Grade 5 Students' Elective Replay After Experiencing Failures in Learning Fractions in an Educational Game: When Does Replay After Failures Benefit Learning?

Qian Zhang
qz89230@uga.edu
University of Georgia
USA

Teomara Rutherford
University of Delaware
USA
teomara@udel.edu

ABSTRACT

Despite theoretical benefits of replayability in educational games, empirical studies have found mixed evidence about the effects of replaying a previously passed game (i.e., elective replay) on students' learning. Particularly, we know little about behavioral features of students' elective replay process after experiencing failures (i.e., interruptive elective replay) and the relationships between these features and learning outcomes. In this study, we analyzed 5th graders' log data from an educational game, ST Math, when they studied fractions—one of the most important but challenging math topics. We systematically constructed interruptive elective replay features by following students' sequential behaviors after failing a game and investigated the relationships between these features and students' post-test performance, after taking into account pretest performance and in-game performance. Descriptive statistics of the features we constructed revealed individual differences in the elective replay process after failures in terms of when to start replaying, what to replay, and how to replay. Moreover, a Bayesian multi-model linear regression showed that interruptive elective replay after failures might be beneficial for students if they chose to replay previously passed games when failing at a higher, more difficult level in the current game and if they passed the replayed games.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; *Computer games.*

KEYWORDS

Educational Games, Replay, Learning Analytics

ACM Reference Format:
Qian Zhang and Teomara Rutherford. 2022. Grade 5 Students' Elective Replay After Experiencing Failures in Learning Fractions in an Educational Game: When Does Replay After Failures Benefit Learning?. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3506860.3506873

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
LAK22, March 21–25, 2022, Online, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506873

1 INTRODUCTION

One important feature of serious educational games that affect students' learning is whether students can replay a game. According to the framework for evaluating educational games by Gunter et al. [8], replaying a game can help students consolidate learned materials or remediate their shortcomings through more practice. Students may replay a game to pass the game or choose to engage in elective replay (ER) of a game that has been successfully passed before. Researchers are most interested in ER due to the complexity of such behaviors and their impact on students' learning [4, 12, 15, 20, 22]. However, little is known about the features of students' ER and the relationships between these features and students' learning outcomes.

This study analyzed 5th graders' log data in a supplemental game-based mathematics tutorial, Spatial Temporal (ST) Math, to construct systematic ER features to describe students' ER process, and applied a Bayesian multi-model linear regression to investigate which ER features are important to predict students' learning. Particularly, we situated our analysis in the context of fraction learning, because it is one of the most important but challenging math topics [13] and ER within fraction learning may have implications for students' exposure to and eventual learning of this important topic.

2 RESEARCH BACKGROUND AND RELATED WORK

2.1 Elective Replay and Interruptive Elective Replay

Researchers are particularly interested in ER because students have various motivations behind ER (e.g., ER for a better score, ER for managing negative emotions), need to make different decisions in ER (e.g., when to ER, what to ER), and their motivations and decisions can be influenced by many factors (e.g., learning environment, game designs) [4, 12, 15, 20, 22]. Despite a growing interest in students' ER in educational games, the effects of ER on students' learning are not conclusive. Theoretically, games that allow for ER can improve students' retention and understanding not only directly through more practice and feedback but also indirectly by affording a sense of autonomy or control, which increases students' motivation and engagement [6, 21]. Empirically, some studies found

that game designs that encouraged ER improved students' motivation and performance [3], but other studies found no relationship between ER and learning gains using game log data [5].

Particularly, a study by Liu and colleagues [12] examined students' ER behaviors in ST Math, a standards-aligned supplemental mathematics tutorial game (more details in the next section). They found that ER had no statistically significant associations with students' post-test performance, but students who had more ER after failing a game (termed *interruptive ER*) performed worse than students who had more ER after passing a game (termed *followed ER*) or who did not have ER. Also, they found that weaker students who had low pretest scores and did poorly in a game were more likely to have interruptive ERs. In contrast, students with high pretest scores and the in-game performance had more followed ERs. Thus, they concluded that low-performing students may not benefit from ER possibly due to their poor decisions in initiating more interruptive ERs. This study provided initial evidence about the associations between two different types of ER and students' learning outcomes, but their conclusion left some unanswered questions. For instance, the study's conclusion suggests a negative effect of interruptive ER. This conclusion was based on a comparison between groups of students whose majority of ER was interruptive or followed. However, the finding that students whose majority of ER is interruptive learned worse than students whose majority of ER is followed does not necessarily mean interruptive ER generally hinders learning. For example, when students fail and get stuck in a game, they could improve conceptual understanding and knowledge application automaticity required to pass the current game by replaying previously passed games that share similar content knowledge. Therefore, interruptive ER might be a good way for knowledge remediation and improvement if used strategically. More importantly, it is what we know about students' interruptive ER process in terms of ER features and how these interruptive ER features specifically improve or hinder learning that informs game design and instruction. Although Liu et al.'s [12] study constructed some ER features and examined relationships between these features and students' ER behaviors, their feature construction did not examine the relationships between ER features and students' learning outcomes. Therefore, research is needed to systematically investigate students' interruptive ER processes and the effects of interruptive ER features on students' learning.

