Peeking through the Classroom Window : A Detailed Data-Driven Analysis on the Usage of a Curriculum Integrated Math Game in Authentic Classrooms

Preya Shabrina
pshabri@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Ruth Okoilu Akintunde
rookoilu@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Mehak Maniktala
mmanikt@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Tiffany Barnes
tmbarnes@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Collin Lynch
cflynch@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Teomara Rutherford
teomara@udel.edu
University of Delaware
Newark, DE, USA

ABSTRACT

We present a data-driven analysis that provides generalized insights of how a curriculum integrated educational math game gets used as a routinized classroom activity throughout the year in authentic primary school classrooms. Our study relates observations from a field study on Spatial Temporal Math (ST Math) to our findings mined from ST Math students' sequential game play data. We identified features that vary across game play sessions and modeled their relationship with session performance. We also derived data-informed suggestions that may provide teachers with insights into how to design classroom game play sessions to facilitate more effective learning.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → Classification and regression trees.

KEYWORDS

Curriculum Integrated Math Games, Game Analytics, Integration

ACM Reference Format:
Preya Shabrina, Ruth Okoilu Akintunde, Mehak Maniktala, Tiffany Barnes, Collin Lynch, and Teomara Rutherford. 2020. Peeking through the Classroom Window : A Detailed Data-Driven Analysis on the Usage of a Curriculum Integrated Math Game in Authentic Classrooms. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3375462.3375525

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375525

1 INTRODUCTION

Studies have shown that educational games can be effective for student learning, both in primary school math [4], and in general [3, 25]. The popularity of educational games have increased over time and they are becoming more common in classrooms to supplement traditional human instruction and existing curricula[2].

Considerable research has studied how to design educational games to support learning better (e.g., analyzing requirements of educational game design in online education and developing a general adaptive game design method [21], measuring students' learning motivation in game-based learning environments [9], integrating concept mapping to improve learning from games [16], game designing involving teachers to develop games more aligned to curricular or situational needs [19]). Other research has focused on identifying students in need of help (e.g., early identification of students who may not complete many game levels in ST Math during a 20-minute session [14], identifying students who are not progressing well using features extracted from eye-tracking and facial expression recognizing sensor data that was generated when students interacted with a game [11]).

On the other hand, only a few studies have focused on how teachers integrate games in traditional classrooms. For example, in a field study, Peddycord-Liu et al. observed classrooms when the teachers conduct gameplay sessions and their activities surrounding the gameplay [23]. Kangas et al. also shed light on teachers' pedagogical activities surrounding digital and non-digital games in their literature review [17]). More research is needed that focuses on the strategies teachers use to integrate educational games into their classroom environment. For year-long curriculum-integrated games that are used on a daily basis, this is even more imperative.

Our study focuses on identifying how a year-long curriculum-integrated educational math game, Spatial Temporal Math (ST Math), is used in elementary/primary school classrooms using data-driven methods. This study was inspired by the field study conducted by Peddycord-Liu et al. [23] observing the practices of teachers who use ST Math alongside traditional instruction in classrooms. We relate the field observations in the classrooms of eight teachers coming from six schools with varied experience levels in using ST Math mentioned in Peddycord-Liu's field study to our findings obtained from a data-driven analysis involving 42,515 students and 1,281 teachers from five different grades and 54 schools. In addition to analyzing ST Math usage embedded in the daily class schedule, we investigated its year-round use. We identified 12 features that varied from session to session and modeled their association with session performance to derive specific suggestions informing teachers about the potential positive or negative impacts of these features. The method presented in the paper can be applied to gameplay data from other games to derive data-driven suggestions for improving and refining the use of each particular game in the classroom.

1.1 Spatial Temporal Math (ST Math)

ST Math is an educational math game that is used as a supplemental program and is aligned to Common Core and other state standards. The game is developed by MIND Research Institute and is currently used by over 1.2 million students and 56,000 teachers in 48 states in the USA. ST Math uses spatial puzzles to illustrate mathematical concepts, restricting the use of text and providing visual simulations and feedback [28]. The game is organized as a hierarchy [Figure 1] of objectives (general mathematical concepts), games, levels, and puzzles. ST Math has around 50 objectives for each grade. Each objective contains 5-10 games and each game contains 1-10 levels. For example, “Division Concepts” is an objective for 3rd graders that has 6 games. The game, “Fair Share,” under this objective has three levels that contain puzzles asking students to equally distribute blocks among animals. The students are given a set of lives at the beginning of each level. To pass a level, the students must solve all the puzzles in that level without losing all of their lives. If the students fail to do so, they are required to restart the level in order to progress. When students finish attempting a level, the number of puzzles passed and the total number of puzzles in that level are recorded to indicate performance. ST Math levels within the same game focus on the same general concept and are organized in order of increasing difficulty. A student must pass all the levels in each game under an objective to pass that objective. Within each ST Math objective, students take a pre-test, training levels, and a post-test [24]. Each district prescribes a sequence of objectives for students to learn, but teachers can rearrange objectives for their class or individual students.

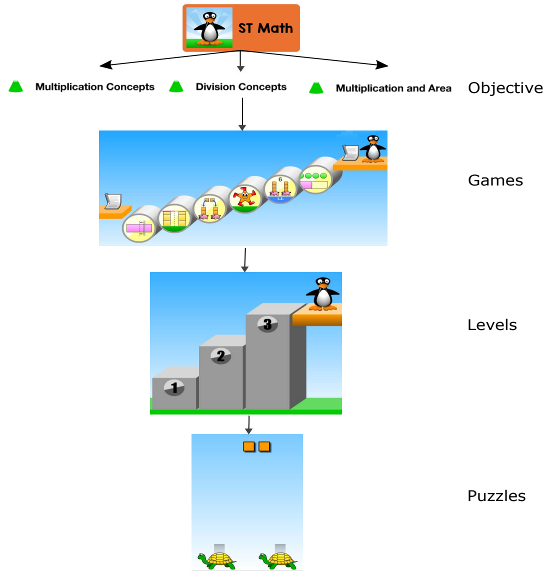


Figure 1: ST Math Hierarchy (© 2019 MIND Research Institute)

2 LITERATURE REVIEW

Curriculum integrated math games are enriched with wide-ranging gaming content to complement human instruction while teaching mathematical concepts. These games can be played throughout the academic year as a day-to-day classroom activity (like ST Math) to go on side by side of traditional curricula [29].

