The Antecedents of and Associations with Elective Replay in an Educational Game: Is Replay Worth It?

Zhongxiu Liu
North Carolina State University
zliu24@ncsu.edu

Christa Cody
North Carolina State University
cncody@ncsu.edu

Tiffany Barnes
North Carolina State University
tmbarnes@ncsu.edu

Collin Lynch
North Carolina State University
cflynch@ncsu.edu

Teomara Rutherford
North Carolina State University
taruther@ncsu.edu

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.

Keywords

Educational Games, Serious Game Analytics, Replayability

1. INTRODUCTION

"Replayability is an important component of successful games." [15] 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 paper, 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. This paper explores 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 [9]. 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 [18]. In the RETAIN model designed by Gunter et al. [10] 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." [10] Meaningful elective replay is often encouraged by game features such as score leaderboards, which inspire students to replay for higher scores [4]. 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 [6, 8].

However, few studies have investigated when students choose to replay, why they do so, or have measured the outcomes associated with elective replay. One reason 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 [14, 5]. 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.

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?

2. RELATED WORK

2.1 Factors Influencing Elective Replay

Few empirical studies have investigated the motivations behind elective replay in educational games. Burger et al. [5] studied the effect of verbal feedback from a virtual agent on replay in the context of a brain-training game. They found that elaborated feedback increases, whereas comparative feedback decreases, the students' interest in future replay. They also found that negative feedback generated an immediate interest in replay, whereas positive feedback created long term interest in the educational content. In another study, Plass et al. [14] compared three conditions in a math game: working individually, competing with another player, or collaborating with a peer. The study showed that both competition and collaboration modes heightened students' intention to replay when compared with the individual mode, with the latter result being statistically significant. However, both studies measured replay via questionnaires asking the students' desire to play the entire game again instead of observed replay behavior. Moreover, these studies sought to understand replay only from the angle of game design, and did not address the connections, if any, between student characteristics and interest in replay.

Other studies suggest elective replay is a habitual behavior that arises from individual need, although these studies did not directly investigate replay. Bartle [3] found one type of player who is primarily motivated by concrete measurements of success. In ST Math, these *achiever-type* players may largely use replay to get better 'scores' (losing fewer lives when passing a level). Mostow et al. [12] observed a student in a reading tutor who used the learner-control features to spend the majority of time replaying stories or writing "junk" stories instead of progressing to new material. Thus, some students may also use replay as a form of work avoidance – playing already passed levels instead of solving the current problem or moving on. Sabourin et al. [17] found that students in an educational game used off-task behaviors to cope with frustration, implying that off-task behavior can be a productive self-regulation of negative emotions. In ST Math, when students get frustrated with the current educational content but still have to play the game in the classroom, they may replay already learned content as a mental break from the current task. These studies showed that the circumstances of replay and students' characteristics influence their decisions to replay and its outcomes.

2.2 The Outcomes of Replay

Despite the believed benefits of replayability [9, 18, 10, 4], few studies have investigated the educational impact of elective replay. Boyce et al. [4] 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 replaying the base game with new features. In another study, Clark et al. [7] 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, while 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. These studies showed that elective replay may lead to increased learning or higher in-game performance. However, more research is needed to understand the potential educational impact of replay in educational games, particularly elective replays initiated solely by the players.

3. GAME, DATA AND FEATURES

3.1 ST Math Game

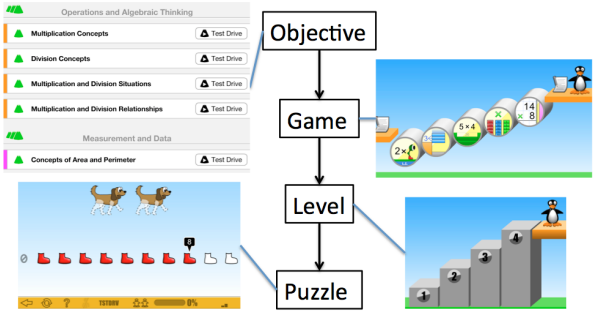

Figure 1: ST Math Content and Examples

ST Math is designed to act as a supplemental program to a school's existing mathematics curriculum. ST Math is mostly played during classroom sessions, but students have the option to play it at home. In ST Math [16], mathematics concepts are taught through spatial puzzles within various game-like arenas. ST Math games are structured at the top level by objectives, which are broad learning topics. Within each objective, individual games teach more targeted concepts through presentation of puzzles, which are grouped into levels for students to play. Students start by completing a series of training games on the use of the ST Math platform and features. They 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. Games represent scenarios for problem-solving using a particular mathematical concept, such as "finding the right number of boots for X animals of Y legs." Each game contains between one and ten levels, which follow the same general structure of the game, but increase in difficulty. Figure 1 illustrates the hierarchy of ST Math content and examples.