2.2 ST Math and Learning of Fractions

ST Math, developed by MIND Research Institute, is a yearly mathematics tutorial game aligned to state standards (Figure 1). Each year's curriculum in ST Math is broken down into different objectives that cover different math topics through multiple games. Each objective starts with a pretest and ends with a post-test to assess students' prior knowledge before game play and their learning after game play. Each game within objectives contains multiple levels with increasing difficulty. Within levels, students solve multiple puzzles that represent the actual math content. Students cannot move on to the next puzzle until they correctly solve the current puzzle. However, students have limited opportunities to try solving puzzles: They typically have two Jiji "lives" within a level; an incorrect answer costs a Jiji life. If students use up all Jiji "lives" before correctly solving all puzzles in a level, they will not pass the level. This is counted as a level "failure" in the game. Thus, the number of level failures students have in a game is an indicator of their in-game performance. When students fail a level of a game, they can try to pass the level again by starting from the first puzzle of the level or they can choose to replay other previously passed levels (i.e., interruptive ER) in any game under any objective.

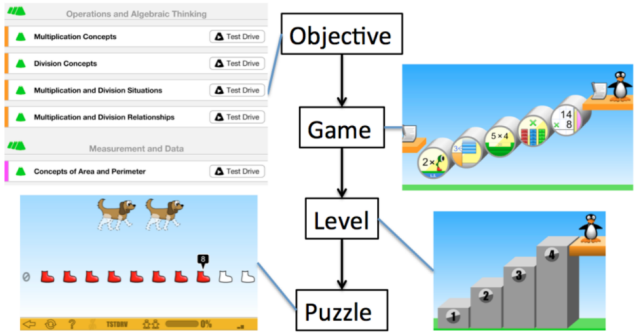

Figure 1: ST Math Content and Examples

The topic of fractions is well-known as one of the most important yet challenging math topics to learn [13]. Research has shown that accurate representation and translations between different representations of fraction magnitudes are crucial to conceptual understanding and arithmetic operations of fractions [23]. The same fraction magnitude can be represented in a symbolic or visualized form [14]. For example, ¾ and 0.75 are both symbolic representations of the same fraction magnitude. The two typical visual representations of fractions are the area model (i.e., fractions are represented as one or more equal areas of circles or grids) and the number line model (i.e., fractions are represented as numbers within a segment of the number line) [17, 25]. Moreover, these two widely-used visualizations are based on well-identified conceptualizations of fractions in literature that represent various mental models in conceptual understanding of fractions [1, 11]. Specifically, the area model is based on the part-whole mental model that conceptualizes fractions as one/more equal parts of a whole or one/more equal objects of a set, and the number line model is based on the measurement mental model that conceptualizes fractions as points on the number line. ST Math games in objective 11 "Fraction on the Number Line" were designed to help students construct the part-whole and measurement mental models of fractions by practicing translating between or within symbolic and visual representations of fractions (i.e., symbolic to visual, visual to symbolic, visual to visual, and symbolic to symbolic). The symbolic representations used in these games include fractions and decimals, and the visual representations include grids/squares, circles, and the number line. Although some studies have investigated how educational games can facilitate students' learning of fractions, these have mostly focused on game design features [7, 16, 25]. To our knowledge, no research has investigated the association between students' behavioral features (e.g., interruptive ER features) during a game and their learning of fractions.

Taken together, the current study addresses the need to systematically investigate behavioral features in students' ER process when playing an educational game and the relationship between these features and students' learning outcomes by answering two research questions: 1 What do interruptive ER features reveal about students' ER decisions of when to start an interruptive ER session, what to ER, and how to ER? 2 Which interruptive ER features predict students' post-test performance after taking into account their prior knowledge and number of failures during the game? We focus on interruptive ER due to its potentiality to improve learning and our lack of understanding of how students might use it and how it might hinder or help students' learning. We also only focus on students' interruptive ER behaviors in objective 11 "Fraction on the Number Line" in ST Math because of the importance and challenge of the math topic.