Curriculum integrated math games, like ST Math, Motion Math, and Astra Eagle, have demonstrated ability to improve students' math test scores [18, 26, 29]. But, their efficacy depends on the effective combination of games with classroom activities and traditional teaching instruction and strategies [4, 8, 15, 23, 27]. Using gaming technology in classrooms is particularly challenging for teachers [23]. A study conducted by Watson et al. asked teachers using video games in classrooms to list the barriers they faced in implementing games in schools. They listed "It is difficult to manage a gaming class" as one of the significant barriers [30]. Zhao et al. have shown that integrating any technology in classrooms is a complex and messy process [31]. The responsibility of successful technology integration is carried out by teachers who often face multi-directional management problems [10], including aligning game content with learning objectives [20, 30] and time and classroom management [20]. These problems can be more frequent in cases of curriculum integrated games when used as a day-to-day classroom activity. To understand and address the multifaceted problems faced by teachers, it is crucial to understand the practices and strategies they follow while integrating digital games in classrooms.

Only a few studies have focused on the practices that teachers have developed to integrate digital games in their classrooms and their impact on students' learning outcomes [17, 23]. Ke conducted a field study to observe the association of classroom structures with students' learning attitude and performance and found that cooperative goal structure classrooms where peers cooperate with each other are most beneficial [18]. A recent study conducted by Callaghan et al. observed how teachers integrate games into instruction and the associations between teaching practices and student achievement [8]. They suggested that school differences may contribute to both teaching practices and student outcomes. Hangh et al. investigated how teachers used educational games in classrooms and found out that teachers shift between roles to meet the demands of different game modalities and situations during a gameplay session [13].

Our study focused on the classroom usage of ST Math while investigating the classroom formats used in gameplay sessions and the session performance profiles in those class formats. We extended Peddycord-Liu's field study by drawing observations on a year-long basis and by extending their descriptive work to link between classroom format and session performance. Peddycord-Liu et al. [23] observed classrooms when ST Math gameplay sessions were conducted and interviewed the teachers who conducted those sessions. They observed that the teachers develop a pedagogical framework to organize gameplay sessions in three different types of class formats: lab, free, and rotation seating. In lab seating, all students used ST Math at the same time in a computer lab with one student at each computer. In free seating, students used portable devices and were allowed to move around the classroom. In rotation seating, only some students play ST Math at a station in the room while other children did other activities. We sought to identify these three formats in ST Math gameplay sessions, but the key difference in our methodology is that we extracted the class formats from students' sequential gameplay data. In addition to identifying ST Math class formats and session performance profiles in those class formats through gameplay data, we identified 12 features that varied across sessions and performed systematic analysis to formalize their associations with session performance, with an aim to provide teachers with suggestions on how to design a gameplay session in a traditional classroom. We also analyzed how ST Math gets used throughout the year in a first-grade classroom.

The field study conducted by Peddycord-Liu et al. was limited in the sense that it focused only on ST Math usage by volunteer teachers in a single district. Also, the field study was conducted at the end of an academic year when teachers were done with most of the ST Math content, leaving open questions around ST Math usage in classrooms throughout the year. In contrast, our study uses students' year-around gameplay data across five grade levels and 54 schools within a single district.

3 DATA

MIND Research Institute provided the researchers with ST Math students' sequential gameplay data. The data contain one transaction for each level attempt made by the students. Each transaction comprises a student identifier, timestamp indicating level attempt completion, login session identifier (When students log in to the system to play, this id is recorded against each level played before logging out), login place (home or lab), objective #, game #, level #, count of prior attempts for that level, count of previously passed levels, number of puzzles passed, and number of total puzzles in that level. MIND collected these data and provided them to the researchers on a yearly basis along with students' basic information (institution id, grade, teacher id, and annual progress on ST Math). Six hundred thirty eight students (0.015% of all students) whose information were missing were excluded from our study. None of the IDs revealed true identification of the human subjects (students or teachers) involved in this research.

4 METHODOLOGY

4.1 Cleaning and Preprocessing

The cleaning step involved removal of duplicate data and data with erroneous puzzle counts that recorded 0 as the count of total puzzles in a level. Our study involved analyzing students' gameplay sessions in authentic classrooms. Thus, we removed all level attempts played from home or outside school hours. The server time zone that recorded the timestamp associated with the level attempts was unknown. To determine school hours in our chosen district for the study, we observed the distribution of students throughout the hours of a day and found that density of students playing ST Math was highest during an 8-hour period (8:00 PM - 4:00 AM), which we identified as the time students spent in school. We aligned all of these ranges to fit normal school hours of 8:00 AM - 4:00 PM. All level attempts outside this range were removed. Performance for each level attempt was calculated as puzzles passed divided by the total number of puzzles.

4.2 Descriptive Statistics

For our study, we used sequential gameplay data of first-fifth grade students from four consecutive academic years (2014-15, 2015-16, 2016-17, 2017-18). Our study involved 42,515 students and 1,281 teachers coming from 54 schools in a single district in Mid-Atlantic USA. Each academic year contains data from August-July. According to the district's public schools' traditional year calendars, a typical academic year starts in the middle of August and ends in the middle of June. Students played ST Math throughout this time in school.

Here we present some general characteristics of the cleaned dataset. Our dataset had 18,670,512 (18M) level attempts among which 9,911,753 (10M) were passed level attempts and 8,758,759 (9M) were failed level attempts. Throughout the academic year, each student made an average of 419 level attempts. 50% of the total students passed a level on their first day with ST Math. On average, a student passed 54.24% of all levels they attempted. Five hundred fifteen students (2014-15 : 57, 2015-16 : 193, 2016-17 : 159, and 2017-18 : 106 students) were found who never failed a level and 167 students (2014-15 : 1, 2015-16 : 77, 2016-17 : 71, and 2017-18 : 18 students) were found who failed all levels attempts. Average student performance recorded was 0.67. The maximum number of level attempts were found in the months of January-March and September-November. The number of level attempts decreased gradually from April to June. July and August had lowest number of level attempts (and these were excluded from the study considering them summer-time gameplay). The number of level attempts decreased in December compared to November. On average, students completed 52% of their assigned ST Math content by the end of an academic year; this is consistent with prior research on ST Math [28]. Only 8,302 students (2014-15 : 1760, 2015-16 : 2844, 2016-17 : 2962, and 2017-18 : 736 students) were found who completed the objectives within the set assigned for the district. 233 students (2014-15 : 34, 2015-16 : 66, 2016-17 : 57, and 2017-18 : 76 students) were found who made no progress throughout the academic year. These students were included in our study, because they participated in classroom gameplay sessions even if they didn't progress.

