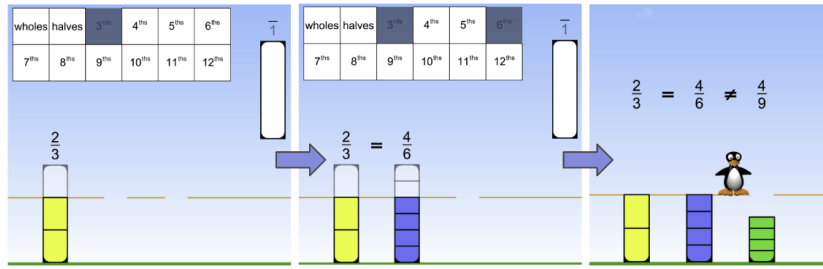


Figure 5.1 An Example of a Too Difficult Level with Flat Learning Curve.

where the denominator of the fraction did not correspond directly to the pieces in the presented model posed the most difficulties for low-achieving students [Han17]. Thus, designers may wish to offer extra scaffolding on these two levels. On the contrary, Game 1 L2 and Game 7 L1 appeared to be too easy because performance is consistently near perfect. For example, in Game 7 L1 students order fractions with the same denominator and different numerators, with visualizations showing the size of these fractions as the widths of bars. This level is too easy because students can, without understanding fractions, visually compare the bar widths. However, this level serves to teach the game mechanics for the subsequent levels in this game. Thus, we suggest designers either reduce the number of puzzles, or use one puzzle in this level as a tutorial for L2, instead of as an independent level.

Six levels follow a non-learning curve pattern. Two levels (Game 1 L1, Game 2 L4) present the same puzzles at the same attempt for all students to make a specific point, such as understanding fractions equal to 1. These non-randomized puzzles are easier, causing a jump in performance. However, we do not suggest changing them due to their educational value. Four levels showed disjointed learning curves. These learning curves revealed cases where students failed to transfer between specific numbers. For example, the first three puzzles of Game 5, L2 require students to locate a fraction $X/8$ on a number line divided into fourths, and the last three require locating $X/4$ on a number line marked with eighths. Game 5, L3 has a similar setup, with $X/6$ and $X/3$. Although these puzzles cover the same concept, the four disjointed learning curves show that students failed to transfer between puzzle sets of different fixed denominators (the four puzzles types shown in Figure 5.2). It could be that some students do not understand the underlying concepts, but learned pattern matching based on specific numbers. Another possible explanation is that the number of partitions prevented transfer between the two puzzle sets. Mitchell and Horne found that some students may incorrectly count the number of lines to determine where a fraction falls on the number line instead of considering the spaces between the lines [MH08]. Thus, game designers may consider providing practice with randomized numbers instead of presenting fixed numbers separately to reduce the ease of content-specific pattern matching or counting.

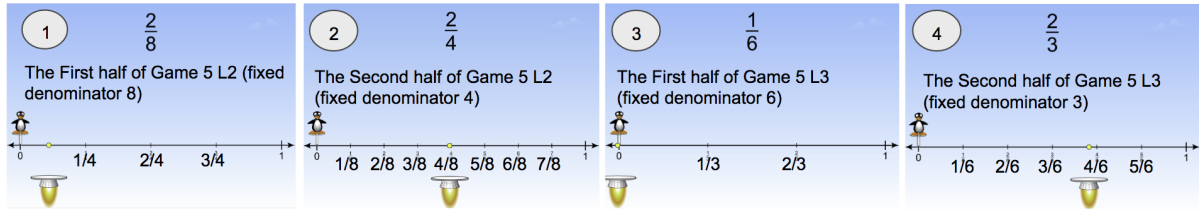


Figure 5.2 An Example Where Students Failed to Transfer. The Above Four Types of Puzzles Showed Four Disjointed Learning Curves.

To summarize, fitting learning curves to each game’s level derived specific game design insights, and helped us better understand how learning is structured in the ST Math environment. However, a general trend was that a number of the learning curves were disjointed. In many games, each level seemed to form its own learning curve instead of forming a single learning curve with the other levels. The lack of connection and the cases of disjointed learning curves implied that 3rd graders may rely on content- and problem-specific procedural knowledge or pattern matching strategies to solve puzzles, instead of transferring the understanding of underlying math concepts [RJ01; LH06]. When new content or problem representations were introduced to practice the same math concept, students treated them as new knowledge components.

5.4.2 Analyzing Levels in Games

In this section, we sought to find similarities between levels by looking for level combinations that would form a continuous learning curve. Based on the previous analyses, we started by considering each level as a separate Knowledge Component (KC), and applied a bottom-up approach to hierarchically combine levels within a game based on learning curve fitting. We then searched for KC pairs that, once combined as the same KC, yielded AFM models with the lowest Bayesian information criterion (BIC) values as compared to other combination choices. A lower BIC means that the model fit was comparatively better considering both the fit (maximum likelihood) and complexity (number of parameters). This approach is similar to Cen and Koedinger’s work [Cen06], except that they used a top-down approach that searches to split one KC into multiple KCs to improve the model. We applied this method to levels within a game instead of across games. We did so because the conjunction of learning curves only makes sense if the different levels involve practicing the same skill (KC). We excluded levels with disjointed or flat learning curves. This is because disjointed learning curves, if split into multiple learning curves, have too few puzzles to study a pattern. Flat learning curves, especially those with high performance, may be appended to the tail of any previous learning curves that reached high performance in the last attempts. Such conjunctions may not have empirical meanings.

Figure 5.3 shows the hierarchical combinations of levels in Game 6. The algorithm first combined

L1&4, which resulted in a lower BIC than the original model in which each level is a different KC. Both L1&4 ask students to compare two fractions with the same denominator and different numerators, with one requiring students to answer with a ladder and the other with math symbols. Similarly, L2&3 require comparing fractions of the same numerator with ladders or math symbols, and were suggested to be combined next. This hierarchy showed that students can transfer easily between ladder and math symbols, but not as easily from comparing fractions with the same denominator to comparing fractions with the same numerator and different denominators. This is likely due to the simplicity of comparing fractions with the same numerator (or denominator)—students only have to compare one number rather than considering the relationship between the numerator and denominator [RJ09]. In other words, comparing fractions was still a difficult skill and it wasn't the symbolic representation of greater than/less than that tripped students up. Thus, we suggest re-positioning L4 to come after L1 or even removing ladders and using math symbols only. However, when introducing comparing fractions with the same numerator and different denominators in the next level, designers should provide other scaffolding to make connections between the two skills.

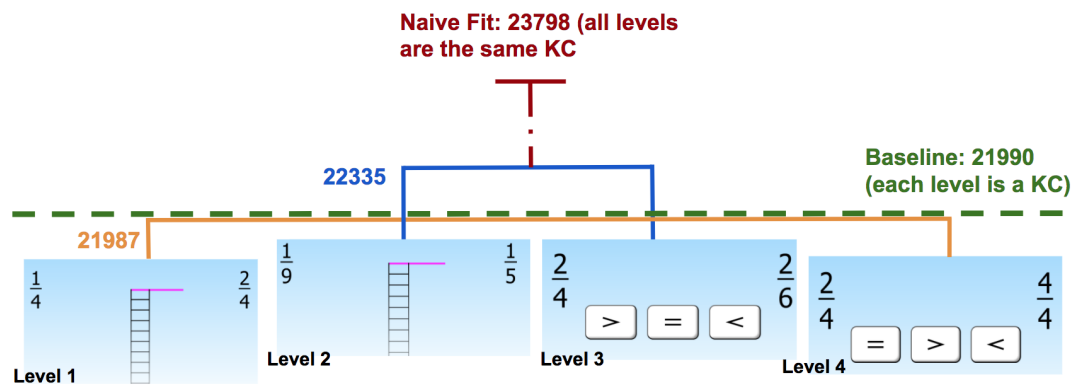


Figure 5.3 Hierarchical Combinations of Game 6 Levels that Led to Models with Different BICs. BIC Similar to Baseline Indicates that the Combined Levels Share a Similar KC.

In the rest of the games, the algorithm suggested combining the following levels without much increase in BIC: L2&3 of Game 2 (BIC 14138 compared to baseline 14127), L1&2 of Game 4 (BIC 14731 compared to baseline 14702), and L3&4 of Game 7 (BIC 14425 compared to baseline 14352). The combination in Game 4 shows that by locating fractions one by one on a number line, students can easily transfer from comparing fractions with the same denominator and different numerators to comparing fractions with the same numerator and different denominators. This implies that locating fractions on a number line is a strategy that facilitates such a transfer (see the integrated theory of numerical development in [Rin17; Sie11]). However, for comparing fractions with the same

denominator, the transfer between fractions smaller than 1 to larger than 1 is much more difficult. It could be that students rely on the numbers in the fractions rather than their magnitude [Han17; Jor17], so when the fraction is greater than 1, they do not have a conceptual understanding of where the fraction is located on a number line in order to use the number line to help with comparisons. Thus, designers may consider more scaffolding to help students make this transfer.

The other suggested combinations are pairs of levels concerning the same skill with the same problem representation. To summarize, hierarchically combining learning curves helped us identify where students may need extra support to transfer between math skills and problem representations.

5.4.3 Analyzing Games in Objective

From previous analyses, we learned that the performance in ST Math is influenced by both targeted math skills and problem representations. Thus, in this subsection, we sought to further investigate how math skills and problem representations interact with each other and influence students' learning across games. We started with designing an expert Q-matrix (mapping from puzzles to knowledge components) with two types of knowledge component: math skill (KC-S) and problem representation (KC-R). Each level was mapped to at least one KC-S and one KC-R, but not all KCs were mutually exclusive, which means a level could contain multiple KC-Ss and/or KC-Rs. We constructed a total of four KC-Rs, including: number line, vertical bar, horizontal bar, and representation containing visual cues that help students solve the puzzle through pattern-matching. We constructed 6 KC-S, including: presenting fractions (e.g., as segments of a bar); finding equivalent fractions; comparing three types of fractions: same numerators, same denominators, and different in both numerators and denominators; and comparing fractions greater than 1.

Next, we fit learning curves on the expert-designed Q-matrix with three assumptions. The **first** assumption was that each KC contributed to performance *additively*, as assumed by the AFM model. The **second** was that each KC contributed to performance *conjunctively* (modeling the conjunctively as a multiplication of skill parameters), as assumed by the CFM model. We used R's `optim` package with BFGS optimization method to estimate the parameters of CFM through maximizing the likelihood function. The **third** assumption was that the KCs interacted neither additively nor conjunctively. This means the same math skill presented in two separate levels would be viewed as two distinct skills that students learned under different representations. Thus, a *combination* of KCs forms a new KC. In other words, each level would be mapped to only one KC, and levels with the same KC-R and KC-S combinations shared the same KC. When each level had one KC, AFM and CFM were equivalent. This assumption yielded 15 KCs (15 different combinations of KC-R and KC-S) across 26 levels.

As shown in Table 5.2, our models have low discrimination ability based on AUROC from 10-fold cross-validation. Thus, there are limitations when applying traditional learning models to serious

Table 5.2 Learning Models Under Different Assumptions of KC Interactions.

KC Assumptions	Model	Accuracy	AUROC	BIC
High Baseline: Each level is a KC (26 KC, each level has 1 KC)	AFM/CFM	0.782	0.622	134831 (+0.0%)
Combined (15 KC, each level has 1 KC)	AFM/CFM	0.775	0.602	137425 (+1.9%)
Additive (9 KC, each level has >1 KC)	AFM	0.768	0.58	141607 (+5.0%)
Conjunctive(9 KC, each level has >1 KC)	CFM	0.761	0.54	149002 (+10.5%)
Low Baseline: Each game is a KC (7 KC, each level has 1 KC)	AFM/CFM	0.756	0.54	150787 (+11.8%)

games, where it is easier for children to guess or pattern match with visual cues in specific game environments. The model based on the third assumption had the best fit, with only 1.9% increase in BIC as compared to the most ideal model assuming each level is a KC. Our result implies that ST Math’s targeted skills and problem representations do not contribute to performance simply through additive or conjunctive relationships. Instead, the same skill would be treated as different skills when combined with other skills and problem representations. It could be that when students play ST Math, they do not use math skills alone. Rather, students develop new skills that are content- and situation-specific, based on the combinations of targeted math skills and problem representations.

5.5 Contribution

In this study, we demonstrated using learning curves as a simple, efficient way to evaluate how well an educational game supports learning. Our results pinpointed problematic levels and cases where students needed extra support to transfer between math skills, content, and problem representations. We derived actionable feedback for ST Math and general insights on fraction learning.

This work has several limitations. First, ST Math is designed as a curriculum-integrated game, but our data does not capture factors in classrooms. Future research will include teacher interviews and classroom observations to better assess the impact of classroom factors. Second, we limited the data to only the number of puzzles required to pass a level which excluded some attempts by low-performing students. Future research may explore methods to separate learning curves for student sub-populations [Mur13] in order to increase external validity.

Our results suggest that students developed new ‘skills’ based on the combination of targeted math skills and problem representations, rather than simply combining them as assumed in the additive or conjunctive factor models. The variety in ST Math’s problem-solving scenarios may

improve students' understanding of math concepts. However, this variety could distract learning if students focus more on content- and situation-specific practice than the underlying math concepts. The literature review by Lehtinen and Hannula-Sormune [LH06] argues that in cases of transfer failures, the (new) situations are not necessarily interpreted as mathematical by children. For example, students may see Game 1-3 as 'selecting divided bars and the number of segments to match a given bar's height/width,' instead of 'understanding fractions as proportions and finding equivalent fractions by multiplying or dividing the numerator and denominator by the same number.' Thus, when bars are replaced with a number line to practice the same math skill, it becomes a different task. Similarly, work by Rau et al. [Rau09] found that providing multiple representations can promote better learning than a single graphical representation, but only when students are prompted to self-explain how the graphics relate to the symbolic fraction representations. With the increasing popularity of mini-games collections, it is important to design scaffolding that facilitates transfer by focusing students on the underlying math concepts instead of reinforcing simple strategies like pattern matching. Such scaffolding should also be considered in other e-learning platforms that offer multiple problem-solving scenarios for young children.