As with many games, the student is given a set number of 'lives' at the start of each 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, they will fail the level and be required to restart the level with a new set of lives. Once a student has passed a level, they can elect to replay it at any time. After a student has passed every level in an objective, they can take the objective post-test. Students cannot progress to the next objective until they have completed the last objective post-test. Both the objective pre- and post-tests consist of 5-10 multiple choice questions related to the objective. The post-tests parallel the pre-tests in both the question format and difficulty of the content. While answering each question in both tests, students indicate their relative confidence in their answer (low/high).

3.2 Data

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. Table 1 summarizes students' demographic information. These demographic data, together with students' state standardized test scores in 2012 and 2013, were matched to gameplay data through anonymized IDs.

Table 1: 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%

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. In the next section, we describe the features we created to analyze ER.

3.3 Features

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 2 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 2: Example of ER 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)

3.3.1 *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).

3.3.2 *Elective Replay Features*

Table 3 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. 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.

Table 3: 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 2	
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 Table2	
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 Immediate Prior Pass Attempts followed or interrupted by ER and ER Session E.g. "Division-lvl3" for all ER Sessions in Table 2	
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.

3.3.3 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 [3], avoiding progress on the current task [12], or taking a mental break from negative emotions [17]. 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 2 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 2, the student's pass attempts on "Division-Level3" was interrupted twice on the third and fifth plays, 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 replays. Thus, this student also belongs to Group DLSO, because the two groupings are independent of each other.

4. METHODS & RESULTS

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%, $c2(1, N=4827) = 17.99$, $p<.001$). 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

Table 4: Mann-Whitney U Tests Comparing Gameplay Characteristics between ER Pattern Student Groups

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 statistically significances 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)

5 columns of Table 4 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.

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.

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 relative 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 were 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 access the trees' accuracies.

Table 5 reports the 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 4, 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 5, 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 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 [4] [7], students may replay the same

Table 5: Decision Trees to Predict Levels whose Pass Attempts were Interrupted or Followed by ER

Condition: interrupted/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 $\geq$ 2.5, Yes
ER from a different objective (N=12,506)	78.7% accuracy 1st attempt performance $\geq$ 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 $\geq$ 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.

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 4, Group SL students who performed most of their ERs after the same level also achieved the highest ER performance.

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.

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 6 reports the percentage of data points that gained, dropped (mainly for avoiding ceiling effect in this data), and did not gain for each

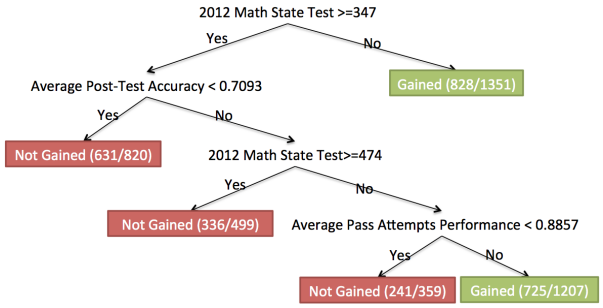

Figure 2: Decision Tree to Predict Whether a Student will Gain in State Standardized Math Test

type of gain based on the Marx and Cummings Normalization method [11].

Table 6: %Observations with Gains, No Gains, and Percentage Dropped for the Three Gains

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

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, while the remainder of the students had only a 25.9% chance. Students with high pre-test confidence ($\leq$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 achieved full scores in accuracy and 54.5% achieved full scores in confidence.

Our decision tree for the student granularity is shown in Figure 2, with a cross-validated accuracy of 57.8%. Students who started with medium level of math abilities (2012 state test math scores $<$474, and $\geq$ 347) improved their scores when they performed well in ST Math (average pass attempts performance $>$ 0.8857). This shows that the game-

play 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.

Table 7: Mann-Whitney U Tests Comparing Gains between ER Pattern Student Groups.

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 $^*p < .001$, $+p < .01$

Finally, we investigated how ER patterns relate to gains. Table 7 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 less gains, which 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 4), 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 pass a level helped students learn, or implied better learning in the previous gameplay.

5. DISCUSSION AND CONCLUSIONS

This work presents a significant extension on prior studies of replay which have typically taken place over a short period of time and have assessed replay via intentional questionnaires not observed behaviors [14, 5]. This work analyzed logged student-initiated elective replay from a sample of 4,827 3rd-5th graders during school year 2012-2013 in ST Math in a natural educational setting. We sought to answer three research questions: Q1: What are the characteristics of students who electively 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?

We concluded that, with over half of students who electively replayed at least one level, ER is a common behavior in ST Math. Moreover, examining elective replay can enhance our understanding about how students play and the characteristics of successful play. For example, we found that students who did poorly on the current level were more likely to electively replay a different level during/after the level's pass attempts. We also found that students who generally engaged in elective replay before passing the current level (Group B) started with lower pre-test scores, did worse during gameplay, and had the lowest objective-level accuracy and confidence gain and math gains. One explanation for this result is that weaker students used ER as a work avoidance tactic, as found in Mostow et al. [12], and that instances of ER stand in for lower motivation or engagement for the objective topic, ST Math, or mathematics overall.