4.3 Identifying Gameplay Sessions

In our study, we defined an ST Math gameplay session as a time within school hours, when students played ST Math with their classmates. Identifying sessions involved grouping level attempts played in the same session and assigning each of the level attempts a session number.

We observed from the data that, most of the time, students played for less than 45 minutes at a stretch. Also, the district's teachers' union contract states that the number of students in elementary school classrooms should not exceed 25. Thus, we focused on finding sessions that were no more than 45 minutes long and had 5-25 participants. We defined the instances where fewer than 5 students participated as "Random Gameplay," considering that they do not represent a significant portion (>= 25% of total students) of a classroom. We eliminated the random gameplays from our analysis to focus on larger classrooms, because management issues in larger classrooms are more challenging than in smaller classrooms [22] and teachers have to adjust pedagogical strategies a great deal to be effective in larger classrooms [6]. Also, we wanted to find groupings that reflected the patterns found in actual classrooms observed by Peddycord-Liu et al. [23]. Although our session marking strategy favors sessions shorter than 45 minutes, it allows longer sessions or sessions with participants below or above our defined limit to form. We filtered these unusual sessions later.

To identify ST Math sessions, we grouped the data using institution id, grade, teacher id, and date. Then, the level attempts that occurred in the same day were sorted in order of increasing timestamp and each level attempt was assigned a session number in the following two phases:

- **Phase 1:** Consecutive level attempts with 10 minutes or less in between were marked with the same session number.
- **Phase 2:** Phase 2 merged sessions identified in Phase 1 using one of the three rules mentioned below and assigned the final session number:
  - **2.1:** If consecutive sessions found in the first phase were within the same 45-minute time frame, they were merged. Else 2.2 or 2.3 were followed.
  - **2.2:** If consecutive sessions have a time gap ranging from >10 minutes to 15 minutes (gaps more than 15 minutes can be indicative of lunch break or recess period) and have the same students (at least 50% common students) then they were merged considering the second session as the continuation of the first session.
  - **2.3:** If consecutive sessions have time gap more than 10 minutes with different sets of students then we assumed that a new session has started because merging the two sessions will result in a session much longer than 45 minutes.

The algorithm we designed to identify sessions, called the session identification strategy, is illustrated in figure 2. We identified 586,026 sessions using our session identification strategy from which we excluded sessions longer than 45 minutes; this resulted in 463,732 sessions. Among those sessions, 65% had 1-2 students (one student sessions: 234,116(50.5%); two students sessions: 67,312(14.5%)); 13.8% had 3-4 students (three students sessions: 38,180(8.2%); four students sessions: 25,653(5.5%)). 162 sessions had greater than 25 students. After excluding sessions with fewer than 5 or greater than 25 students, we had 98,309 sessions (21.2% of total sessions).

4.4 Feature Generation

We generated 12 features that we call the components of classroom integration for each of the 98,309 classroom-based gameplay sessions as derived in section 4.3. These components vary from session to session and the teachers might need to take a decision regarding them while conducting gameplay sessions.

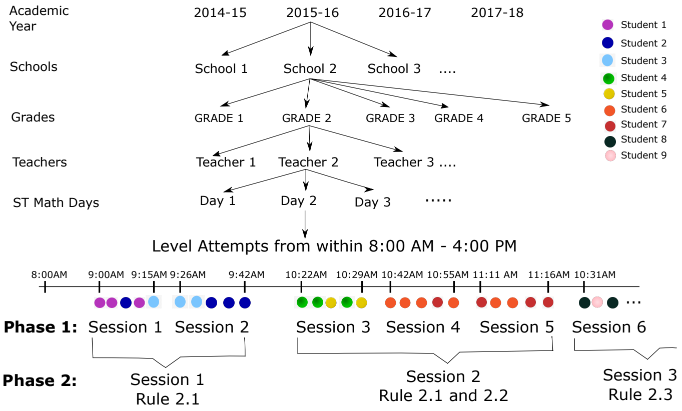


Figure 2: Identifying ST Math Gameplay Sessions

Table 1: Classroom Session features

Feature	Mean	SD	cutoff
Session length (min)^2	28.05	9.64	> 60
Class size^2	9.43	4.25	> 21
Avg. # of level attempts^1	5.35	2.49	> 14
Avg. min. played/student ^{1}	7.78	5.96	> 28
Max. % of class ^{2}	0.77	0.2	< 0.1
Start time variance (min.)^2	77.48	71.23	> 312
Finish time variance (min.)^2	63.56	69.27	> 296
# previous ST Math days	.	.	.
# previous sessions	.	.	.
% students practiced	0.13	0.26	.
Binary Features	**No**	**Yes**	.
Disjoint small group?	78341	5317	.
Before-session practice?	43779	39879	.

^{1}Cutoff at 3 standard deviations ^{2}Cutoff using Interquartile range

We also calculated session performance, which is the average performance of all students in a session. Table 1 lists the features and details on outlier removal that resulted in 83,658 sessions (2014-15 : 16,805, 2015-16 : 25,760, 2016-17 : 28,257, and 2017-18 : 12,836 students) that we used in two different studies. In our first study, we assigned each of the sessions with a class format to get an insight of how the sessions were conducted in authentic classrooms and analyzed ST Math usage throughout the academic year. In our second study, we modeled the association between the components of classroom integration and session performance.

5 STUDY 1: IDENTIFYING CLASS FORMATS IN ST MATH SESSIONS

Peddycord-Liu et al. [23] observed three different class formats and how they related with students' communication during gameplay: lab-seating, free-seating, and rotation-seating. In lab-seating, most of the students play throughout the whole class sitting in a fixed spot and the teachers are able to see their computer screens. In free seating, most of the students play throughout the whole class sitting as individuals or in groups and are allowed to move freely around the class while they play on portable devices. In rotation-seating, disjointed groups of students play in rotation.

Because our analysis involved identifying class formats from gameplay data, we formalized their definition for ease of computation. We defined those class formats as lab-seating where all students' start time and finish time varied by no more than 5 minutes and renamed the format as Low Variance Start/End Times format. The intuition behind this definition was that, because students all must leave their classroom to go to a lab, and then return, their start and end time would have a lower variance. On the other hand, in free-seating, students can use mobile devices, which to some extent limits teacher's ability to monitor their screens. Moreover, students can move around the class and likely spend time finding their desired seating position. Thus, in free-seating, it's likely that students' start and finish times will vary to a greater extent. We marked those classrooms as free-seating where students' start and finish times varied by more than 5 minutes and renamed the format as High Variance Start/End Times format. For rotation-seating, our definition is the same as Peddycord-Liu et al., which refers to the format where small groups of students play in rotation. However, we tried to find less restrictive rotational patterns where we include in the Small Group Rotation format groups that contain some overlap among students.