We learned several lessons from applying learning curves to this serious game environment. Learning in game environments is inseparable from the games' mechanisms, structures, and designs. Researchers should consider starting analyses at a low granularity, such as the individual level we used here. Understanding students' learning at a low granularity would help illuminate factors that contribute to learning, and help structure analyses at higher granularity where these factors may combine or evolve. Moreover, game performance does not solely comprise learning. This means traditional learning modeling methods may have limited power in serious games. Thus, researchers should be flexible with different models and assumptions to work within specific game environments. Regardless, researchers should triangulate results with human interpretations and the literature to make sure the results do not derive from unforeseen game scenarios or the large amount of data. For example, Harpstead and Alevan [HA15] used both data and human judgment to examine the fit of learning curves in a physics game and identified an unforeseen pattern matching strategy. Liu et al. [PL17] mined predictive relationships between ST Math objectives, and used both human interpretation and literature to suggest game design feedback. Thus, when analyzing serious game data, it is extremely important to not solely focus on the performance of models, but also consider the models' interpretation and practical value.

CHAPTER

6

STUDY 4: TEACHER-FOCUSED FIELD STUDY ON THE CLASSROOM USE OF A CURRICULUM-INTEGRATED GAME

Peddycord-Liu, Z., Catete, V., Vandenberg, J., Barnes, T., Lynch, C., and Rutherford, T. 2019. "Jiji Dying Makes Them Very Upset" and Other Insights: A Teacher-centered Field Study of a Curriculum-integrated Digital Math Game. In submission to the ACM CHI Conference on Human Factors in Computing Systems.

Abstract This work presents a new framework describing how teachers use ST Math, a curriculum-integrated educational game, in 3rd-4th grade classrooms in a Southern U.S. district. We combined authentic classroom observations with teacher interviews to identify needs and practices of teachers using ST math. Our findings extended and contrasted with prior work by identifying critical differences that arise from aspects of long-term use and curricular integration. Based on our findings, we suggested practical ways that curriculum-integrated games can be better designed to accommodate teachers to support effective classroom practice.

6.1 Introduction

Although educational games are becoming more popular, more research is needed on how best to integrate games into real-world classrooms. One recent literature review concluded that the crucial role of teachers has been neglected or marginalized, in both research and in game design [Mol17], with most research focusing on barriers that prevent teachers from adopting educational games [San06; Mol17; Che14; Mif13; ES13; Lim11; DT10]. There is limited focus on the practices of using educational games (e.g., [Wat11; ES13; Kan17b; Nan18]), especially for curriculum-integrated games that are intended to be used throughout an entire year ([Cal18]). There is a critical need to understand how games are used in classrooms, and to use this understanding to inform game design and benefit student learning.

In this paper, we studied the classroom use of ST Math—a curriculum supplement game used by over one million students in hundreds of schools. We conducted 43 field observations of the classrooms of eight teachers with varied levels of experience with ST Math. We combined these with semi-structured interviews to explore teacher use in detail and derive suggestions for game design. We were specifically interested in understanding how teachers integrated ST Math into their classrooms and teaching practice, and how teachers orchestrated student use of the game.

6.2 Background

Chapter 2 describes a more extensive literature review related to teachers' use of educational games. This section gives a brief reminder of the main findings of the related work. A limited number of prior works have investigated teacher's use of educational games, as previously described in Chapter 2. Among these studies, Kangas et al. conducted a literature review and presented a framework that summarizes teacher activities for classroom use of digital and non-digital games into five categories: planning, orientation, playing, elaborating and reflecting [Kan17b]. However, Kangas et al. pointed out that many studies included in their work lasted only a few sessions long [Kan17b]. There were open questions on applying this existing framework to a large-scale, curriculum-integrated games like ST Math.

Prior work has also conducted controlled experiments and suggested that digital games are more effective when combined with structured classroom activities, such as focused discussions to connect game content with classroom content [Bak15; Row17; BB11]. However, these studies focused on evaluating pre-designed conditions instead of understanding authentic classrooms and teacher practices.

Other researchers have conducted case studies, surveys, and interviews to investigate game use in authentic classroom settings [Wat11; ES13; Nan18; Cal18]. However, these studies have not focused on finding patterns across teachers or on discussing the results from a game design perspective.

There is a need to conduct field studies on teachers' use of ST Math in authentic classrooms, to expand upon prior theory and to inform the design of large-scale, curriculum integrated games.

6.3 Method

This study focuses on one school district in the southern U.S., where ST Math is a mandatory intervention. Principals at all 81 elementary schools were invited to participate, and seven volunteered within a two-week recruitment period. All 3rd and 4th grade teachers at these schools were invited to complete a screening survey. Eight female teachers from six schools responded to and participated. Teachers' experience levels with ST Math fell into three categories, High (**H1-4**), those who have been teaching with ST Math since it was first introduced in the district over six years ago, Moderate (**M1-2**), those using ST Math for 1-3 years, and Low (**L1-2**), those using ST Math for less than a year. L1 recently transferred to the county, and L2 was a first year teacher.

We took a grounded theory case-study approach to understand the experiences of the eight teachers, specifically focusing on their use of ST Math's teacher features and their classroom integration of ST Math. We involved different perspectives in this process. The first and last authors have had years of experience researching ST Math and experience playing the game as students, manipulating the game as teachers, and viewing teacher resources. The remaining authors and field observers come from education and computer science, with varied experience in educational research, learning technology design, and K-12 outreach. The third author is an experienced teacher for the same age group.

Initial classroom observations were followed by a teacher interview and subsequent observations. Our goal of classroom observations was to experience and record how ST Math use varied across teachers and different classroom formats, to inform our interviews and later analyses. We conducted a total of 43 observations, with 2 teachers being observed 4 times, 1 teacher with 5 observations, and 5 teachers with 6. Classroom observations were conducted in teachers' regular math sessions with planned ST Math use. We ensured that these observations covered diverse scenarios of using ST Math both in class and in lab. Each observation lasted from 20 minutes up to an hour, depending on class schedules. The observers were trained on the classroom observation protocols to improve consistency between observers. In most observation sessions, we assigned one observer to focus on the teacher and the remaining observers focused on students. The teacher observer took field notes on the classroom set-up, the major instructional activities, and detailed all of the interactions between students and teachers during the ST Math session. The student observers recorded student affect and behaviors and any interactions of note. After each session, observers conferred and noted interesting results, relating them to prior observations, literature, and theory. We rotated observers for the same teacher to facilitate diverse perspectives.

We conducted teacher interviews after several observations to facilitate a meaningful conversa-

tion about the teacher's perspective and classroom context. Each teacher participated in a 30-50 minute semi-structured interview after school, during which we incorporated questions based on our field observations, with field observers present at the interviews. The first author, who conducted the interviews, observed at least one session of each class format for each teacher before interviews. We informed teachers that the data would be anonymized and researchers were not ST Math stakeholders. During the interview, we asked teachers about their activities and rationales before, during, and after the use of ST Math in the classroom, their perceptions of the students' use of ST Math, and their suggestions to improve the system. The semi-structured interview facilitates a guided discussion where the interviewer can help teachers expand their thoughts and reflect on their decision-making. The interviews were audio-recorded with strategic note taking to help with later interpretations, and were transcribed verbatim.

To establish a better understanding of actual classroom use, we conducted grounded theory analyses of teacher interviews following Saldana's method of two rounds of coding, where the first round is used to identify and describe codes, and the second round is used to identify common patterns [Sal15]. We carefully used our field notes and experiences to understand the context for each interview. The interview transcripts were coded by the first 2 authors who were also field observers. The first author, who was also the interviewer and primary teacher observer, compiled and verbally summarized the field notes to establish the context, evaluate teacher statements, and facilitate discussion between the two coders. Initially, the researchers coded two interviews independently, using four types of coding as described by Saldana: *structural coding* to relate data to specific research questions, *process coding* to identify activities and actions, *descriptive coding* to summarize each statement, and *evaluation coding* to describe teachers' comments and feedback on ST Math. Then the researchers collaborated to refine codes into a codebook. Next, each coder independently coded three different interviews, checked each other's codes, and then met to discuss and collaboratively update the codebook. The coders then collaboratively conducted *pattern coding* [Sal15] to identify developed themes and patterns across interviews.

6.4 How were Teachers Using ST Math?

From classroom observations and interviews, we identified four types of activities related to teacher's classroom use of ST Math: *Preparation*, *Integration*, *Intervention* and *Data-informed Practice*. In Table 6.1, we have aligned these with activities identified in Kangas et al.'s literature review on teacher activities surrounding digital and non-digital games: *planning*, *orientation*, *playing*, *elaboration*, and *reflection* [Kan17b]. Our work both extends and contrasts with the Kangas et al.'s game-based learning framework, highlighting key differences that arise from a school-year-long curriculum-integrated digital game, rather than short-lived games whose content may not directly support curricular standards. In the following sections, we describe each type of teacher activity, connect

these activities to prior literature, and recommend game designs that could better support teachers and students using curriculum-integrated digital games.

Table 6.1 Categories of Teacher Activities to Orchestrate a Curriculum-Aligned Digital Game in Classrooms.

<p>Preparation</p> <ul style="list-style-type: none"> - Observe Student Gameplay - Use Online Resources - Self Gameplay - Attend Professional Development 	<p>Integration</p> <ul style="list-style-type: none"> - Set up for Classroom Formats [<i>planning</i>] - Foster Gameplay Culture [<i>planning</i>] - Order the Curriculum [<i>planning</i>] - Connect Game Content with Classroom Content [<i>planning, orientation, elaboration</i>]
<p>Intervention</p> <ul style="list-style-type: none"> - Cognitive Intervention [<i>playing</i>] - Affective Intervention - Organize Peer Tutors [<i>playing</i>] - Progress Management [<i>playing</i>] 	<p>Data-informed Practice</p> <ul style="list-style-type: none"> - Identify Struggling Students Combining in- and out-of-game Assessments - Identify Math Content to Review/Reteach - Communicate with Students [<i>elaboration</i>]

6.4.1 Preparation

Preparation is the work that teachers do to get ready to use a curriculum-integrated game. In this category, teachers themselves learn about ST Math to prepare for using it in classrooms. This preparation category was not explicitly identified by Kangas et al [Kan17b]. We identified three major activities in this category: observing students' gameplay, playing the game themselves, and attending Professional Development (PD). All teachers stated that they relied heavily upon observing students' gameplay to learn the game. This is supported by our classroom observations, showing that many teachers asked students to explain what they were doing as students played. Three highly experienced teachers also learned by playing the game, with H3 in charge of ST Math at her school, and H1 and H2 playing ST Math with their children at home. The remaining teachers stated that they relied mostly on figuring it out during students' gameplay. H1 and H4 mentioned that having previous experience using ST Math helped them accumulate knowledge across years.

Teachers who recently attended PD (M1-2, L2) commented that the required county PD was too basic. For example, M1 stated that the county PD was primarily about "how the game looks," with "very little about the reports, or like the teacher information that's on there." L2 noted that she played the game during PD but didn't know "how to manipulate" and "why we use it." L1 had recently

transferred into the district and missed the PD. On the other hand, more experienced teachers stated that MIND's supplemental online resources (accessible from a 'ST Math Central' button) were effective, but the two new teachers, L1 and L2, were not aware that this resource existed.

6.4.1.1 Design Implication

Our results support prior research on educational technology in classrooms, which concluded that both teachers and students continuously help each other in understanding and using the game [Nan18]. This could be especially true for games with large amounts of content that are intended to be played over a long period of time. Our findings also back up a previous ST Math study [Cal18] and literature review of broader ICT training [BA12] that teachers need PD that focuses on the purpose of the game and game integration practices, instead of specific interface elements that can be learned from observing students. We reiterate Giroto's recommendation that teacher preparation should be integrated into their existing workflows with educational software, providing teachers just-in-time information needed to make decisions and support students [Gir16]. For example, the software could require that teachers watch mandatory short videos as soon as they register in ST Math. Moreover, as the two new teachers expressed ignorance regarding the 'ST Math Central' button, it may be helpful to have other mechanisms to help teachers notice where they can find and re-access PD resources.

6.4.2 Integration

Integration is the work that teachers do to fit the the game into the classroom. Our Integration category closely aligns with Kangas et al.'s *planning* category, where teachers create a pedagogical frame to prepare students to learn, organize gameplay situations, and fit the game into the curriculum [Kan17b]. Based on our analyses, ST Math teachers conduct the following Integration activities: organize gameplay into different class formats, foster a gameplay culture, reorder ST Math objectives, and connect the game content with math content. The last activity is also related to Kangas et al's *orientation* and *elaboration* phases, but we did not observe these activities directly before or after the gameplay sessions.

6.4.2.1 Class Format

We observed three class formats that each appeared to significantly impact gameplay interaction: lab-seating (18 observations, 6 teachers), free-seating (9 observations, 3 teachers), and rotation-seating (16 observations, 5 teachers). In *lab-seating*, all students played ST Math during the whole class, by sitting at designated spots in a lab-like environment. Teachers could clearly see students' screens, and were able to pay closer attention to ST Math gameplay as compared to other formats. In *free-seating*, students played ST Math the whole time, but were free to move around the classroom, sitting

as individuals or in groups, using tablets or laptops. Student-to-student interactions were more frequent, and teachers had to move around continuously to see all students' screens. Consequently, students who sat by themselves, sat in corners, or were reluctant to seek teachers' help, were often ignored. In *rotation-seating*, the students were split into groups with one group doing ST Math while the other group(s) did other activities, such as engaging in teacher-led instruction. Students rotated groups after a set time. In this format, the teachers were often busy with students doing other activities, and the "self-sufficient" ST Math groups rarely received attention. Our interviews made clear that the class formats were influenced by school policy, scheduling, and the availability of technologies in classrooms.