On the other hand, compared to students who didn't ER, students who mostly electively replayed after passing the current level (Group A) started with pre-test scores that were not significantly different, did better in the game, and had higher learning and confidence gains. One reason could be that these students electively replayed for a better score, as we also found that students who mostly replayed the same level immediately after passing it (Group SL) had the highest ER performance. This association is especially true among achiever-type players [3] that prefer to gain concrete measurements of success. Because losing fewer lives in ST Math requires better mastery of the math content, ER may have helped these students learn. Another explanation is that these students' ERs could imply better learning during prior gameplay, as Table 4 also shows that Group A students had better pass attempt performance. Possibly, successful prior performance motivated these students to electively replay more of the game. Moreover, because successful prior performance feeds self-efficacy [2, 13], confidence gains in Group A students, who chose more ER, may be linked to electively replaying levels they have already mastered.

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 [17] 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 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-post- tests may be too easy for students, as 18.9% of pretests achieved a full score in accuracy, and 54.5% achieved 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 how many lives a student actually lost or 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.

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 [1]. This work presents prospects for both examining more detailed characteristics of replay and utilizing experimental manipulations.

6. ACKNOWLEDGEMENTS

This work was supported by NSF grant IUSE #1544273 "Evaluation for Actionable Change: A Data-Driven Approach" Teomara Rutherford PI, Tiffany Barnes & Collin F. Lynch Co-PIs.

7. REFERENCES

[1] V. Aleven, E. Stahl, S. Schworm, F. Fischer, and R. Wallace. Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73(3):277–320, 2003.

[2] A. Bandura. Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, 28:117–148.

[3] R. Bartle. Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD research*, 1(1):19, 1996.

[4] A. Boyce, K. Doran, A. Campbell, S. Pickford, D. Culler, and T. Barnes. Beadloom game: Adding competitive, user generated, and social features to increase motivation. In *the 6th International Conference on Foundations of Digital Games*, pages 139–146. ACM, 2011.

[5] C. Burgers, A. Eden, M. D. van Engelenburg, and S. Buningh. How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, 48:94–103, 2015.

[6] S. L. Calvert, B. L. Strong, and L. Gallagher. Control as an engagement feature for young children's attention to and learning of computer content. *American Behavioral Scientist*, 48(5):578–589, 2005.

[7] D. B. Clark, B. C. Nelson, H. Y. Chang, M. Martinez-Garza, K. Slack, and C. M. D'Angelo. 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):2178–2195, 2011.

[8] D. I. Cordova and M. R. Lepper. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4):715, 1996.

[9] J. P. Gee. *What video games have to teach us about learning and literacy.* St. Martin's Griffin - Macmillan, New York, USA, 2007.

[10] G. A. Gunter, R. F. Kenny, and E. H. Vick. Taking educational games seriously: Using the retain model to design endogenous fantasy into standalone educational games. *Educational Technology Research and Development*, 56(5-6):511–537, 2008.

[11] J. D. Marx and K. Cummings. Normalized change. *American Journal of Physics*, 75(1):87–91, 2007.

[12] J. Mostow, J. Beck, R. Chalasani, A. Cuneo, P. Jia, and K. Kadaru. A la recherche du temps perdu, or as time goes by: Where does the time go in a reading tutor that listens? In *In International Conference on Intelligent Tutoring Systems*, pages 320–329, 2002.

[13] F. Pajares. Self-efficacy beliefs in academic setting. *Review of Educational Research*, 66:543–578.

[14] J. L. Plass, P. A. O'keefe, B. D. Homer, J. Case, E. O. Hayward, M. Stein, and K. Perlin. The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology*, 105(4):1050, 2013.

[15] M. Prensky. Computer games and learning: Digital game-based learning. In *Handbook of computer games studies*. The MIT Press, Cambridge, MA, USA, 2005.

[16] T. Rutherford, G. Farkas, G. Duncan, M. Burchinal, M. Kibrick, J. Graham, L. Richland, N. Tran, S. Schneider, L. Duran, and M. Martinez. A randomized trial of an elementary school mathematics software intervention: spatial-temporal math. *Journal of Research on Educational Effectiveness*, 7(4):358–383, 2014.

[17] J. L. Sabourin, J. P. Rowe, B. W. Mott, and J. C. Lester. Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *JEDM-Journal of Educational Data Mining*, 5(1):9–38, 2013.

[18] S. Thomas, G. Schott, and M. Kambouri. Designing for learning or designing for fun? setting usability guidelines for mobile educational games. *Learning with mobile devices: A book of papers*, pages 173–181, 2004.