5.1 Identified Class Formats

Based on our definition of the three class formats, we marked each of the 83,658 sessions with one of the class formats and calculated the average session performance for each of the formats. Our results are summarized in table 3. The three formats are illustrated using one randomly selected sample session from each category in figure 3. In each of the three figures, student ids are represented along the y-axis and the times of their level attempts are represented along the x-axis with blue dots of a different shade for each student.

From figure 3a, we can see the level attempts of students in a Low Variance Start/Finish Times Session where each of the students started in between 12:17PM - 12:22PM and finished in between 12:33PM - 12:37 PM. Figure 3b plots the level attempts of students in a High Variance Start/Finish Times session where students with ids 2, 3, 4, 6, and 8 started around the same time whereas student with ids 1, 5, and 7 joined almost after 13 minutes after student 8 had started. Figure 3c demonstrates the Small Group Rotation format where we can see three disjointed groups of students played after one another with each group having a gameplay session of around 10 minutes.

Our results [table 3] suggested that the district's elementary schools favored High Variance Start/End Times format. The results demonstrate that the Low Variance Start/End Times Format had the highest average performance. Average class size and class length were also higher in this class format than the other two class formats.

5.2 Gameplay Sessions from Teachers' Perspective

We grouped the sessions by teachers and prepared summary statistics for the components of classroom integration from teachers' perspective. The statistics are shown in table 2. The statistics suggest that, on average, teachers conduct classes that are 28.5 minutes long where students play around 6 levels in 13.82 minutes. The statistics also revealed, although each teacher had about 21 students per grade, the average number of students per session was approximately 9. This is indicative of the fact that most of the time they do not engage all the students they instructed in a particular grade in a particular academic year. We calculated the maximum number of students who played at the same time at some point during a session as a fraction of total participating students to get an estimate of availability of technology in elementary classrooms. The average value of this feature is 0.77, and we found only 5,249 sessions where all of the students participating in a session played at the same time at some point, which might be indicative of insufficient technology availability in elementary classrooms.

5.3 How Students Played ST Math throughout the year

In order to get an understanding of year-long gameplay in ST Math, we visualized the way one classroom played ST Math throughout the year.

5.3.1 ST Math Days. Figure 4 shows how one first grade classroom played the game in 2014 with days of the week on the x axis and number of students on ST Math during that day on the y-axis. The bulk of students logged on together on Fridays from October through January and logged on together on Tuesdays and Fridays from February to April. Notice that from May to June, students played ST-Math together almost everyday.

5.3.2 ST Math Times. We also extracted the times when students played ST-Math together on Tuesdays and Fridays. We noted that students played together between 2pm and 3pm on Fridays from November to January and played together between 10am and 1pm on Fridays and from January to April they played on Tuesdays, but no pattern in their play time could be found. This may indicate that the class mainly played ST-Math on Fridays during these months and Tuesday play was supplemental. In addition, ST-Math was played for a longer period by majority of the students on Tuesdays, Wednesdays, Thursdays, and Fridays in May and June.

Our observations suggest that ST Math gameplay sessions were not evenly distributed throughout the year. Play appeared to ramp up throughout the year, starting with one day, moving to two days a week during the middle of the school year, and ending with nearly daily play.

6 STUDY 2: MODELING THE ASSOCIATION BETWEEN COMPONENTS OF CLASSROOM INTEGRATION AND SESSION PERFORMANCE

6.1 Session Performance as a Continuous Variable

To model the relationship between the components of classroom integration shown in table 1 and session performance as a continuous variable, we used two different mixed effect models : Mixed Effect

Table 2: Statistics of Gameplay Sessions in Classrooms Grouped by Teachers

	Students Assisted	Number of Sessions Conducted	Class Length (minutes)	Level Attempts	Time Spent in Game Play (minutes)	Number of Students per Session	Maximum Participants as a Fraction of Total Students in the Session	Class Performance
Mean	~21	~78	28.5	~6	13.82	~9	0.77	0.68
Median	21	61	28.45	~5	13.50	~9	0.77	0.68
Max	~70	366	43.63	~11	27.72	~20	1	1
Min	5	1	6.37	~1	1.14	5	0.42	0.20
Standard Deviation	6	~64	3.28	~1	3.07	~2	0.08	0.09

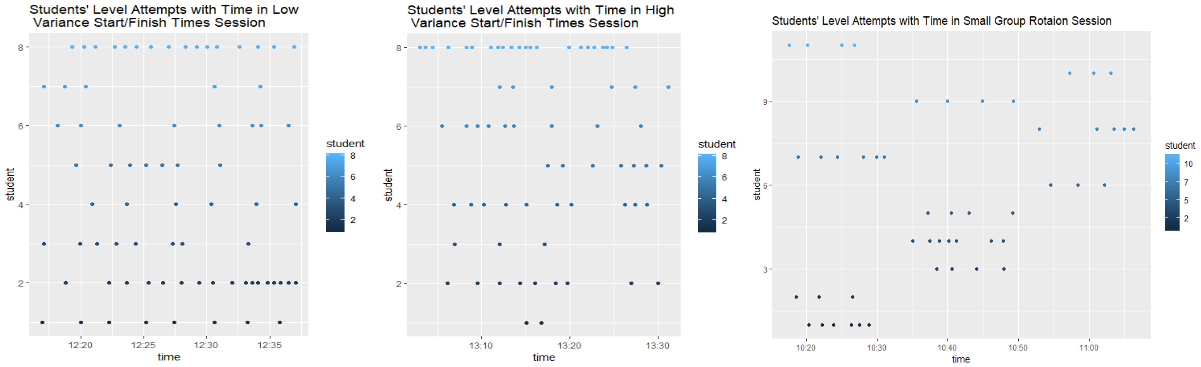


Figure 3: a) Low Variance Start/Finish Times Session; b) High Variance Start/Finish Times Session and c) Small Group Rotation Session