In interviews, teachers reflected on the pros and cons of different class formats. M1 and L1 liked that the rotation-seating allowed them time to focus on students who needed help by putting the rest on ST Math. H2, H4, and L2 believed that students got more chances to work with each other when left on their own. For lab- and free-seatings, M2 liked that students got more time to make sustained progress in ST Math. This is because when students quit before completing a level, they must restart it, losing any progress they have made and their remaining lives. Restarting happens frequently in shorter gameplay sessions like those found in rotation-seating. Children were observed to be frustrated due to losing lives or progress, and multiple teachers (H1-2, H4, M1-2, L1) were observed giving warnings several minutes before gameplay ended, specifically instructing students who made above 50% progress to continue and the rest to quit. H2, M2, and L2 appreciated that they were able to provide more help in the lab- and free-seating formats. However, two highly experienced teachers (H1, H2) believed in having a variety of formats because students learned different things from more teacher-centered and more student-centered classroom formats.

6.4.2.2 Foster Gameplay Culture

Teachers viewed a motivating and student-centered culture as crucial to ST Math success. Because ST Math is curriculum-integrated, teachers foster a culture around its use, rather than framing the game during each classroom session. Some teachers (H4, M2, L1-2) encouraged more quiet, focused, and independent work, whereas others (H1-3, M1) allowed students to communicate with each other and/or move around. Teachers fostered gameplay culture based upon their own beliefs about how students learn, and how specific student groups "function." Some gameplay cultures were observed across classrooms. Many teachers (H1-2, M1, L1) encouraged students to share their progress on ST Math, to compete with one another, and to celebrate their achievements. To foster this culture, many teachers (H1, M1, L1) incentivized students by giving small prizes, and they perceived this technique to be effective. All teachers valued student autonomy, encouraging students to use classroom resources like manipulatives or the whiteboard and to notify teachers when they need help. H1 stated that students love ST Math due to "the environment that we have

built about the importance of ST Math."

Because the county requires ST Math usage, one common factor among classrooms was a focus on making significant progress (levels completed) in ST Math. The county had previously set an expected weekly progress threshold, against which teachers are evaluated based upon their students' average performance. All teachers communicated about progress with students through classroom posters and/or during gameplay. However, it is not clear if progress is a measurement that benefits students' learning, as students were observed and described to 'rush through' resulting in low performance in the game. All teachers but L2 reported they frequently assigned objectives as homework, to give students access to gameplay at home. Teachers did so primarily to promote progress and help students catch up with the class. H2 mentioned an instance of healthy competition between siblings to complete ST Math at home. H2 also mentioned that by matching homework to the curriculum, parents could be involved in good conversations about math learning.

Although all teachers valued setting a classroom culture and were required to promote ST Math progress, only two highly experienced teachers (H1, H3) purposely fostered a culture of playing the game with a goal of reaching "productive struggle." Productive struggle, "the kind of effortful learning that develops grit and creative problem-solving," [Kan] is a key pedagogical goal of ST Math [War15; Rol14]. H1 described "never give up" as a mindset that she wanted students to have, and we observed her praising students for engaging in productive struggle. H3 stated that the students knew when she would "not help them" and instead let them solve problems by themselves. From our classroom observations, it was not clear that productive struggle was sufficiently emphasized, as many students raised their hand as soon as they got wrong answers. H1, who attended MIND's PD when ST Math was first launched, wrote in her survey response that "Over the years, [the focus on productive struggle] seems to have shifted away," and expanded in her interview that many teachers "front-load the information" instead of letting students figure out how to learn from the game on their own. H2 commented that sometimes she "do[es] too much when we are in lab" and appreciates how the rotation-seating class format forces her to "take a step back."

6.4.2.3 Order the Curriculum

ST Math objectives are high level topics such as 'fraction' and 'division.' ST Math comes with a pre-defined objective sequence, but the county has produced its own recommended ordering based on the local math curriculum. Some schools specify orderings for certain grade levels, and within the game, teachers have the flexibility to reorder objectives for their individual classes. During our interview, L2 stated that the school enforced some specific reorderings to ensure that difficult objectives (e.g., fractions) were played at the same time as the concepts were taught in math class, or were covered before the state assessments.

We noted differences between the highly experienced and new teachers. The highly experienced

teachers (H1-4) stated that they frequently reordered objectives to match what they taught in class. They believed that by practicing the same math concept twice (once in the classroom and once in ST Math), students would benefit from “visually understanding” and “reinforcing” what they learned in class. H2 stated that besides reordering “based on common sense like what naturally builds on each other,” she frequently reordered objectives to “keep it fresh” even when some students did not finish the current objective. She believed that in this way the students could “hit every strand” in ST Math without necessarily completing the curriculum, which would better prepare them for the next grade. In contrast, M1 referred to reordering as more a school’s grade-level decision. M2, L1 and L2 rarely reordered with M2 stating that “you don’t want to mess it up.” M1 stated that some teachers in her school didn’t understand that they could reorder the objectives.

6.4.2.4 Connect with Classroom Content

Except for a special case of a MIND research study in H1’s classroom, we did not observe teachers discussing specific game content or its purpose before or after gameplay, as with the Kangas et al.’s *orientation* and *elaboration* activities [Kan17b]. This may be because our study was conducted near the end of the year, when teachers already finished teaching curriculum content and students were spread out in their ST Math progress. In the interviews, however, some of the more experienced teachers did reflect on using ST Math to assist their classroom teaching. Regarding introducing the game’s purpose, H2 stated that during the first half of the school year, she frequently used ST Math to introduce math concepts, believing that it helps children understand new concepts before they must learn new vocabulary; M1 stated that she used ST Math’s teacher mode to demonstrate puzzles as a whole-group activity. Regarding discussions after gameplay, H1 stated that she instructed the whole class or a small group on objectives where she saw “a lot of the students” struggling. H4 reflected on paying attention to students’ progress in ST Math relative to their instructional progress, and took care to draw connections between the in-class math content and the topics covered in ST Math. During the discussion of game content, many teachers reflected on ST Math objectives that were difficult for many students, such as fractions (H2, H4, M2, L2) and place-value (H1, H2, H3).

6.4.2.5 Design Implication

Our results from interviews and classroom observations suggest that explicit supports for classroom formats, fostering gameplay culture, and curriculum ordering could be beneficial for curriculum-integrated games. This suggestion is in line with Dillenbourg and Jermann’s theory that teachers serve as “orchestrators” when using technologies in the classroom [DJ10]. One important aspect of being orchestrators is planning how and when students learn within the allotted classroom formats and times, which as we observed, may not be under direct teacher control. Thus, teachers could benefit greatly from ST Math configurations and suggestions that support lab-, rotation-,

and free-seating formats. We also suggest that ST Math could provide teachers with visualization tools such as a dashboard [Dia17a], real-time augmented reality monitoring [Hol18], or alternative communication channels such as email or text alerts when a student needs help.

Our *integration* results also showed that ST Math teachers were sensitive to the impact of lost progress on students, both when Jiji loses lives, and when students must quit due to time constraints. We recommend that ST Math should design to allow students to save current progress before logging out. This could reduce teacher and student frustration when they are forced to pause due to rotations or time constraints. This is also supported by Dillenbourg and Jermann's theory that teaching time is segmented into time slices and a technology that works well must orchestrate the time issues and provide teachers with flexibility in time [DJ10].

Our results on fostering gameplay culture reveal that teachers were not equally prepared to support the designed intent of ST Math to promote productive struggle. However, experienced teachers in our study, and in other research [Kha17], recognized the importance of class culture in the successful adoption of non-traditional teaching approaches. Therefore, we recommend that curriculum-integrated games should explicitly build in supports that promote the target classroom culture and mindset, such as in-game awards for progress, or encouraging students to try again upon first failures.

Our results suggest that only experienced teachers reorder the ST Math curriculum. Previous research shows that reordering could benefit learning, both when teachers reported doing it to tailor their classroom [Cal18], or when it is evaluated based on student data [PL17]. Like ST Math, a curriculum-integrated game should be designed to have appropriate curricular sequencing, and allow teachers and districts the flexibility to adapt these sequences. On the other hand, as beneficial re-orderings are discovered, they could be recommended within such a system.

The lack of orientation and elaboration before and after gameplay suggests that teachers are more likely to use ST Math to conveniently assist their classroom teaching, as found in prior work [Nan18], rather than designing teaching centered around ST Math. 'Ideal' game-based learning lessons [Kan17b; Wat11; ES13] that incorporate these lead-in and reflection discussions require time to plan, familiarity with the game and the curriculum, and knowing the current knowledge and skills of most students in the class. Achieving all three of these is difficult for large-scale games that are intended to be played year-long. Teachers have little time to plan or learn the game, some may be new to the curriculum and teaching, and some may not have access to appropriate PD. ST Math is also self-paced, making it difficult to gauge a whole classroom. Thus, we argue that game designers should be aware of these limitations, and design tools that could help teachers achieve 'ideal' integration for specific difficult concepts. Shorter, high-level introductions to specific game content within an objective, and salient mappings to the standardized curriculum content could help teachers more quickly and consistently learn about games and prepare to connect them to the classroom. Rather than reports geared only towards grading, curriculum-integrated games could

also help teachers plan for teaching by suggesting common misconceptions or difficulties based on gameplay performance.

6.4.3 Intervention

Intervention is what teachers do to interact with and support students during gameplay. We identified four types of teacher interventions from interviews and classroom observations: cognitive intervention, affective intervention, peer tutoring, and progress management. This category resembles the *playing* category identified by Kangas et al. [Kan17b] where teachers actively serve as tutors, facilitators, and organizers during gameplay. However, we only observed such teacher activities with a few students at a time. We did not observe teachers pausing the game to seize the ‘teachable moments’ [Wat11], probably because of the diverse classroom activities and levels of the students at the time of observation. Importantly, we observed a great deal of one type of teacher activity not captured by Kangas et al.—affective intervention. Both in classroom observations and teacher interviews, teachers clearly felt that affective support was critical for their grade 3-4 children playing ST math.

Another important new contribution this work makes is to identify how the intervention patterns we observed were directly related to the previously-described classroom formats. In lab-seatings, ST Math students got more help because teachers could pay more attention to them and could conveniently see their computer screens. In free-seatings, the raised hand indicator clicked by students to ask for help, frequently went unnoticed because teachers could not easily see student screens. Teachers were also more likely to check on and help students who were sitting in large groups, or those seated near the center of the classroom. In rotation-seatings, teachers seldom intervened with the ST Math group unless those students exhibited distracting behaviors. Although the observers were not as familiar with the individual students’ proficiency levels or personalities, we generally concluded that the students who sought help, expressed frustration verbally, or were obviously off-task, got more teacher attention than students who struggled quietly.

6.4.3.1 Cognitive Intervention

ST Math is specifically designed to help students learn math concepts through problem-solving using visual-spatial reasoning skills and the animated feedback. Our teacher observations and interviews showed that teachers realized the affordances of the game and made specific efforts to help students benefit from problem-solving. In noting how students could learn from ST Math, M2 stated that “it’s really about if they can be observant” to ST Math’s animated feedback, and teachers often reminded students to watch Jiji in the animation. However, we observed that many children went off-task during the 5-10 seconds-long animations after each answer. H3 stated that younger children love these animations but older children just “look around.” Teachers themselves stated that they needed

to watch Jiji fail a couple of times before understanding how to help. Although teachers often told students that they were learning together, we did not often observe teachers explaining how they learned from watching the gameplay or animations. When we observed teachers providing help, their cognitive interventions focused on guiding questions and scaffolding by breaking problems down into smaller steps.

Teacher interviews showed that teachers recognized that learning in ST Math was more student-centered than traditional classroom teaching. Both classroom observations and teacher interviews revealed that teachers encouraged students to find resources to support their own learning, such as consulting peers, using classroom resources such as multiplication charts, or using pencil and paper. In classrooms, we commonly observed students transferring ST Math problems into other familiar formats using classroom resources, even without teacher prompting. For example, one student was observed struggling with an ST Math puzzle representing fractions on a numberline; after several failures, the student decided to use paper and pencil to re-represent the fraction numbers with a pie chart and, after understanding the proportions that way, progressed through the level.

Classroom observations showed that children were often impatient. We observed many teachers (H2, M1, L2) telling students to “slow down and think.” Students often ignored animated feedback, and rushed through pre- and post-tests. Many students asked for help as soon as they got a wrong answer, with some asking teachers to “just give me the answer.” Such help seeking, in which students ask teachers for help before attempting problems on their own, has been observed in classrooms across contexts [Arb98], and has also been related to the exploiting of help in intelligent tutors [Ale04; Bak08].

6.4.3.2 Affective Intervention

We observed that children often needed help in regulating emotions during gameplay. We often observed frustration leading to off-task behaviors, some of which disturbed the classroom activities. In teacher interviews, **all teachers** mentioned frustration when describing student gameplay. Many teachers (H4, M1-2, L2) stated that it was difficult for children to work out such frustration without teachers’ affective interventions. Teachers (H2, H4, M1, L1) stated that children at times said “the computer is tricking me” and they needed to explain to children “Jiji is a very precise penguin.” Children were also very afraid of Jiji dying, as observed across classrooms and heard from interviews. For example, H4 intervened a student “what happens if Jiji dies? Nothing! You got to try this again!”

From our observations and interviews, teachers believed the importance of affective support, helping students regulate their emotions, so they can benefit from the game. Teachers stated the need to distinguish between students who needed cognitive and affective support. Affective help was commonly used to intervene with students who “need[ed] attention,” needed a “security blanket,” needed a “constant reminder” to stay focused, or were too frustrated to proceed. Teachers also

intervened by physically moving next to students, speaking no words at all. We observed teachers H1, H3, H4 and M1 helping frustrated students by using "teacher mode," logging in as the teacher on the student's computer so the student could attempt a puzzle as many times as needed without losing lives. Teachers H1-H3 allowed students to replay previously-passed levels [Liu17], stating that this was partially to alleviate frustration experienced at the current level. As H2 stated, she believed students replayed because they "like that feeling of success." Another commonly-observed affective intervention was praising students' progress. H1-2 and M1 considered ST Math itself as an affective support that helped English language learners gain confidence in math.