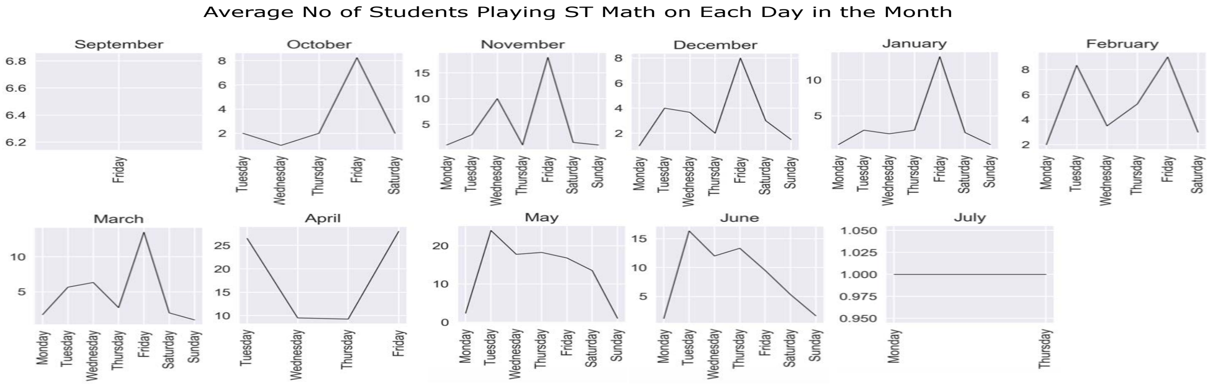


Figure 4: How One First Grade Classroom Played ST-MATH in 2014

Linear Regression(MELR) [1] model and Mixed Effect Random Forest(MERF) Model [1, 12]. While modeling the association between certain features and some target variable using clustered data in mixed effect models, each cluster is considered as random effect and the features are considered as fixed effect variables. Mixed effect models estimate the association between the fixed effect variables and the target variable independent of the random effect variables. The MELR model assumes a linear model of the form: $Y = aX + b_i Z + e$. In this equation, Y refers to the target variable, X are the fixed effect features and Z are the random effect features that introduce clusters in the data. MELR learns the fixed effect coefficients a, random effect coefficients b_i and noise e using iterative algorithms. On the other hand, the MERF model assumes a non-linear model of the following form $Y = f(X) + b_i Z + e$. The notations used in this equation are similar to MELR, except for $f(X)$ which is a non-linear function that MERF learns using Random

Table 3: Class Formats found in the Gameplay Sessions

	Low Variance Start/End Times Session	High Variance Start/End Times Session	Small Group Sessions
Count	8,787	69,554	5,317
Mean Session Performance	0.71	0.69	0.67
Mean Class Size	7.3	9.8	7.7
Mean Class Length (min)	16.2	29.8	25.4

Forest. In addition to training the mixed effect models, we extracted p-values (p-value>0.05 indicates significant association) from MELR and feature importance as % increase in MSE from MERF to identify which features had statistically significant associations with session performance.

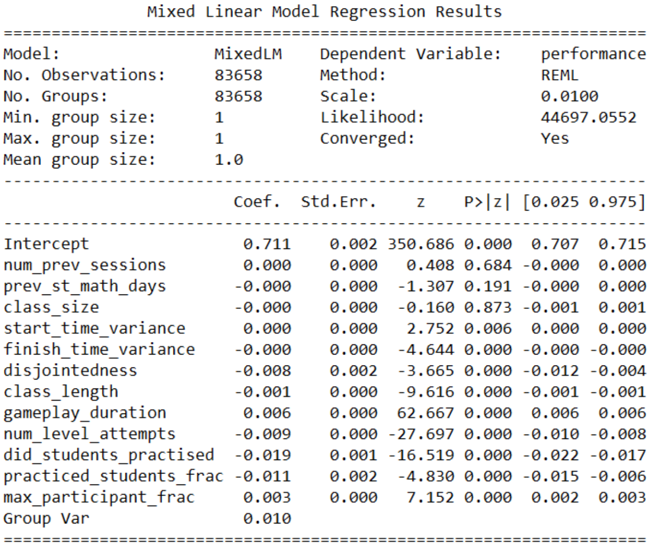
In our study, we used grade, teacher id, student id, objective content label as random effect variables to estimate association between components of classroom integration [table 1] and session performance net of any variance due to clustering. We used 70% of the total data to train the models and rest of the data were used for testing.

6.1.1 Findings. The reported mean squared error (MSE) for our trained MELR and MERF on test set was 0.0198 and 0.0161 respectively. The MSEs suggest that MERF generated a slightly better fit. The p-value calculated by the MELR model suggested all of the features apart from num_prev_sessions, prev_st_math_days and class_size had statistically significant associations with session performance. The summary of the generated model is shown in figure 5.

To get feature importance from our trained MERF model, we calculated % increase in MSE for each feature using the permutation importance technique [7] with the equation: importance of feature X = (MSE on test set with noise in place of feature X - MSE on test set with all features) / MSE on test set with all features. The feature importance values obtained using this method are shown in table 4. According to these values, MERF identified gameplay_duration, num_level_attempts, practised_students_frac, did_students_practised, and max_participant_frac as the most important features.

Table 4: p-values Obtained for the 12 Session Features from MERF Model)

Feature	Feature Importance (% increase in MSE)
gameplay_duration	**1.3028**
num_level_attempts	**1.3093**
practiced_students_frac	**1.3173**
did_students_practised	**1.3138**
max_participant_frac	**1.3271**
num_prev_sessions	0.0008
prev_st_math_days	0.0290
class_size	0.0350
start_time_variance	0.0378
finish_time_variance	0.0600
disjointedness	0.0598
class_length	0.0969



```
               Mixed Linear Model Regression Results
===================================================================
Model:                 MixedLM    Dependent Variable:   performance
No. Observations:      83658      Method:               REML
No. Groups:            83658      Scale:                0.0100
Min. group size:       1          Likelihood:           44697.0552
Max. group size:       1          Converged:            Yes
Mean group size:       1.0
-------------------------------------------------------------------
                        Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------------------
Intercept                0.711    0.002 350.686 0.000  0.707  0.715
num_prev_sessions        0.000    0.000   0.408 0.684 -0.000  0.000
prev_st_math_days       -0.000    0.000  -1.307 0.191 -0.000  0.000
class_size              -0.000    0.000  -0.160 0.873 -0.001  0.001
start_time_variance      0.000    0.000   2.752 0.006  0.000  0.000
finish_time_variance    -0.000    0.000  -4.644 0.000 -0.000 -0.000
disjointedness          -0.008    0.002  -3.665 0.000 -0.012 -0.004
class_length            -0.001    0.000  -9.616 0.000 -0.001 -0.001
gameplay_duration        0.006    0.000  62.667 0.000  0.006  0.006
num_level_attempts      -0.009    0.000 -27.697 0.000 -0.010 -0.008
did_students_practised  -0.019    0.001 -16.519 0.000 -0.022 -0.017
practiced_students_frac -0.011    0.002  -4.830 0.000 -0.015 -0.006
max_participant_frac     0.003    0.000   7.152 0.000  0.002  0.003
Group Var                0.010
===================================================================
```

Figure 5: Summary of MELR Model

6.2 Prediction of Performance as Binary

We generated a binary feature perf_rank that had the value 1 for the sessions with performance greater than 0.7 and 0 otherwise. 41,781 sessions had perf_rank 1 and 41,877 had perf_rank 0.

We split our entire data set into train and test sets in a 70% – 30% ratio. Then, using the train set, we trained a Logistic Regression model with recursive feature elimination and Random Forest models (with Averaging, Adaptive Boosting [Ada Boost], and Gradient Boosting with embedded feature selection based on impurity reduction) with 10-fold cross validation using the 12 components of classroom integration to predict perf_rank for each of the sessions and extracted the ranking and contribution of each feature in the models. The ranking and performance summary for each model are shown in table 5 and 6. The summary demonstrated that the random forest models suffered from poor recall for perf_rank 0 and poor precision for perf_rank 1, which indicates that although the features collectively were quite successful in capturing the perf_rank 1 cases, they fell short to capture the 0 cases i.e., poor performance. From this we concluded that poor session performance occurs due to factors beyond the components of classroom integration.

6.3 How Each Feature Impacts Session Performance

We performed Pearson's correlation tests between each feature and session performance and used the p-value (p-value < 0.05 indicates significant correlation) obtained from the test to identify statistically significant correlations and plotted smooth curves of session performance against each of the features to reason about how the features have an impact on performance and how they affect the models.

The p-value for start_time_variance and finish_time_variance were 7.432e-15 and 2.2e-16 respectively, indicating that these features are highly correlated with performance. From smoothed curves of performance against start_time_variance and finish_time_-variance we observed that lowest performance in a session increases with these two features, whereas highest performance in a session shows the opposite trend (lower highest performance with higher variances). In the case of the boolean feature, disjointedness (presence of small disjoint groups playing in Small Group Rotation format), we observed that in 54% of the instances where disjointedness was 1, session performance was less than 0.7, which makes this feature a potential predictor of perf_rank 0. We concluded this as the reason why Logistic Regression performed better than the random forest models in identifying perf_rank 0 when it ranked disjointedness as one of the top features. The p-value for class_size was 2.2e-16, indicating a statistically significant correlation with session performance; the smoothed curve of session performance against class_size demonstrated the same trend as start_time_variance and finish_time-_variance (increased lowest performance and decreased highest performance with larger class_size). num_prev_session and prev_st_math_days had p-values of 1.121e-04 and 1.165e-08. Both of the features showed a positive correlation with lowest session performance and negative correlation with highest session performance. Because these two features increase as the year progresses, increasing lowest session performance with these features indicates weaker students improved throughout the semester, whereas decreasing highest session performance indicates that either the stronger students finished their assigned content early and didn't participate in the later sessions or they encountered more difficult content that better matched their capabilities. p-values of class_length (2.422e-08), practised_students_frac (2.2e-16) and max_participant_frac (2.2e-16) indicated statistically significant correlations with session performance. However, smoothed curves of session performance against these features didn't show any visually identifiable trend. Binary feature did_students_practice also failed to demonstrate a manually identifiable trend.

6.4 Session Performance with Gameplay Duration and Number of Level Attempts

Because the mixed effect models agreed on the significant impact of gameplay_duration and num_level_attempts, and these two features had a ranking between 1-5 for all the binary predictors, we further analyzed our data to gain insight into how these two features are associated with students' session performance. We grouped the data using these two features and calculated average performance for each value of gameplay_duration and num_level_attempt. We then plotted smoothed curves of performance against these two features, which are shown in figure 6. The curve of performance against gameplay_duration suggest that students' session performance is lower when gameplay_duration is less than 7 minutes than when gameplay_duration is more than 7 minutes. The curve of performance against num_level_attempts showed a decrease in session performance when more than 10 levels are played in a session. To further analyze the impact of these two features we manually constructed a decision tree [figure 7] and observed its different branches. The decision tree suggests when students play for less than 5 minutes in a session the chances of getting a session performance above 0.7 is much lower than when students play for more than 5 minutes. The branch corresponding to gameplay_duration<5 minutes suggests that playing for a very short time or attempting too many levels in a short time is associated with poor performance; this is consistent with a scenario wherein students make level attempts without trying to progress or learn–referred to as wheel spinning (a state of high effort, but little progress) [5]. The decision tree showed a maximum of 62% and 65% probability of achieving performance > 0.7 in its branches corresponding to gameplay_duration of 5-10 minutes with num_level_attempts 5-10 and gameplay_session of >10 minutes with num_level_attempts 5-14. A gameplay of moderate length (5-10 minutes) where students attempt a moderate number of levels (5-10) is associated with comparatively higher session performance. Also, the last branch of the decision tree suggests that students can do well when they play more than 10 levels if they are provided with sufficient time (no less than 10 minutes). However, the branches corresponding to num_level_attempts > 10 have an average performance of 0.64, which is less than other branches that corresponds to the curve segment with negative slope in the plot of performance against the number of level attempts [figure 6].

Table 5: Feature Ranking by the Binary Predictors

Feature	Rank			
	Logit	Random Forest (Avg.)	Ada Boost	Grad Boost
did_students_practised	1	11	11	11
disjointedness	2	12	12	12
practiced_students_frac	3	10	9	10
num_level_attempts	**4**	**2**	**5**	**4**
gameplay_duration	**5**	**1**	**1**	**1**
max_participant_frac	6	8	8	8
class_size	7	9	10	9
class_length	8	4	4	5
prev_st_math_days	9	7	6	7
start_time_variance	10	5	3	3
finish_time_variance	11	3	2	2
num_prev_sessions	12	6	7	6

7 CONCLUSION AND RECOMMENDATIONS

We analyzed the usage of a curriculum-integrated math game in traditional classrooms embedded as a classroom activity as well as its usage throughout the academic year. As part of this work we identified three different types of class formats that teachers adopt while conducting a gameplay session and noted performance profiles for the class formats. Also, We identified features that vary across sessions and modeled their association with session