6.4.3.3 Peer Tutor

All teachers but L1 stated that they encouraged peer tutoring. Peer tutoring was observed in the classes of H1, H3, M1, and L2 when teachers sent an "ST Math Helper" to support others when they were engaged with other children, commonly in rotation-seatings. Peer tutoring was also observed in free-seatings where children sat in groups and helped each other. Teachers reported that they recognized that: children knew more about the levels; children were more likely to understand and follow peer's instructions than teacher's; teaching itself helped learning; and that children build confidence and leadership through peer tutoring. Teachers purposefully selected peer tutors and guided the peer tutoring process. ST Math Helpers were usually children who made more progress, were more proficient in math, or those who were proficient in math but needed to build their confidence. Teachers also believed that "explaining instead of telling answers" is a skill that needed to be cultivated; students lacking such ability was the main reason L1 noted that she didn't encourage peer tutoring. However, we also observed cases in which children tried to solve the puzzle by looking at a peers' computer. In one case, teacher caught students who switched laptops to play games on their friend's account.

6.4.3.4 Progress Management

Teachers acknowledged that ST Math's self-paced structure would leave students at different levels of progress. Only H3 stated a preference to keep children at similar progress levels so they could help each other. However, teachers did intervene with students at the two extremes, those far ahead and far behind. Students who were falling behind received more interventions during gameplay. Teachers (H2-3, M1) acknowledged outside factors that influenced these students, such as scheduling conflicts with ST Math time, limited internet access at home, and previous exposure to school. Only H2 stated skipping a struggling student ahead once, noting that the student needed to get exposure to other objectives. For students who were too far ahead, H4 and M1 moved the challenge objective (an objective for spatial puzzles without targeting specific math concepts) forward, to stop them from progressing into what they hadn't learned in class. Other teachers (H2, H4, M1-2, L1) assigned

different activities, such as code.org. Two highly experienced teachers (H2-3) commented that it was important to distinguish students who were moving quickly due to their math proficiency from “rushers,” who completed objectives but had low in-game performance. H3 sent these students back to re-do rushed objectives, and H4 stated that she wanted to do that. One rushing behavior observed across classrooms was that some students randomly clicked instead of thoughtfully answering the objectives’ pre-post multiple choice tests, a behavior that had no impact on their measured game progress.

6.4.3.5 Design Implication

This *intervention* category of teacher behaviors highlights the need for various ways to support students during gameplay. The long-term use combined with the self-paced nature of curriculum-integrated games causes student differences to become more salient and important for teachers to manage. This increases the burden on teachers to provide affective and cognitive support to individual students, and makes it difficult for individual students to get help, especially in large classes. Thus, we suggest that such games should invest in features that provide individual interventions, and support teachers in directing their attention to students who most need support during gameplay.

For individualized affective interventions, ST Math could use reward mechanisms [Rav17] to celebrate progress and encourage productive struggle. Our previously-suggested features to allow students to save the current progress, may help frustrated students take a break and refocus again [Sab11].

Our results suggests new features that help students access spatial-temporal feedback at low cost. After each puzzle attempt, ST Math provides an animation that demonstrates whether the student answer solves the problem. When teachers come to help, they often need to deduce what was done by observing students fail. However, incorrect answers cause Jiji to die and students lose their progress on the level. This causes some students to avoid attempting puzzles even if they could learn from the resulting animation, and some teachers to front-load information about how students should play, so they can avoid wasting Jiji’s lives. Thus, to promote learning from failure, ST Math could allow students and teachers to replay the animations from previously-incorrect attempts, and pause during the animation to discuss. Such a feature would help teachers understand why students failed and would allow students to practice spatial-temporal reasoning without costing another life.

Observations and interviews revealed that students often ignore the ST Math animations that show after every answer. We suggest that designers should allow students to shorten or skip animations if they have already reached a certain proficiency, and thus help students pay attention to animations when they are most important for learning. Another related student behavior is the

transforming of the ST Math problems into other (physical) formats. This pattern is a double-edged sword: it bridges ST Math content with classroom content; however, it may not take advantage of the various game representations and scenarios designed to foster spatial-temporal reasoning [PL18]. Thus, the system should be designed to help students pay attention to the feedback that reinforces the important spatial-temporal reasoning the game was designed to foster.

Our results suggest that children generally like to work together and compete with each other. The system could integrate leaderboards to foster healthy competition while allowing children to opt-out from competition [Rav17]. However, the addition of such features should be mindful of structuring competition in a manner that supports mastery goals [US06]. Otherwise, students may rush without thinking, as we observed, or develop their own strategies to game the system [Ale04; Bak08] to progress without necessarily learning.

Lastly, we suggest game designers should personalize features for students at the two progress extremes, on whom teachers spent a significant amount of effort to manage. For fall-behind students, teachers themselves requested features including a scaffolding mode or a misconception report on common wrong answers from pre-post tests. For far-ahead students, easier access to more advanced math content was requested by teachers. We suggest game designers also consider features that encourage peer tutoring, such as real-time suggestions of who needs help and who can help [Dia17b].

6.4.4 Data-Informed Practice

Data-Informed Practice is what teachers do based on educational game data. This category is not explicitly identified by Kangas et al. [Kan17b] but most related to how teachers reflect on their practices in game-based learning. However, our results from interviews suggest teachers primarily used ST Math reports to inform teaching practices outside the game, instead of on improving student performance in the game.

In our interviews, all teachers stated that they use ST Math reports to identify struggling students. All teachers believed that performance in ST Math reflected students' math ability. However, teachers also identified cases where ST Math performance didn't always reflect math skills, due to scheduling conflicts, internet access, or even gaming-the-system behavior in pre- and post-tests (H1-3). Teachers (M1, L1) noted for some students, ST Math performance didn't map to FSA or other math assessments, because the language-heavy nature of non-ST Math assessments posed additional difficulties. Thus, M1, H2, and H4 noted combining in-game assessments with out-of-game assessments to understand whether students were stuck on the reading comprehension elements of traditional assessments or the math concepts, and used this distinction to instruct students accordingly.

When discussing how they used ST Math reports, less experienced teachers focused primarily on progress. L1 and L2 had limited understanding regarding the access and use of reports, with

L2 requiring children to write down progress on paper charts after each gameplay session. On the other hand, M1-2, and H1-4 identified both the struggling students and the content they struggled on, to inform which concepts needed to be retaught and reviewed in class, and which students needed focused instruction. H2 and H4 also stated communicating with students on what they could see in the report, to set expectations on productive learning behaviors and discourage “thinking it’s just a game.” H2 also noted sharing good performance with parents as a way to praise and encourage students. This communication of gameplay data is similar to Kangas et al’s *elaboration* where teachers help students make connections from gameplay back to learning.

6.4.4.1 Design Implication

Our results showed that experienced teachers use ST Math reports for self-reflection, communication with students, and planning future teaching. This use of ST Math reports is similar to results by Kharrufa et al. who studied what kinds of information experienced teachers would like from educational reports [Kha17]. However, our results also suggest that teachers, especially new teachers, need support in understanding ST Math reports and transforming the reports into actions. Although ST Math has a website for report explanation and suggested actions, these are not available directly from the reports page. We suggest game designers integrate information so teachers can see reports, read what they mean, and consider suggested real-time actions in one place, together with clear instructions. Such integration corresponds to Dillenbourg’s technology design suggestions on supporting continuity in teacher’s workflow [DJ10]. Our results also showed that ST Math helped teachers discern whether some students were struggling with math concepts or English reading comprehension on traditional assessments. Therefore, game designers should leverage this opportunity to explore educational games’ potential to assess learning from another perspective.

6.5 What Did Teachers Suggest?

When asked what kinds of supports teachers would like within ST Math, **all teachers** wanted better mapping between ST Math and math standards—“if you’re teaching this specific standard, these are the games you want to assign.” In ST Math, objectives are mapped to standards in a specific gameplay report. However, this mapping is not visible where teachers manipulate ST Math objectives.

Teachers wanted gameplay feedback through multiple communication channels. H2 and L2 suggested post-gameplay reports via email, to check on their devices at convenient times without necessarily logging into ST Math. Moreover, teachers expressed it was often difficult to see students’ screens and notice current alert features given limited attention and class formats. H1-2, H4, M1, and M2 wanted live alert/emails sent to their screen or phone during gameplay. For example, teachers mentioned the need to be informed of the ST Math group’s progress while instructing other groups

in the rotation-seating class format. Lastly, H2 displayed the report to students during gameplay and wanted the report to show positive messages to motivate students, such as the number of students on task.

Teachers also wanted to use reports to inform their teaching. H1-3 suggested detailed answers and time taken on pre/post tests, having a class-level summary on common misconceptions to reteach, and being able to identify students who simply clicked through. H2-3, M1, and L2 wanted reports on estimated idle time. They wanted to tell whether spending long amounts of time meant learning or behavior problems—"either he's not getting it, or he's just not focused," and they wanted alerts for students who were significantly delayed in starting gameplay and for those who were doing other computer activities instead of ST Math. Moreover, when kids are stuck, teachers (H1, H4, M1, L1-2) wanted to understand "what the puzzle looks like," on which content they are struggling, and why. This ties in to our suggestion that students and teachers be able to view an animation of any previously-failed puzzle attempts. M1 wanted help translating percentage progress to give students goals around learning of concepts, such as "I need you to master this fraction concept by the end of next week." L2 wanted the average scores of other classes, to know whether the game was difficult, or she needed to teach this objective better.

Several teachers suggested making ST Math more personalized. H3 and M1 wanted far-ahead students to be able to do upper-grade objectives, as these students were also owed "a year of math growth" even if they started out higher performing. For far-behind students, H1 and M1 wanted to assign individualized objective orderings or homework without affecting the whole class. L2 suggested assigning a "snippet" of objectives, so slower children could play a shorter version of later objectives they won't be able to get to in time for the state test. H4 and L1 suggested scaffolding some games. Other suggestions included: better PD that focuses on understanding and teaching with ST Math more than just understanding the user interface (H1, M1, L1-2); More gamified features such as in-game awards (M1).

To summarize, teachers' self-identified needs corresponded to design suggestions derived from our previous analyses and prior literature. Teachers had high expectations and hopes for the educational game, and their needs went beyond simply making students perform well in-game. They wanted the game to make their own teaching more effective, both for the whole classroom and for individual students. They wanted conveniently-accessible information and actionable feedback from the game to inform teaching.

6.6 Contribution

Through combining field observations in authentic classroom contexts with teacher interviews, we identified and four types of important teacher activities around the use of curriculum-integrated games: *preparation*, *integration*, *intervention*, and *data-informed practice*. We related our findings

to the Kangas et al.'s framework [Kan17b] of teachers' activities for game-based learning, identifying critical differences that arise from aspects of long-term use and curricular integration, such as the need to support student affect, and to help teachers use data to inform their pedagogical practice. Finally, we have derived several insights that can help guide practical game design to benefit both teachers and students. Curriculum-integrated games should provide:

- In-game PD, reports, and just-in-time supports for teacher workflows, to help teachers learn how and why to use the game and its data to inform effective classroom practices.
- Salient mappings of game content to curricula and flexible content reordering.
- Support for different classroom formats and teacher needs for feedback.
- Affective supports for individual students and flexible ways to direct student and teacher attention.
- More flexible ways for students and teachers to use the game together.

There are several notable limitations. This study was conducted on a single game in a single school district during a short 3-week period. There may also be selection bias—teachers who volunteered to participate may be more passionate about ST Math than those who did not. In addition, observations were at the end of the school year. Future work may observe new patterns at the beginning of the school year, when classes are introducing new math content and are getting used to ST Math.

CHAPTER

7

CONCLUSION

To summarize, this dissertation is a pioneering work on curriculum-integrated educational games, a type of game integrated into school classrooms with increasing demand and popularity. My research combined data-driven methods and qualitative approach to derived practical, actionable feedback for game design and broader math education. This dissertation has made the following contributions:

- Pioneering research on curriculum-integrated educational math games, a type of game integrated into school activities with increasing market demand and popularity,
- Data-driven and qualitative methods and insights that inform the game design of ST Math, which could potentially benefit 1,200,000+ students currently using ST Math across the U.S.,
- New data-driven and qualitative methods and insights that are applicable to educational games and other e-learning platforms, benefiting students beyond ST Math.

Together, these contributions support a new methodology of integrating human insights with data-driven methods, combining qualitative and quantitative research to derive actionable, practical insights. In this chapter, I will synthesize how the work in previous chapters contributed to answering my research questions.

7.1 RQ1

How can we derive practical insights on gameplay and learning, with a focus on gameplay and outcomes that arise from use over time?

My dissertation emphasizes on ***applying interpretable data-driven methods and evaluating results against educational literature and human interpretations***, as demonstrated in studies 1-3. For example, Study 3 identified game levels where students failed to follow the ideal learning curves, and evaluated these levels against common fraction mistakes according to educational literature. One such mistake found by data and by the literature is the common mistake of relying on numbers rather than the magnitude or relationship between numerators and denominators. From the analyses in study 3, I suggested what could help mitigate such mistakes in ST Math, such as comparing fractions by locating them on a number line. Moreover, through applying interpretable learning models grounded in educational theory, I suggested that young students sometimes developed content- and context-specific skills instead of transferring knowledge across problem-solving scenarios. This work called for educational games to focus on learning the math skills instead of reinforcing pattern matching. Through applying and innovating data-driven methods in context and evaluating them against literature, this work resulted in generalizable methods to combine data analytics with the literature to derive educational insights across domains.

My dissertation also demonstrated that ***game learning analytics should not only be used to investigate the outcome of learning, but to analyze learning as a progression arising from use over time***. Study 1 analyzed the replay behavior arising from gameplay over time, and studies 2 and 3 derived insights on how students learned and transferred between math objectives over time. For example, Study 2 analyzed the different orders in which students played math objectives (math concepts). From year-long gameplay data, study 2 suggested clusters of objectives and multiple game-design and educational insights. These insights included: calculating with money helps students understand fractions; playing games with virtual manipulatives in fractions helps students understand division; too many steps for manipulating game elements distracts students from focusing on the math content; representing different units with the same number line confuses students; and placing difficult objectives too close to each other poses difficulties. Such educational insights may not be derived without analyzing learning as a process and progression over time.