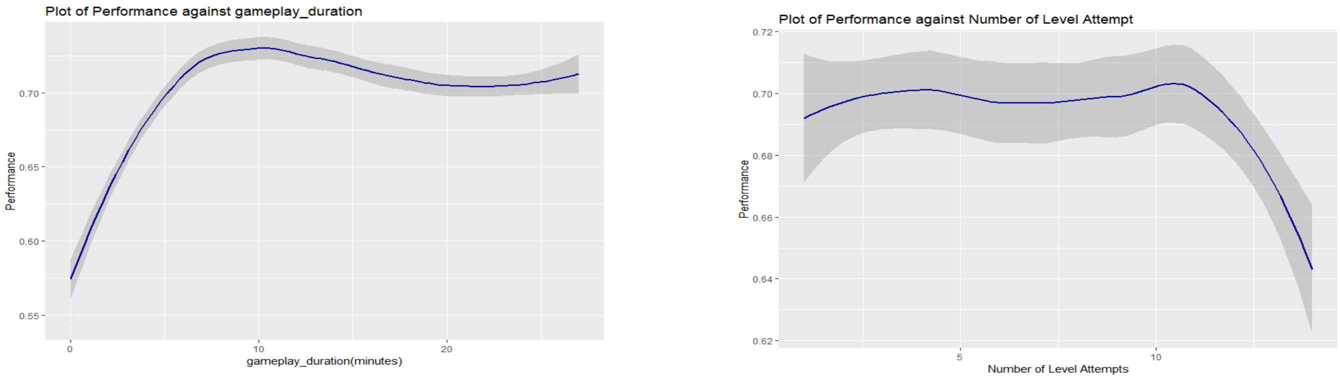


Figure 6: Change in Performance with increasing gameplay_duration and num_level_attempts

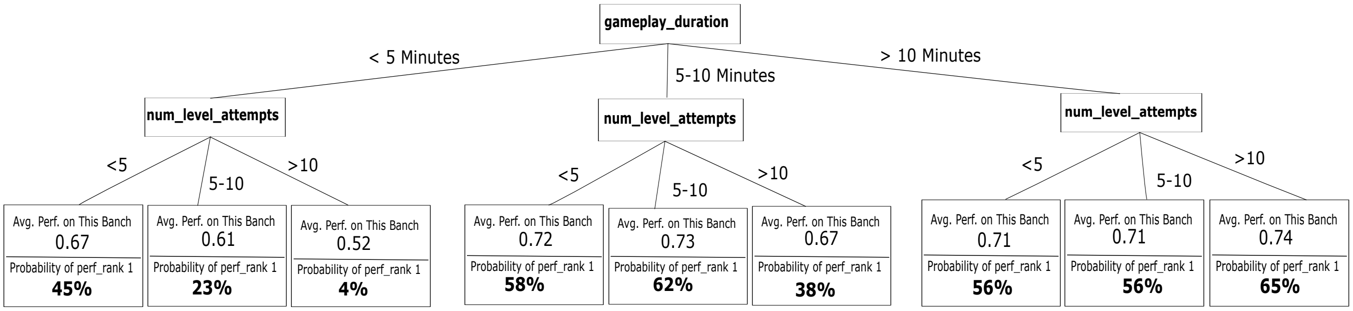


Figure 7: Decision Tree with gameplay_duration and num_level_attempts

Table 6: Performance of Different Models in Predicting Performance

Model	Perf-Rank	Precision	Recall	F1 -score	Accuracy
Logistics Regression	0	0.58	0.63	0.60	59%
	1	0.60	0.55	0.57	
Random Forest	0	0.68	0.56	0.61	65%
	1	0.62	0.73	0.67	
ADA Boost	0	0.64	0.55	0.59	61.86%
	1	0.60	0.68	0.64	
Gradient Boosting	0	0.70	0.52	0.60	64.8%
	1	0.61	0.78	0.69	

performance. Based on the findings from our study, we give the following suggestions:

- Teachers usually adopted High Variance Start/Finish Times format while conducting gameplay sessions, whereas, Low Variance Start/Finish Times formats had the highest average performance (Section 5.1). However, we observed that lowest performance in a session improved with increasing variance in start and finish times, indicating that weaker students performed better in High Variance Start/Finish Times sessions (Section 6.3). To optimize performance, teachers likely need to carefully consider student backgrounds and needs when choosing a format.
- Our study found that gameplay duration and number of level attempts have associations session performance and thus likely need careful tailoring while conducting gameplay sessions (Section 6.4). Based on our generated decision tree, we suggest that teachers should let students play a moderate number of level attempts (5-10) in a moderate amount of time (5-10 minutes). If the students need to play more levels, they should be given sufficient time to complete the game to avoid rushing and failing to improve their skills.
- Our observations on students' year-long gameplay activity demonstrated that students play a lot in a few short bursts throughout the year (Section 5.3) and our findings showed that students do not do well when they play too much in a short time period (Section 6.4). Thus, we suggest that gameplay sessions should be evenly distributed throughout the year so that students can focus more on learning rather than just completing the assigned content.
- The components of classroom integration fell short in accurately predicting poor performance (section 6.2), which suggests that poor performance has many causes beyond the classroom format used for a gameplay session.

8 LIMITATIONS AND FUTURE WORK

This study has several limitations. The correctness of the analysis conducted in this study is completely dependent on the accuracy of the session identification strategy. Data containing information on what actually happened in classrooms during ST Math sessions would be helpful to verify and adjust the strategy we used to determine which level attempts constituted classroom sessions. Moreover, we used teacher id to identify the students who belong to the same classroom. There was some indication that this was not a perfect variable and we may have classified a small portion of students as belonging to the same class when they likely had different teachers. Also, our study explored only 21.2% of the overall data and 60.6% of the 174,030 sessions with 3-20 students. We excluded all the sessions with 1-4 students, which represented 78.8% of the data, and only analyzed sessions with 5-25 students to focus on larger classrooms and those in line with those classes observed in ST Math classrooms in Z. Liu's field study (2019). Future work can combine data analytic methods with contemporaneous field studies to identify when and why sessions with 1-4 students occur and how those sessions are conducted.

Our combination of all grade levels, 1-5 and all schools in analysis presents another limitation. It is also possible that specific grade levels or schools have very different session features.

Also, the data contained only one timestamp for each transaction indicating the end time of a level attempt. No information was available regarding how long the student spent on that level. We estimated class_length and game_play_duration by calculating the difference in timestamps associated with level attempts. These features can be further perfected with additional information on how much time student spent on each level. With additional data, future work can revise the study and provide more specific suggestions to teachers on how to effectively design a gameplay session for a year-long curriculum integrated game.

ACKNOWLEDGMENTS

This research is made possible by support of the National Science Foundation under Grant No. 1544273. We would like to thank Veronica Catete, Rachel Harred, Zhongxiu Peddycord-Liu, and Ge Gao for their sincere review.

REFERENCES

[1] [n. d.]. https://towardsdatascience.com/mixed-effects-random-forests-6ecbb85cb177.