Next, the arc of my dissertation showed that ***combining data-driven results with contextual knowledge from field research significantly benefits the understanding and interpretation of gameplay behaviors and learning***. For example, Study 1 was the first study on ST Math when I have just started to understand the data. Through data mining and statistical analyses, I found that replay before passing current levels was associated with low game performance and learning. Teachers' insights from the field-study backed up our hypotheses on what might be happening. Most teachers stated that they observed the replay behaviors I identified in my study. Two of the most experienced

teachers considered replay as a student-initiated behavior to regulate emotions. For example, one stated that she allowed students to replay to some extent because she believed students “like that feeling of success.” Two teachers specifically stated that they believed students replay when they were frustrated and tried to avoid playing the current level. Combining the knowledge gained from data-driven methods and the field study, I confirmed my hypothesis that replay before passing the current game level is a sign of struggling on current math content, and an affective need for a mental break. However, because many studies, including my study 4, suggest that students need affective support during games, we could design a new study to investigate if ST Math should give affective interventions on detection of replay patterns and if we can suggest students to replay the right content for the right amount to obtain both affective and instruction benefits. This example shows that we gained a better understanding of gameplay behaviors and learning, and can derive more empirical suggestions through combining data-driven and qualitative research.

Lastly, my dissertation emphasizes that ***field study is an irreplaceable part of understanding students’ gameplay behaviors and learning in context.*** For example, from classroom observations and teacher interviews in Study 4, I found that the children involved in our study were extremely sensitive to success and failure in the game. Teachers spent a significant amount of effort providing just-in-time affective support in the classrooms, sometimes even before instructional help. From this interesting finding, I suggested that the game design should be more sensitive to young students’ emotions, and implement features such as encouraging messages that could alleviate the teacher’s burden. Another example is that we observed many students not paying attention to spatial-visual animations after the answer, which gets in the way of developing spatial reasoning and math skills. This observation provides some interpretation on why students failed to transfer, and focused on underlying math concepts and calls for game design changes, such as reducing animations when students have already reached mastery and making replay of animated feedback from the previous wrong attempt available for students and teachers to reflect on prior work.

To summarize, my answers to how can we derive practical insight of gameplay and learning, with a focus on gameplay and outcomes that arise from use over time are:

- Apply interpretable data-driven methods, and evaluate results against educational literature and human interpretations.
- Apply game learning analytics to not only investigate the outcome of learning, but consider learning as a progression that arises from use over time.
- Combine data-driven results with contextual knowledge from the game and field research, and
- Conduct field study as an irreplaceable part of understanding students’ gameplay behavior and learning in context.

7.2 RQ2

What data-driven methods can we design to evaluate game content against educational outcomes, and to inform designs that maximize students' learning?

Through the creation of three data-driven methods in studies 1-3, my work demonstrated that actionable data-driven methods should be interpretable and generalizable, and their applications must consider a game's design and other contextual information.

First, ***actionable methods must be designed considering a game's content, gameplay mechanics, application context, and other characteristics***. For example, ST Math is designed with a drill-and-practice mechanism, in which students play the same skills (puzzles) over and over again until completing all puzzles with a given number of lives in a level. Item-response theory, the cognitive theory behind learning curves, is based on such repetitive practices. Thus, in Study 3, I applied learning curves to determine whether the game design resulted in the predicted learning gains in time and/or the number of practices. Moreover, I applied the analyses following the hierarchy of ST Math starting with puzzles within levels, then levels within games, and games within objectives. This approach helped me understand learning patterns from the lowest granularity, and then to design methods to investigate the transfer within the same game and objective. Similarly, the self-paced gameplay nature of ST Math, along with the expected potential of games to provide benefits through replay [Gee07], inspired the methods to investigate student voluntary replay behaviors in Study 1. The differences reordering of objectives across classes created diverse objective sequences, which inspired the methods to mine predictive relationships between objectives in Study 2.

My work suggests that such ***data-driven methods could be generalizable across content within the game, and ideally across game systems***. In modern games with many mini-games like ST Math, each game may have a different design. Methods that depend on features to describe context/content-specific elements and related actions can be extremely costly. Instead, having data-driven methods that were designed to utilize shared mechanics and gameplay patterns across games can help evaluate games in an efficient way.

Moreover, as many games were designed with similar mechanics, such as drill-and-practice, good data-driven methods should be generalizable across domains. In study 1, I designed features around students' voluntary replay behaviors, and recognized that grouping of students based on frequencies of types of replay yielded gameplay and learning insights. This analysis can be applied in self-paced learning environments where students have the freedom of choosing and replaying already-passed content. In study 2, I designed a method to mine predictive relationships between math objectives. The method pinpointed where particular objective orderings may not best benefit students, resulting in suggestions about the grouping and sequencing of math content and specific game design insights about where students met difficulties and why. This method is applicable to a wide range of educational technologies with content that may need better sequencing. In study

3, I designed learning curve analyses for ST Math levels, games, and objectives under different educational hypotheses, and pinpointed specific game levels that may need redesign. I designed a new method that hierarchically combines learning curves to identify where knowledge transfer support is needed. My results called for game designs that focus on educational content instead of reinforcing simple pattern-matching. This method is also applicable to games with drill-and-practice mechanics, which are widely applied in modern math games.

Again, results from my data-driven methods were evaluated against educational literature and human interpretation. These methods were proven to have practical value based on field observations. For example, teacher interviews revealed that they have observed students using replay as found in my data-driven analysis, and added the potential interpretation of some replay behavior being used as an affective intervention. Reordering of math objectives was frequently conducted by districts and experienced teachers; data-driven knowledge on the relationships between these objectives can benefit practitioners. Moreover, some students were observed to develop their own strategies for specific games using pens, papers, and manipulatives instead of focusing on transferring knowledge on the underlying math concepts. Pinpointing where learning support is needed would inform game design that can benefit students. Thus, *to derive actionable insights, it is important for data-driven methods to be interpretable and to provide practical value.*

To summarize, I designed three data-driven methods: *Replay Antecedent Analyses*, which captured the circumstances before student-initiated replay and associated them with learning outcomes; *Curricular Sequencing Analyses*, which mined the predictive relationships between curricular contents and suggested reordering of specific contents; and *Hierarchical Learning Curve Analyses*, which pinpointed where learning is not supported by consecutive practices, and evaluated transfer between game contents. These three methods demonstrated that data-driven methods designed to evaluate game content and inform designs should be:

- Designed with awareness of the game’s content, mechanics, application context, and other characteristics,
- Generalizable across different games, and ideally, across learning platforms, and
- Interpretable and designed to provide practical value.

7.3 RQ3

How can we understand the use of ST Math in classrooms, to inform game designs that could benefit practitioners and learners?

Study 4 used a qualitative approach to understand the use of ST Math in classrooms, through combining semi-structured interviews and field observations across various classroom formats and

levels of teacher experience. Based on these qualitative investigations, I identified critical needs and practices of teachers and students and suggested actionable interventions for game design and implementation. I extended and contrasted the study findings with a prior framework on teachers' activities for game-based learning, by identifying critical differences that arise from long-term use and curricular integration. Many of the insights helped interpret the data-driven findings in this dissertation and have inspired future data-driven research.

Study 4 demonstrated three key aspects on methods that can help understand the practical use of an educational game. First, my study suggests that educational game designers and researchers should ***consider studying authentic classroom settings and be aware of the limitations of the methods used and/or theories derived from pre-designed lab-settings***. For example, we originally planned to collect fine-grained affective data by observing student behaviors, facial expressions, and gameplay. However, we found many authentic classroom formats did not support such a method, as students' faces and screens may not be easily observed depending on where they sat and how often they switched seats and activities. In our study, we found direct field notes provided much richer information. Another example of the benefits of field research is that Study 4 identified that teachers spend significant effort providing students with affective support in the classroom. This affective support was not found in a prior literature reviews of teacher activities to support game-based learning. This may be because having ST Math as part of students' school activities make students more emotionally engaged as compared to short-lived pre-designed game sessions in previous studied literature. Alternatively, the novelty effect of a short-lived game could lessen children's behavioral issues during gameplay, or researchers may have considered that teacher interventions to manage student affect during such games were not related to the game-based learning. Thus, studying authentic classroom settings could reveal important factors that influence the practical use of games, that may not occur in pre-designed studies in idealized lab settings.

Second, my study demonstrated that ***it is valuable to cover diverse cases, from both study participants and application scenarios***. For example, my study revealed differences between experienced and new teachers, such as the finding that experienced teachers were more open to re-ordering objectives, and that new teachers didn't know how. From this finding, we concluded that teachers, especially newer teachers, need information (on how to best use ST Math) to be better integrated into teachers' existing workflows instead of on a separate webpage. We also concluded that a curriculum-integrated game should offer a pre-designed curriculum order but also provide flexibility for reordering to support teachers of varied experience. Another important finding was how the student-teacher interactions changed in different classroom formats, an insight that was often neglected in prior research. Teacher interviews revealed that classroom formats may not often be under the teacher's control. Thus, being exposed to diverse classroom formats helped us suggest game design insights that could be more practical and adaptive to classroom needs.

Third, ***my study emphasized combining field observations with teacher interviews***. Field ob-

servations helped researchers gain exposure and experience with empirical scenarios and contextual information before talking to the teachers. Such experience helped us better understand the rationale behind teacher actions, such as why the teachers switched objectives and focused on interacting with specific students. After field observation, having direct conversations with the teachers is of crucial importance. For example, semi-structured interviews revealed how school scheduling and district policies played roles in how teachers organized their classes and set goals for gameplay. These semi-structured interviews also helped show what teachers themselves suggested for the game. Their common suggestions showed that that teachers' needs went beyond making students perform well in the game.

To summarize, this dissertation shows several ways to understand the use of an educational game in classrooms:

- Conduct field studies to understand the rich contextual information in classrooms.
- Consider studying authentic classroom settings, being aware of the limitations of tools, data, and theories derived from pre-designed lab-settings and/or controlled experiments.
- Cover diversity of participants and classroom formats.
- Engage in direct conversation with teachers through interviews to understand the non-visible factors that influence teachers' decisions and rationales from their own perspectives. Conduct field observations to gain exposure and experience in the field before talking to teachers.

7.4 Future Work

This work is the first collaboration between North Carolina State's Computer Science Department, College of Education, and MIND Research Institute. It builds a foundation for our research group to make new insights and design productive systems and experiments. It establishes a productive relationship that we can use to build new knowledge about math education and educational games.

Because study 4 was conducted in the last stage of the project, we have not integrated field study insights into the design of data-driven methods. Future work should investigate gameplay behaviors and design data-driven methods to help solve the practical gaps observed in the field. For example, inspired by the diverse class formats and scheduling we observed in the field and how they influenced students' behaviors, the author is working with Rachel Harred to distinguish class formats from data and apply data-driven methods to determine how long and how frequently students should play ST Math to most benefit their learning. The author is also working with Rachel Harred to distinguish teacher-informed student replay motives from the data, and analyze if these replay patterns reflect student performance or could be used to predict students' consequent gameplay. Moreover, we will

be using the field insights to design large-scale teacher surveys that we can combine with gameplay data to contribute to understanding of teacher practice at a larger scale.

Future work should also implement game design suggestions made based on insights from this dissertation and evaluate their effectiveness. Some design implementations that could make a difference for student learning include: redesigning specific games pinpointed by our data-driven methods; improving teacher reports; experimenting with in-game affective or instructional help to alleviate the teacher's burden; investigating alternative ways to present in-game visual-spatial feedback; and designing gamified features to be sensitive to children's emotions and encourage productive struggle. With new schools adopting ST Math near NC State University, where this research group is located, it would be extremely beneficial to design local user studies and experiments to evaluate these features. Because of the complex factors in particular classrooms and the delays in producing, linking, and sharing anonymized data between software, tests, and research groups, we were unable to triangulate our qualitative field studies with in-game data. Due to this limitation of authentic classroom settings, we may also need to conduct pre-designed experiments where we could strongly tie quantitative and qualitative data to investigate specific features or new game designs.

7.5 Final Canapés

Pursuing the doctoral degree is a fun and rewarding journey. At the end of the dissertation, the author would like to share the most valuable lessons she learned from her journey in a stylish fashion.

- Pick a good advisor first, by interviewing their students not just their research papers.
- Ph.D. is about managing your advisor. Doing whatever your advisor says leads to getting nothing done. To get things done, you need to say both yes and no, and give feedback to your advisor on how to best help you.
- No one cares about your stuff more than you do. "If you don't speak up for what you want, you might end up with a disappointing sandwich made with love."
- When you see smart people come up with ideas and papers like fish giving birth to eggs, you only see the tip of an iceberg. All good research takes time. Take the time to accumulate what's under the water.
- You will fail, frequently. Learn to identify success from failure, and make a paper out of it.
- Ph.D. is more than a dissertation. Find things that make you happy along your journey. Flowers do not bloom on gloomy days.

BIBLIOGRAPHY

- [Adk17] Adkins, S. “2017-2022 global game-based learning market”. *Proceedings of the Serious Play Conference*. 2017.
- [Ale03] Aleven, V. et al. “Help seeking and help design in interactive learning environments”. *Review of Educational Research* **73.3** (2003), pp. 277–320.
- [Ale04] Aleven, V. et al. “Toward tutoring help seeking”. *International Conference on Intelligent Tutoring Systems*. Springer. 2004, pp. 227–239.
- [Arb98] Arbreton, A. “Student goal orientation and help-seeking strategy use.” (1998).
- [Arr17] Arroyo, I. et al. “Wearable learning: multiplayer embodied games for math”. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. 2017, pp. 205–216.
- [BH13] Backlund, P. & Hendrix, M. “Educational games-are they worth the effort? A literature survey of the effectiveness of serious games”. *Proceedings of the 5th International Conference on Games and virtual worlds for serious applications (VS-GAMES)*. 2013.
- [Bai12a] Bai, H. et al. “Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students”. *British Journal of Educational Technology* **43(6)** (2012), pp. 993–1003.
- [Bai14] Bailey, D. et al. “Early predictors of middle school fraction knowledge”. *Developmental Science* **17.5** (2014), pp. 775–785.
- [Bai12b] Bailey, D. H. et al. “Competence with fractions predicts gains in mathematics achievement”. *Journal of Experimental Child Psychology* **113.3** (2012), pp. 447–455.
- [BI14] Baker, R. S. & Inventado, P. S. “Educational data mining and learning analytics”. *Learning Analytics* (2014), pp. 61–75.
- [Bak07] Baker, R. S. et al. “Modeling the acquisition of fluent skill in educational action games”. *proceedings of the International Conference on User Modeling*. 2007, pp. 17–26.
- [Bak08] Baker, R. et al. “Why students engage in “gaming the system” behavior in interactive learning environments”. *Journal of Interactive Learning Research* **19.2** (2008), pp. 185–224.
- [Bak15] Bakker, M. et al. “Effects of playing mathematics computer games on primary school students’ multiplicative reasoning ability”. *Contemporary Educational Psychology* **40** (2015), pp. 55–71.

- [BB11] Barendregt, W. & Bekker, T. M. “The influence of the level of free-choice learning activities on the use of an educational computer game”. *Computers & Education* **56.1** (2011), pp. 80–90.
- [Bar05] Barnes, T. “The Q-matrix Method: Mining Student Response Data for Knowledge”. *Proceedings of the 20th National Conference on Artificial Intelligence*. 2005.
- [Bar96] Bartle, R. “Hearts, clubs, diamonds, spades: Players who suit MUDs”. *Journal of MUD research* **1.1** (1996), p. 19.
- [Bau17] Bauer, A. et al. “Analysis of problem-solving behavior in open-ended scientific-discovery game challenges”. In *Proceedings of the 9th international conference on educational data mining (EDM)*. 2017.
- [Bes17] Beserra, V. et al. “On-Task and Off-Task Behavior in the Classroom: A Study on Mathematics Learning With Educational Video Games”. *Journal of Educational Computing Research* (2017), p. 0735633117744346.
- [Bok15] Bokosmaty, S. et al. “Learning geometry problem solving by studying worked examples: Effects of learner guidance and expertise”. *American Educational Research Journal* **52.2** (2015), pp. 307–333.
- [Boy11] Boyce, A. et al. “BeadLoom game: Adding competitive, user generated, and social features to increase motivation”. *the 6th International Conference on Foundations of Digital Games*. ACM. 2011, pp. 139–146.
- [Boy16] Boyle, E. A. et al. “An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games”. *Computers & Education* **94** (2016), pp. 178–192.
- [Bre15] Brezovszky, B. et al. “Developing adaptive number knowledge with the Number Navigation game-based learning environment”. *Describing and studying domain-specific serious games*. Springer, 2015, pp. 155–170.
- [Bru11] Brunskill, E. “Estimating Prerequisite Structure From Noisy Data.” *EDM*. 2011, pp. 217–222.
- [BA12] Buabeng-Andoh, C. “Factors Influencing Teachers’ Adoption and Integration of Information and Communication Technology into Teaching: A Review of the Literature.” *International Journal of Education and Development using Information and Communication Technology* **8.1** (2012), pp. 136–155.
- [Bur15] Burgers, C. et al. “How feedback boosts motivation and play in a brain-training game”. *Computers in Human Behavior* **48** (2015), pp. 94–103.

- [Cal18] Callaghan, M. N. et al. "How teachers integrate a math computer game: Professional development use, teaching practices, and student achievement". *Journal of Computer Assisted Learning* **34.1** (2018), pp. 10–19.
- [Cal05] Calvert, S. L. et al. "Control as an engagement feature for young children's attention to and learning of computer content". *American Behavioral Scientist* **48.5** (2005), pp. 578–589.
- [Cas17] Casalino, G. et al. "Q-matrix Extraction from Real Response Data Using Nonnegative Matrix Factorizations" (2017), pp. 203–216.
- [Cas14] Castellar, E. N. et al. "Improving arithmetic skills through gameplay: Assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes". *Information sciences* **264** (2014), pp. 19–31.
- [Cen06] Cen, H. et al. "Learning factors analysis-a general method for cognitive model evaluation and improvement". *Intelligent Tutoring Systems* **4053** (2006), pp. 164–175.
- [Cen08] Cen, H. et al. "Comparing two IRT models for conjunctive skills". *International Conference on Intelligent Tutoring Systems (ITS)*. 2008, pp. 796–798.
- [CS92] Chandler, P. & Sweller, J. "The split-attention effect as a factor in the design of instruction." *British Journal of Educational Psychology* **62(2)** (1992).
- [Che14] Chee, Y. S. et al. "Facilitating dialog in the game based learning classroom: Teacher challenges reconstructing professional identity". *Digital Culture & Education* **6.4** (2014), pp. 298–316.
- [Che15] Chen, Y. et al. "Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining". *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*. 2015.
- [Che16] Chen, Y. et al. "Joint discovery of skill prerequisite graphs and student models". *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*. 2016.
- [Che12] Chen, Z.-H. et al. "Influence of Game Quests on Pupils' Enjoyment and Goal-pursuing in Math Learning." *Journal of Educational Technology & Society* **15.2** (2012).
- [Cla11] Clark, D. B. et al. "Exploring newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States". *Computers & Education* **57(3)** (2011), pp. 2178–2195.
- [CS14] Clements, D. & Sarama, J. *Learning and teaching early math: The learning trajectories approach*. Routledge, 2014.
- [Con02] Conati, C. "Probabilistic Assessment of User's Emotions in Educational Games". *Applied Artificial Intelligence* **16(7-8)** (2002), pp. 555–575.

- [CZ04] Conati, C. & Zhao, X. "Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game". *the 9th international conference on Intelligent user interfaces*. 2004, pp. 6–13.
- [Con17a] Confrey, J. et al. "Learning trajectories: a framework for connecting standards with curriculum". *ZDM Mathematics Education* **46(5)** (2017), pp. 719–733.
- [Con17b] Confrey, J. et al. "Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories". *ZDM Mathematics Education* (2017), pp. 1–18.
- [CL96] Cordova, D. I. & Lepper, M. R. "Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice". *Journal of Educational Psychology* **88.4** (1996), p. 715.
- [DC13] Davoodi, A. & Conati, C. "Degeneracy in Student Modeling with Dynamic Bayesian Networks in Intelligent Edu-Games". 2013, pp. 120–123.
- [Dav13] Davoodi, A. et al. "Understanding Users' Interaction Behavior with an Intelligent Educational Game: Prime Climb". *AIED 2013 Workshops Proceedings Volume 2 Scaffolding in Open-Ended Learning Environments (OELEs)*. 2013, p. 9.
- [DT10] Demirbilek, M. & Tamer, S. L. "Math teachers' perspectives on using educational computer games in math education". *Procedia-Social and Behavioral Sciences* **9** (2010), pp. 709–716.
- [Der16] Derboven, J. et al. "Playing educational math games at home: The Monkey Tales case". *Entertainment Computing* **16** (2016), pp. 1–14.
- [DN13] Desmarais, M. C. & Naceur, R. "A Matrix Factorization Method for Mapping Items to Skills and for Enhancing". *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED2013)*. 2013, pp. 441–450.
- [Des06] Desmarais, M. C. et al. "Learned student models with item to item knowledge structures". *User Modeling and User-Adapted Interaction* **16.5** (2006), pp. 403–434.
- [Dia17a] Diana, N. et al. "An instructor dashboard for real-time analytics in interactive programming assignments". *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM. 2017, pp. 272–279.
- [Dia17b] Diana, N. et al. "Automatic Peer Tutor Matching: Data-Driven Methods to Enable New Opportunities for Help." *EDM*. 2017.
- [DJ10] Dillenbourg, P. & Jermann, P. "Technology for classroom orchestration". *New science of learning*. Springer, 2010, pp. 525–552.

- [Don07] Dondlinger, M. J. "Educational Video Game Design : A Review of the Literature Educational Video Game Design : A Review of the Literature". *Journal of Applied Educational Technology* **4(1)** (2007), pp. 21–31.
- [Dor16] Doroudi, S. et al. "Sequence matters, but how exactly? A method for evaluating activity sequences from data". *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*. 2016.
- [Dun07] Duncan, G. et al. "School readiness and later achievement". *Developmental Psychology* **43.6** (2007), p. 1428.
- [Eag12] Eagle, M. et al. "Interaction Networks: Generating High Level Hints Based on Network Community Clustering". 2012.
- [ES13] Eastwood, J. L. & Sadler, T. D. "Teachers' implementation of a game-based biotechnology curriculum". *Computers and Education* **66** (2013), pp. 11–24.
- [Fin15] Findings, S. U. N. *From Print to Pixel: The role of videos, games, animations and simulations within K-12 education*. <http://www.tomorrow.org/speakup/SU15AnnualReport.html>. 2015.
- [Fre16] Freire, M. et al. "Game learning analytic: Learning analytics for serious games." *Learning, Design, and Technology* (2016).
- [Gal13] Galant, J. "Selecting and sequencing mathematics tasks: Seeking mathematical knowledge for teaching." *Perspectives in Education* **31(3)** (2013), pp. 34–48.
- [Gea08] Geary, D. et al. "Development of number line representations in children with mathematical learning disability". *Developmental Neuropsychology* **33.3** (2008), pp. 277–299.
- [Gee07] Gee, J. P. *What video games have to teach us about learning and literacy*. New York, USA: St. Martin's Griffin - Macmillan, 2007.
- [Gia13] Giannakos, M. N. "Enjoy and learn with educational games: Examining factors affecting learning performance". *Computers & Education* **68** (2013), pp. 429–439.
- [Gir16] Giroto, V. et al. "Lessons learned from in-school use of rtag: A robo-tangible learning environment". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 919–930.
- [Gre18] Gresalfi, M. S. et al. "Design matters: explorations of content and design in fraction games". *Educational Technology Research and Development* **66.3** (2018), pp. 579–596.
- [Gun08] Gunter, G. A. et al. "Taking educational games seriously: Using the RETAIN model to design endogenous fantasy into standalone educational games". *Educational Technology Research and Development* **56(5-6)** (2008), pp. 511–537.

- [HA11] Habgood, M. J. & Ainsworth, S. E. “Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games”. *The Journal of the Learning Sciences* **20.2** (2011), pp. 169–206.
- [Han17] Hansen, N. et al. “Identifying learning difficulties with fractions: A longitudinal study of student growth from third through sixth grade”. *Contemporary Educational Psychology* **50** (2017), pp. 45–59.
- [HA15] Harpstead, E. & Aleven, V. “Using empirical learning curve analysis to inform design in an educational game”. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)*. ACM, 2015, pp. 197–207.
- [Har14] Harpstead, E. et al. “Using Extracted Features to Inform Alignment-driven Design Ideas in an Educational Game”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Toronto, Ontario, Canada: ACM, 2014, pp. 3329–3338.
- [Hic16] Hick, D. et al. “Using game analytics to evaluate puzzle design and level progression in a serious game”. *Proceedings of the 6th international conference on learning analytic and knowledge (LAK)*. 2016.
- [Hie17] Hieftje, K. et al. “An Evaluation of an Educational Video Game on Mathematics Achievement in First Grade Students”. *Technologies* **5.2** (2017), p. 30.
- [Hol18] Holstein, K. et al. “The classroom as a dashboard: co-designing wearable cognitive augmentation for K-12 teachers”. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM. 2018, pp. 79–88.
- [Imb07] Imbo, I. et al. “The role of working memory in carrying and borrowing”. *Psychological Research* **71.4** (2007), pp. 467–483.
- [Jan13] Jansen, B. R. et al. “Math practice and its influence on math skills and executive functions in adolescents with mild to borderline intellectual disability”. *Research in developmental disabilities* **34.5** (2013), pp. 1815–1824.
- [Jon09] Jonker, V. et al. “The motivational power of mini-games for the learning of mathematics”. *Third European Conference on Gamebased Learning (ECGBL)*. 2009, pp. 202–210.
- [Jor10] Jordan, N. et al. “The importance of number sense to mathematics achievement in first and third grades”. *Learning and Individual Differences* **20.2** (2010), pp. 82–88.
- [Jor13] Jordan, N. et al. “Developmental predictors of fraction concepts and procedures”. *Journal of Experimental Child Psychology* **116.1** (2013), pp. 45–58.
- [Jor17] Jordan, N. C. et al. “Delaware longitudinal study of fraction learning: Implications for helping children with mathematics difficulties”. *Journal of Learning Disabilities* **50.6** (2017), pp. 621–630.

- [Kan] Kang, H. *Teaching Habits that Block Productive Struggle in Math Students*. Last Retrieved September 10, 2018.
- [Kan17a] Kang, J. et al. "Using gameplay data to examine learning behavior patterns in a serious game". *Computers in Human Behavior* **72** (2017), pp. 757–770.
- [Kan17b] Kangas, M. et al. "A qualitative literature review of educational games in the classroom: the teacher's pedagogical activities". *Teachers and Teaching* **23.4** (2017), pp. 451–470.
- [Ke09] Ke, F. "A qualitative meta-analysis of computer games as learning tools". *Handbook of Research on Effective Electronic Gaming in Education* **1** (2009), pp. 1–32.
- [Ke06] Ke, F. "Classroom goal structures for educational math game application". *Proceedings of the 7th international conference on Learning sciences*. International Society of the Learning Sciences. 2006, pp. 314–320.
- [Ke08] Ke, F. "A case study of computer gaming for math: Engaged learning from gameplay?" *Computers & education* **51.4** (2008), pp. 1609–1620.
- [KG07] Ke, F. & Grabowski, B. "Gameplaying for maths learning : cooperative or not ?" *British Journal of Educational Technology* **38.2** (2007), pp. 249–259.
- [Keb10] Kebritchi, M. et al. "The effects of modern mathematics computer games on mathematics achievement and class motivation". *Computers & education* **55.2** (2010), pp. 427–443.
- [Kha17] Kharrufa, A. et al. "Group Spinner: recognizing and visualizing learning in the classroom for reflection, communication, and planning". *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 5556–5567.
- [KK18] Kiili, K. & Ketamo, H. "Evaluating cognitive and affective outcomes of a digital game-based math test". *IEEE Transactions on Learning Technologies* **11.2** (2018), pp. 255–263.
- [Kin11] King, A. "Using interactive games to improve math achievement among middle school students in need of remediation". PhD thesis. The George Washington University, 2011.
- [Kli11] Klinkenberg, S. et al. "Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation". *Computers & Education* **57.2** (2011), pp. 1813–1824.
- [Kos18] Kosmas, P. et al. "Moving Bodies to Moving Minds: A Study of the Use of Motion-Based Games in Special Education". *TechTrends* (2018), pp. 1–8.
- [Kou17] Kourakli, M. et al. "Towards the improvement of the cognitive, motoric and academic skills of students with special educational needs using Kinect learning games". *International Journal of Child-Computer Interaction* **11** (2017), pp. 28–39.