[2] SS Adkins. 2017. The 2017-2022 Global Game-based Learning Market. In *Serious Play Conference*. 1–20.

[3] Per Backlund and Maurice Hendrix. 2013. Educational games-are they worth the effort? A literature survey of the effectiveness of serious games. In *2013 5th international conference on games and virtual worlds for serious applications (VS-GAMES)*. IEEE, 1–8.

[4] Marjoke Bakker, Marja van den Heuvel-Panhuizen, and Alexander Robitzsch. 2015. Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability. *Contemporary Educational Psychology* 40 (2015), 55–71.

[5] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *International conference on artificial intelligence in education*. Springer, 431–440.

[6] Sid Bourke. 1986. How smaller is better: Some relationships between class size, teaching practices, and student achievement. *American Educational Research Journal* 23, 4 (1986), 558–571.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Melissa N Callaghan, JJ Long, EA van Es, Stephanie M Reich, and Teomara Rutherford. 2018. How teachers integrate a math computer game: Professional development use, teaching practices, and student achievement. *Journal of Computer Assisted Learning* 34, 1 (2018), 10–19.

[9] Ching-Hsue Cheng and Chung-Ho Su. 2012. A Game-based learning system for improving student's learning effectiveness in system analysis course. *Procedia-Social and Behavioral Sciences* 31 (2012), 669–675.

[10] Pierre Dillenbourg and Patrick Jermann. 2010. Technology for classroom orchestration. In *New science of learning*. Springer, 525–552.

[11] Ramla Ghali, Claude Frasson, and Sébastien Ouellet. 2016. Towards real time detection of learners' need of help in serious games. In *The Twenty-Ninth International Flairs Conference*.

[12] Ahlem Hajjem, François Bellavance, and Denis Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84, 6 (2014), 1313–1328.

[13] Thorkild Hanghøj and Christian Engel Brund. 2010. Teacher roles and positionings in relation to educational games. In *Proceedings of the 4th European conference on games based learning*. 116–122.

[14] Rachel Harred, Christa Cody, Mehak Maniktala, Preya Shabrina, Tiffany Barnes, and Collin Lynch. 2019. How Long is Enough? Predicting Student Outcomes withSame-Day Gameplay Data in an Educational Math Game. In *EDM in Games Workshop, 2019*.

[15] Cristyne Hébert and Jennifer Jenson. 2017. Digital game-based pedagogy: Exploring teaching strategies for classroom teachers in the use of video games in K-12 classrooms. In *ECGBL 2017 11th European Conference on Game-Based Learning*. Academic Conferences and publishing limited, 227.

[16] Gwo-Jen Hwang, Li-Hsueh Yang, and Sheng-Yuan Wang. 2013. A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education* 69 (2013), 121–130.

[17] Marjaana Kangas, Antti Koskinen, and Leena Krokfors. 2017. A qualitative literature review of educational games in the classroom: the teacher's pedagogical activities. *Teachers and Teaching* 23, 4 (2017), 451–470. https://doi.org/10.1080/13540602.2016.1206523 arXiv:https://doi.org/10.1080/13540602.2016.1206523

[18] Fengfeng Ke. 2006. Classroom goal structures for educational math game application. In *Proceedings of the 7th international conference on Learning sciences*. International Society of the Learning Sciences, 314–320.

[19] Javier Melero, Davinia Hernández-Leo, and Josep Blat. 2014. Teachers Can Be Involved in the Design of Location-based Learning Games. In *Proceedings of the 6th International Conference on Computer Supported Education-Volume 3*. SCITEPRESS-Science and Technology Publications, Lda, 179–186.

[20] Gerhard Molin. 2017. The role of the teacher in game-based learning: A review and outlook. In *Serious games and edutainment applications*. Springer, 649–674.

[21] Pablo Moreno-Ger, Daniel Burgos, Iván Martínez-Ortiz, José Luis Sierra, and Baltasar Fernández-Manjón. 2008. Educational game design for online education. *Computers in Human Behavior* 24, 6 (2008), 2530–2540.

[22] David Pedder. 2006. Are small classes better? Understanding relationships between class size, classroom processes and pupils' learning. *Oxford Review of Education* 32, 02 (2006), 213–234.

[23] Zhongxiu Peddycord-Liu, Veronica Cateté, Jessica Vandenberg, Tiffany Barnes, Collin F Lynch, and Teomara Rutherford. 2019. A Field Study of Teachers Using a Curriculum-integrated Digital Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 428.

[24] Zhongxiu Peddycord-Liu, Christa Cody, Sarah Kessler, Tiffany Barnes, Collin F Lynch, and Teomara Rutherford. 2017. 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*. ACM, 195–204.

[25] Josephine M Randel, Barbara A Morris, C Douglas Wetzel, and Betty V Whitehill. 1992. The effectiveness of games for educational purposes: A review of recent research. *Simulation & gaming* 23, 3 (1992), 261–276.

[26] Michelle M Riconscente. 2013. Results from a controlled study of the iPad fractions game Motion Math. *Games and Culture* 8, 4 (2013), 186–214.

[27] Elizabeth Rowe, Jodi Asbell-Clarke, Ryan S Baker, Michael Eagle, Andrew G Hicks, Tiffany M Barnes, Rebecca A Brown, and Teon Edwards. 2017. Assessing implicit science learning in digital games. *Computers in Human Behavior* 76 (2017), 617–630.

[28] Teomara Rutherford, George Farkas, Greg Duncan, Margaret Burchinal, Melissa Kibrick, Jeneen Graham, Lindsey Richland, Natalie Tran, Stephanie Schneider, Lauren Duran, et al. 2014. A randomized trial of an elementary school mathematics software intervention: Spatial-temporal math. *Journal of Research on Educational Effectiveness* 7, 4 (2014), 358–383.

[29] Teomara Rutherford, Melissa Kibrick, Margaret Burchinal, Lindsey Richland, AnneMarie Conley, Keara Osborne, Stephanie Schneider, Lauren Duran, Andrew Coulson, Fran Antenore, et al. 2010. Spatial Temporal Mathematics at Scale: An Innovative and Fully Developed Paradigm to Boost Math Achievement among All Learners. *Online Submission* (2010).

[30] William Watson and Sha Yang. 2016. Games in schools: Teachers' perceptions of barriers to game-based learning. *Journal of Interactive Learning Research* 27, 2 (2016), 153–170.

[31] Yong Zhao, Kevin Pugh, Stephen Sheldon, and Joe L Byers. 2002. Conditions for classroom technology innovations. *Teachers college record* 104, 3 (2002), 482–515.