- [Ku14] Ku, O. et al. "The Effects of Game-Based Learning on Mathematical Confidence and Performance : High Ability vs . Low Ability". *Educational Technology & Society* **17(3)** (2014), pp. 65–78.
- [Lan14] Lan, A. S. et al. "Sparse Factor Analysis for Learning and Content Analytics". *Journal of Machine Learning Research* **15.1** (2014), pp. 1959–2008.
- [LJ17] Lazem, S. & Jad, H. A. "We Play We Learn: Exploring the Value of Digital Educational Games in Rural Egypt". *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 2782–2791.
- [Lee04] Lee, J. et al. "More than Just Fun and Games : Assessing the Value of Educational Video Games in the Classroom 1101 Beal Avenue". *ACM CHI extended abstracts on Human factors in computing systems*. 2004, pp. 1375–1378.
- [LH06] Lehtinen, E. & Hannula, M. M. "Attentional processes, abstraction and transfer in early mathematical development". *Instructional psychology: Past, present and future trends. Fifteen essays in honour of Erik De Corte* **49** (2006), pp. 39–55.
- [LIA12] LIAO, C. C. et al. "UNFOLDING LEARNING BEHAVIORS: A SEQUENTIAL ANALYSIS APPROACH IN A GAME-BASED LEARNING ENVIRONMENT." *Research & Practice in Technology Enhanced Learning* **7.1** (2012).
- [Lim11] Lim, C. P. et al. "Employing an activity-theoretical perspective to localize an educational innovation in an elementary school". *Journal of Educational Computing Research* **44.3** (2011), pp. 319–344.
- [Lin11] Lindström, P. et al. "Matching and mismatching between the pedagogical design principles of a math game and the actual practices of play". *Journal of Computer Assisted Learning* **27.1** (2011), pp. 90–102.
- [Liu17] Liu, Z. et al. "The Antecedents of and Associations with Elective Replay in An Educational Game: Is Replay Worth It?" *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*. 2017.
- [Loh15] Loh, C. S. et al. *Serious game analytics. Methodologies for performance, measurement, assessment, and improvement*. Cham, Switzerland: Springer, 2015.
- [Lom13] Lomas, D. et al. "Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments". *proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2013, pp. 89–98.
- [MN17] Maas, H. L. van der & Nyamsuren, E. "Cognitive analysis of educational games: The number game". *Topics in cognitive science* **9.2** (2017), pp. 395–412.
- [MM04] Magnussen, R. & Misfeldt, M. "Player transformation of educational multiplayer games." *Other Players, Copenhagen, Denmark*. 2004.

- [MS05] Martin, T. & Schwartz, D. “Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts”. *Cognitive Science* **29.4** (2005), pp. 587–625.
- [Mar15] Martin, T. et al. “Learning Fractions by Splitting: Using Learning Analytics to Illuminate the Development of Mathematical Understanding”. *Journal of the Learning Sciences* **24.4** (2015), pp. 593–637.
- [MC07] Marx, J. D. & Cummings, K. “Normalized change”. *American Journal of Physics* **75.1** (2007), pp. 87–91.
- [Mas17] Masek, M. et al. “Improving mastery of fractions by blending video games into the Math classroom”. *Journal of Computer Assisted Learning* **33.5** (2017), pp. 486–499.
- [McL17] McLaren, B. M. et al. “A computer-based game that promotes mathematics learning more than a conventional approach”. *International Journal of Game-Based Learning (IJGBL)* **7(1)** (2017), pp. 36–56.
- [Mif13] Mifsud, C. L. et al. “Attitudes towards and effects of the use of video games in classroom learning with specific reference to literacy attainment”. *Research in Education* **90.1** (2013), pp. 32–52.
- [Min17] Min, W. et al. “Inducing Stealth Assessors from Game Interaction Data”. 2017, pp. 212–223.
- [MH08] Mitchell, A. & Horne, M. “Fraction number line tasks and the additivity concept of length measurement” (2008), pp. 353–360.
- [Mol17] Molin, G. “The role of the teacher in game-based learning: A review and outlook”. In *serious games and edutainment applications*. Springer International Publishing, 2017, pp. 649–674.
- [Mor14] Moretti, A. et al. “Data-driven curriculum design: Mining the web to make better teaching decisions”. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. 2014.
- [Mos02] Mostow, J. et al. “A la recherche du temps perdu, or as time goes by: Where does the time go in a reading tutor that listens?” In *International Conference on Intelligent Tutoring Systems*. 2002, pp. 320–329.
- [Mp16] Moyer-packenham, P. et al. “The role of affordances in children’s learning performance and efficiency when using virtual manipulative mathematics touch-screen apps”. *Mathematics Education Research Journal* **28.1** (2016), pp. 79–105.
- [Mur13] Murray, R. et al. “Revealing the learning in learning curves”. *Proceedings of the International Conference on Artificial Intelligence in Education*. 2013, pp. 473–482.

- [Nan18] Nanavati, A. et al. “Speak Up: A Multi-Year Deployment of Games to Motivate Speech Therapy in India”. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 318.
- [Nat01] Nathan, M. et al. “Expert blind spot: When content knowledge eclipses pedagogical content knowledge”. *Proceedings of the 3rd International Conference on Cognitive Science*. 2001.
- [NR81] Newell, A. & Rosenbloom, P. S. “Mechanisms of skill acquisition and the law of practice.” *Cognitive skills and their acquisition* **1** (1981), pp. 1–55.
- [Och16] Ochoa, X. “Simple Metrics for Curricular Analytics”. *Proceedings of the 1st Learning Analytics for Curriculum and Program Quality Improvement Workshop*. 2016.
- [Par12] Pareto, L. et al. “A teachable-agent-based game affording collaboration and competition: Evaluating math comprehension and motivation”. *Educational Technology Research and Development* **60.5** (2012), pp. 723–751.
- [Pav15] Pavlik, P. et al. “A Measurement Model of Microgenetic Transfer for Improving Instructional Outcomes”. *International Journal of Artificial Intelligence in Education* **25.3** (2015), pp. 346–379.
- [PT14] Pechenizkiy, M. & Toledo, P. A. “Learning to teach like a Bandit.” *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. 2014.
- [PL17] Peddycord-Liu, Z. et al. “Using Serious Game Analytics to Inform Digital Curricular Sequencing: What Math Objective Should Students Play Next?” *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI-Play)*. 2017, pp. 195–204.
- [PL18] Peddycord-Liu, Z. et al. “Learning Curve Analysis in a Large-Scale, Drill-and-Practice Serious Math Game: Where Is Learning Support Needed?” *International Conference on Artificial Intelligence in Education*. Springer. 2018, pp. 436–449.
- [Pla13] Plass, J. L. et al. “The Impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation”. *Journal of Educational Psychology* **105(4)** (2013), p. 1050.
- [Pol10] Polycarpou, I. et al. “Math-City : an educational game for K-12 mathematics”. *Procedia Social and Behavioral Science* **9** (2010), pp. 845–850.
- [Pre05] Prensky, M. “Computer games and learning: Digital game-based learning”. *Handbook of computer games studies*. Cambridge, MA, USA: The MIT Press, 2005.
- [RS11] Ramani, G. & Siegler, R. “Reducing the gap in numerical knowledge between low-and middle-income preschoolers”. *Journal of Applied Developmental Psychology* **32.3** (2011), pp. 146–159.

- [Ran92] Randel, J. M. et al. "The effectiveness of games for educational purposes: A review of recent research". *Simulation & gaming* **23.3** (1992), pp. 261–276.
- [Rau09] Rau, M. A. et al. "Intelligent Tutoring Systems with Multiple Representations and Self-Explanation Prompts Support Learning of Fractions". *proceedings of the International Conference on Artificial Intelligence in Education*. 2009, pp. 441–448.
- [Rav17] Ravyse, W. S. et al. "Success factors for serious games to enhance learning: a systematic review". *Virtual Reality* **21.1** (2017), pp. 31–58.
- [Ric13a] Richards, J. et al. *Games for a digital age: K-12 market map and investment analysis*. Joan Ganz Cooney Center at Sesame Workshop, New York. 2013.
- [Ric13b] Riconscente, M. M. "Results from a controlled study of the iPad fractions game Motion Math". *Games and Culture* **8(4)** (2013), pp. 186–214.
- [Rin17] Rinne, L. et al. "Development of fraction comparison strategies: A latent transition analysis." *Developmental Psychology* **53.4** (2017), pp. 713–730.
- [RJ01] Rittle-Johnson, B. et al. "Developing conceptual understanding and procedural skill in mathematics: An iterative process." *Journal of Educational Psychology* **93.2** (2001), pp. 346–362.
- [RJ09] Rittle-Johnson, B. et al. "Students' fraction comparison strategies as a window into robust understanding and possible pointers for instruction." *Educational Studies in Mathematics* **72.1** (2009), pp. 127–138.
- [Rol14] Roll, I. et al. "On the benefits of seeking (and avoiding) help in online problem-solving environments". *Journal of the Learning Sciences* **23.4** (2014), pp. 537–560.
- [RV13] Romero, C. & Ventura, S. "Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3.1** (2013), pp. 12–27.
- [Ros03] Rosas, R. et al. "Beyond Nintendo: design and assessment of educational video games for first and second grade students". *Computers & Education* **40.1** (2003), pp. 71–94.
- [Row17] Rowe, E. et al. "Computers in Human Behavior". *Journal of Educational Psychology* (2017).
- [Rut14] Rutherford, T. et al. "A Randomized trial of an elementary school mathematics software intervention: spatial-temporal math". *Journal of Research on Educational Effectiveness* **7(4)** (2014), pp. 358–383.
- [Rut10] Rutherford, T. et al. "Spatial Temporal Mathematics at Scale: An Innovative and Fully Developed Paradigm to Boost Math Achievement among All Learners." *Available from ERIC database (ED510612)* (2010).

- [Sab13] Sabourin, J. L. et al. “Considering alternate futures to classify off-task behavior as emotion self-regulation: A Supervised learning approach”. *JEDM-Journal of Educational Data Mining* **5.1** (2013), pp. 9–38.
- [Sab11] Sabourin, J. et al. “When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments”. *International Conference on Artificial Intelligence in Education*. Springer. 2011, pp. 534–536.
- [Sal15] Saldaña, J. *The coding manual for qualitative researchers*. Sage, 2015.
- [San06] Sandford, R. et al. *Teaching with Games*. The proceedings of the JISC Online Conference: Innovating e-Learning. Cheltenham. Direct Learn Services Ltd. 2006.
- [Sch14] Schenke, K. et al. “Alignment of game design features and state mathematics standards: Do results reflect intentions?” *Computers & Education* **76** (2014), pp. 215–224.
- [Sch16] Schenke, K. et al. “Construct confounding among predictors of mathematics achievement”. *AERA Open* **2.2** (2016), p. 2332858416648930.
- [SK12] Shavelson, R. & Kurpius, A. “Reflections on learning progressions”. *Learning progressions in science*. Springer, 2012, pp. 13–26.
- [Sie11] Siegler, R. et al. “An integrated theory of whole number and fractions development.” *Cognitive psychology* **62.4** (2011), pp. 273–296.
- [Sie12] Siegler, R. et al. “Early Predictors of High School Mathematics Achievement”. *Psychological Science* **23.7** (2012), pp. 691–697.
- [SB12] Siemens, G. & Baker, R. S.J. d. “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration”. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. 2012, pp. 252–254.
- [Sjö17] Sjöden, B. et al. “Can a Teachable Agent Influence How Students Respond to Competition in an Educational Game?” *International Conference on Artificial Intelligence in Education*. Springer. 2017, pp. 347–358.
- [Sno15] Snow, E. L. et al. “The Dynamical Analysis of Log Data Within Educational Games”. *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*. Ed. by Loh, C. S. et al. Springer International Publishing, 2015, pp. 81–100.
- [Sto11] Stone, K. et al. “Electronic Collaborative Learning in Math-City”. *International Conference on Frontiers in Education: Computer Science and Computer Engineering*. 2011, pp. 295–301.
- [Str14] Straatemeier, M. “Math Garden: A new educational and scientific instrument”. PhD thesis. University of Amsterdam, 2014.

- [TV14] Takeuchi, L. M. & Vaala, S. *Level up Learning: A National Survey on Teaching with Digital Games*. Joan Ganz Cooney Center, New York. 2014.
- [Tal17] Talandron, M. M. P. et al. “Modeling the Incubation Effect Among Students Playing an Educational Game for Physics”. *International Conference on Artificial Intelligence in Education*. Springer. 2017, pp. 371–380.
- [Ter18] Ternblad, E.-M. et al. “Do Preschoolers ‘Game the System’? A Case Study of Children’s Intelligent (Mis) Use of a Teachable Agent Based Play-&-Learn Game in Mathematics”. *International Conference on Artificial Intelligence in Education*. Springer. 2018, pp. 557–569.
- [Tho04] Thomas, S. et al. “Designing for learning or designing for fun? Setting usability guidelines for mobile educational games”. *Learning with mobile devices: A book of papers* (2004), pp. 173–181.
- [Tor15] Torbeyns J., S. M.X.Z.S.R. S. “Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents”. *Learning and Instruction* **37** (2015), pp. 5–13.
- [TJ17] Tucker, S. I. & Johnson, T. N. “I thought this was a study on math games: Attribute modification in children’s interactions with mathematics apps”. *Education Sciences* **7.2** (2017), p. 50.
- [US06] Urdan, T. & Schoenfelder, E. “Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs”. *Journal of school psychology* **44.5** (2006), pp. 331–349.
- [Van17] Vandercruysse, S. et al. “Content integration as a factor in math-game effectiveness”. *Educational technology research and development* **65.5** (2017), pp. 1345–1368.
- [Ven13] Ven, S. H. van der et al. “Visuospatial working memory and mathematical ability at different ages throughout primary school”. *Learning and Individual Differences* **27** (2013), pp. 182–192.
- [Vru15] Vrugte, J. ter et al. “How competition and heterogeneous collaboration interact in pre-vocational game-based mathematics education”. *Computers & Education* **89** (2015), pp. 42–52.
- [Vru17] Vrugte, J. ter et al. “Computer game-based mathematics education: Embedded faded worked examples facilitate knowledge acquisition”. *Learning and instruction* **50** (2017), pp. 44–53.
- [Vuo11] Vuong, A. et al. “A method for finding prerequisites within a curriculum”. *Proceedings of the 4th International Conference on Educational Data Mining (EDM)*. 2011.

- [War15] Warshauer, H. K. “Productive struggle in middle school mathematics classrooms”. *Journal of Mathematics Teacher Education* **18.4** (2015), pp. 375–400.
- [Wat11] Watson, W. R. et al. “A case study of the in-class use of a video game for teaching high school history”. *Computers & Education* **56.2** (2011), pp. 466–474.
- [WR07] Witzel, B. & Riccomini, P. “Optimizing math curriculum to meet the learning needs of students”. *Preventing School Failure: Alternative Education for Children and Youth* **52.1** (2007), pp. 13–18.
- [Wou13] Wouters, P. et al. “A meta-analysis of the cognitive and motivational effects of serious games”. *Journal of Educational Psychology* **105(2)** (2013), pp. 249–265.
- [Yan15] Yang, Y. et al. “Concept graph learning from educational data”. *Proceedings of the 8th International Conference on Web Search and Data Mining*. ACM, 2015.

APPENDIX

APPENDIX

A

MATH GAME LITERATURE

Table A.1 Articles of Studies on Math Educational Games.

Study & Math Game	Learning Objective	Game Mechanics	Method	Sample	Duration	Data Collection	Outcomes (if control-experiment)
Randel et al. (1992). NA [Ran92]	NA	NA	literature review	8 math game/67 games	NA	NA	7/8 positive. lack rigorous in experiment design
Rosas et al.(2003): Nintendo [Ros03]	Selected Math topics and Reading Comprehension	Nintendo mini-games, role-playing	control-experiment, user study	1274 1-2nd graders	30 daily over 3 months	pre/post-tests, survey, field observations	Mixed
Conati and Zhao (2004): Prime Climb [CZ04]	Number Factorization	pair-wise competition	comparative study	20 7th graders	20 minutes	pre/post-tests	NA (no control group)
Lee et al. (2004): Skill Arena [Lee04]	Addition & Subtraction	pair-wise competition, drill-and-practice	user study	39 2nd graders	19 days	in-game logs, field observations	NA

Table A.1 (continued).

Magnussen and Misfeldt (2004): Matematrix [MM04]	Number Pattern	team competition, simulation	user study	8-12 years old	one session	field observations	NA
Ke (2006): Astra Eagle [Ke06]	Academic Standard Aligned	mini-games, drill-and-practice	comparative study	124 5th graders	two 40 minutes sessions	pre/post-tests, survey	NA
Ke and Grabowski (2007): Astra Eagle [KG07]	Academic Standard Aligned	mini-games, drill-and-practice	control-experiment, comparative study	125 5th graders	two 40 minutes sessions	pre/post-tests, survey	positive
Baker et al. (2007): Zombie Division [Bak07]	Number Pattern	simulation, drill-and-practice	game learning analytics	15 elementary	135 minutes across 4 weeks	in-game logs	NA
Ke and Grabowski (2008): Astra Eagle [Ke08]	Academic Standard Aligned	mini-games, drill-and-practice	user study	15 4-5th graders	two 40 minutes sessions	pre/post-tests, survey, think-aloud, field observations	NA

Table A.1 (continued).

Stone et al. (2010): Math City [Sto11]	basic math operation, fraction, resource management	simulation, collaboration	user study, comparative study	14 elementary	two 20 minutes sessions	field observations, survey, in-game performance	NA
Kebritchi et al. (2010): DimensionM [Keb10]	algebra	simulation, collaboration, competition	control-experiment	193 9-10th graders	30 per week across 18 weeks	pre/post-tests, survey, interview	Positive in math test, null in attitude
Rutherford et al. (2010): ST Math [Rut10]	Academic Standard Aligned, spatial-visual reasoning	mini-games, drill-and-practice	control-experiment	2-5th graders from 34 schools	multi-years	pre/post-tests, survey	positive in math test and attitude
Lindstrom et al. (2010): Squares Family [Lin11]	place-value, spatial reasoning, strategical thinking	teachable agent, collaboration, pair-wise competition	user study	40 8-10th graders	7 30 minutes sessions	pre/post-tests, field observations	NA

Table A.1 (continued).

Habgood and Ainsworth (2011): Zombie Division [HA11]	Number Pattern	simulation, drill-and-practice	comparative study	58 7-8 years old; 16 9-11 years old	135 minutes	pre/post-tests, in-game logs, interview	NA
King (2011): VmathLive [Kin11]	Academic Standard Aligned	competition, drill-and-practice, role-playing	control-experiment	128 7th graders	18 weeks	pre/post-tests	positive in math test
Klinkenberg et al. (2011): Maths Garden [Kli11]	selected math topics, visospatial working memory	mini-games, drill-and-practice	game learning analytics	3648 primary school	10 months	game-play logs	NA
Bai et al. (2012): DimensionM [Bai12a]	algebra	simulation, collaboration, competition	control-experiment	437 8th graders	18 weeks	pre/post-tests, survey	Positive in math test and motivation
Pareto et al. (2012): Squares Family [Par12]	place-value, spatial reasoning, strategical thinking	teachable agent, collaboration, pair-wise competition	control-experiment	38 3rd graders	7 34 minutes sessions	pre/post-tests, survey	positive in math test, null in attitude

Table A.1 (continued).

Chen et al. (2012): My-Pet-My-Quest [Che12]	academic standard aligned	role-playing, mini-games, drill-and-practices	comparative study	53 4th graders	2 40 minutes sessions	in-game logs, survey	NA
Liao et al. (2012): My-Pet-My-Quest [LIA12]	academic standard aligned	role-playing, mini-games, drill-and-practices	game learning analytics	29 3rd graders	4 montsh	in-game logs	NA
van der Ven et al. (2013): Maths Garden [Ven13]	selected math topics, visospacial working memory	mini-games, drill-and-practice	game learning analytics	22,731 primary	46 weeks	game-play logs, rating logs, human-tagged game characteristics	NA
Jansen et al. (2013): Maths Garden [Jan13]	selected math topics, visospacial working memory	mini-games, drill-and-practice	control-experiment	58 12-15 years old	5 weeks	pre/post-tests	Mixed
Giannakos (2013): Gem-Game [Gia13]	calculation	role-playing	control-experiment, user study	41 and 46 2nd graders	4 weeks	pre/post-tests, survey	NA

Table A.1 (continued).

Riconscente (2013): Motion Math [Ric13b]	academic standard aligned, motor skills	mini-games, drill-and-practice	control-experiment	122 5th graders	20 minutes for 5 days	pre/post-tests, survey	positive in math test, self-efficacy and attitude
Davoodi et al. (2013): Prime Climb [Dav13]	Number Factorization	pair-wise competition	game learning analytics	45 5-6th graders	NA	in-game logs	NA
Rutherford et al. (2014): ST Math [Rut14]	Academic Standard Aligned, spatial-visual reasoning	mini-games, drill-and-practice	control-experiment	2-5th graders from 52 schools	two years	pre/post-tests, survey	null
Schenke et al. (2014): ST Math [Sch14]	Academic Standard Aligned, spatial-visual reasoning	mini-games, drill-and-practice	control-experiment	10860 3-5th graders	two years	pre/post-tests, survey	positive in math test
Ku et al. (2014): Battle Ship & Math Kicker [Ku14]	calculation, strategy	mini-games, simulation, competition	control-experiment	51 4th graders	two years	pre/post-tests, survey	positive in math test for low-ability, positive in confidence for all

Table A.1 (continued).

Castellar et al. (2014): Monkey Tales [Cas14]	a set of math topics, visual-sort memory, multi-tasking, spatial-skills	mini-games, drill-and-practice	control-experiment	74 2nd graders	3 weeks	pre/post-tests survey	positive math test and affective outcomes
Bakker et al. (2015): Rekenweb [Bak15]	a set of math topics	mini-games, drill-and-practice	control-experiment, comparative study	1661 2-3rd graders	10 weeks	pre/post-tests, in-game logs	mixed
Martin et al. (2015): Refraction [Mar15]	fraction	problem-solving puzzles	game learning analytics	24 10-11 years old; 4128 8-9 years old	140 minutes	pre/post-tests, in-game logs	NA
Vrugte et al. (2015): Zeldenrust [Vru15]	proportions, reasoning	role-playing, simulation	comparative study	242 11-15 years old	200 minutes over 4 sessions	pre/post-tests, in-game logs	NA
Derboven et al. (2016): Monkey Tales [Der16]	a set of math topics, visual-sort memory, multi-tasking, spatial-skills	mini-games, drill-and-practice	user study	8 10-11 years old	6 months	diaries, interviews	NA

Table A.1 (continued).

Brezovszky et al. (2016): the Number Navigation Game [Bre15]	number pattern, number operation, strategy	simulation, problem-solving puzzle	user study	23 6th graders	7 45 minutes sessions	pre/post-tests	NA
van der Mass and Nyamsuren (2017): Maths Garden [MN17]	selected math topics, visospacial working memory	mini-games, drill-and-practice	game learning analytics	177880 on-line players	6 months	in-game logs, in-game survey	NA
Tucker and Johnson (2017): Motion Math [TJ17]	academic standard aligned, motor skills	mini-games, drill-and-practice	user study	10 5th graders	2 around 30 minutes session	interview, field observations	NA
Vandercruyssen et al. (2017): Zeldenrust [Van17]	proportions, reasoning	role-playing, simulation	comparative study	64 14017 years old	4 50 minutes sessions	pre/post-tests, survey	NA
Vrugte et al. (2017): Zeldenrust [Vru17]	proportions, reasoning	role-playing, simulation	comparative study	93 12-15 years old	200 minutes over 4 sessions	pre/post-tests, in-game logs	NA

Table A.1 (continued).

Mclaren et al. (2017): Decimal Points [McL17]	fraction	mini-games	control-experiment	153 6th graders	7 45 minutes sessions	pre/post-tests, survey	positive in math test and attitude, null in confidence
Masek (2017): Abydos [Mas17]	curriculum-aligned fraction	simulation, role-playing	control-experiment	131 6th graders	8 20 minutes sessions over 4 weeks	pre/post-tests	positive in math test
Hieftje (2017): Knowledge Battle [Hie17]	academic standard aligned	mini-games, role-playing, drill-and-practice	control-experiment	134 1st graders	8 8-10 hours across 4 weeks	pre/post-tests, survey	Null in math test
Arroyo et al.(2017): EstimateIT! [Arr17]	measurement, number sense	tangible manipulatives, problem-solving puzzle, collaboration	user study	various small groups (<53) 4-6th graders	short sessions in multiple cites	pre/post-tests, survey, interview, field observations	NA
Bessera et al.(2017): NA [Bes17]	basic arithmetic	mini-games, drill-and-practice	user study	110 2nd graders	14 30 minutes sessions	pre/post-tests, survey, interviews, field observations	NA

Table A.1 (continued).

Lazem and Jad (2017): Multiplication Facts Game [LJ17]	multiplication facts	drill-and-practice, tangible manipulatives, pair-wise competition	user study	10 8-9 years old, 2 4th-6th graders	8 days	pre/post-tests, survey, interviews, field observations, in-game logs	NA
Sjoden et al. (2017): Squares Family [Sjö17]	place-value, spatial reasoning, strategical thinking	teachable agent, collaboration, pair-wise competition	game learning analytics	163 4th graders	7 weeks 40-50 minutes per week	in-game logs	NA
Kourakli et al. (2017): Kinect Mathloons [Kou17]	math calculation, motor skills	mini-games, drill-and-practice	user study, game learning analytics	20 6-11 years old	7 weeks	pre/post-tests, interviews, in-game logs	NA
Gresalfi et al. (2018): Motion Math [Gre18]	academic standard aligned, motor skills	mini-games, drill-and-practice	control-experiment	90 3rd graders	2 30 minutes sessions	pre/post-tests, survey, interview	null in math test, positive in enjoyment

Table A.1 (continued).

Ternblad et al. (2018): Magical Garden [Ter18]	number concept	mini-games, teachable agent	user study	43 3 years old	20 minutes sessions 2-3 times per week for 5 weeks	in-game logs	NA
Kiili et al. (2018): Semideus [KK18]	fraction concept	mini-games, drill-and-practice	comparative study	51 6th grade	30 minutes	survey, in-game logs	NA
Kosmas et al. (2018): Kinect [Kos18]	selective math skills, visual-audio-spatial memory, concentration, attention, speed	mini-games	user study	31 6-12 years old	5 months	pre/post-tests, survey, in-game logs, interview	NA