3 METHOD

3.1 Participants

The MIND Research Institute provided gameplay data from 5,521 5th graders who played ST Math during the 2018-2019 school year within the Beachside District [1] in California. We further identified students who had interruptive ER behaviors in objective 11 "Fractions on the Number Line". Among the 5,521 students, 2,808 played objective 11, and 358 of those students engaged in interruptive ER behaviors in objective 11. Table 1 compares students' demographic information among all 5th graders in the dataset from the Beachside District, students who played objective 11, and students who had interruptive ER behaviors in objective 11.

3.2 Measure

3.2.1 Students' Prior Knowledge, In-game Performance and Learning Outcomes. For each objective, students took a pretest before playing the objective and a post-test after completing the objective. Pretests and post-tests had five multiple-choice questions each and were parallel forms, question by question. Therefore, we used students' performance in the pretest before playing objective 11 as a measure of prior knowledge and we used students' performance in the post-test after playing objective 11 as a learning outcome. We also counted the number of students' failures during the game as a measure of their in-game performance.

3.2.2 Interruptive ER Features. We first generated five interruptive ER features by following students' sequential behaviors after they failed at a level in a game (Figure 2). First of all, students do not always ER after they fail a level; they may, instead, immediately replay the failed level or to quit the game. Therefore, we generated the first feature to describe the percentage of failures after which students choose to ER (%Failure). Also, levels in a game were designed to be increasingly difficult [18]. Students may ER after they fail at lower/easier levels or higher/harder ones. Therefore, we generated the second feature to describe students' failed levels before ER (Level before interruptive ER). It should be noted that the number of levels differs across games. For example, game 1 has four levels and game 2 has five levels. To make the level locations comparable across games, we normalized the level numbers to be between 0 and 1, so that, in all games, levels closer to 1 are harder than levels closer to 0. After students decide to ER after failing a level, they need to decide which objective, game, and level to replay, and they may replay a level within a game that is more or less similar to the game of the level they just failed. Therefore, we generated the third feature to measure the similarity between the games of students' replayed levels and failed levels immediately before ER (Game similarity). Once students start an interruptive ER, they still need to decide how many times to replay before they go back to make progress in the previously failed level. Therefore, we generated the fourth feature to measure the times of consecutive ER once students start an ER session (Consecutive ER times). Finally, because students have previously solved all the puzzles in an ERed level of a game (i.e., passed the level), students should be able to pass the level again, but there are incidences where they do not pass the replayed level. Therefore, we generated the fifth interruptive ER feature to measure the percentage of levels students passed when they engaged in ER (ER level passed%). Besides the five features, prior research on ER in ST Math demonstrated that students who have more interruptive ER have lower performance than students who have more followed ER or no ER [12], which implies that the percentage of interruptive ER out of all ER might matter. Therefore, we generated an additional feature of the percentage of interruptive ER (Interruptive ER%).

We constructed all features at the student granularity and some features at the game-level granularity (see Table 2). Each level of granularity is used to answer different research questions.

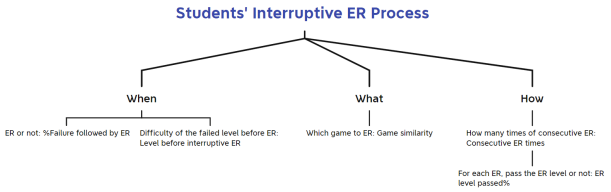

Figure 2: Students' Sequential ER Behaviors after Failures

3.2.3 Game Similarity Coding. We used a fine-grained measure of similarity at the student granularity and a general measure of similarity at the game-level granularity.

First, we coded each game in terms of representations, tasks, fraction conceptualizations, and knowledge required. The symbolic representations include fractions and decimals. The visual representations include grids/squares, circles, and the number line. The task in each game is translating between/within symbolic and visual representations (i.e., the game gives students one representation and asks them to generate another representation). The knowledge required in each game includes visual-symbol translation; symbol-symbol translation; visual-visual translation, fraction addition, fraction-decimal addition, decimal addition, and digital position. The fraction conceptualizations include the part-whole model and the measurement model, corresponding to different visual representations. Next, we compared each pair of games (failed games and ER games following failed games) and scored their similarities. For comparison between each game pair, we assign 1 if all their

[1] The district name is a pseudonym

Table 1: Comparison of Demographics Information

	Grade 5 students	Grade 5 students who played objective 11	Grade 5 students who have interruptive ERs in objective 11
N	5,521	2,882	358
Girls	46.44%	50.42%	45.81%
	Na: 5.72%	Na: 4.41%	Na: 5.31%
African American	10.90%	10.00%	17.60%
Asian	6.47%	7.36%	3.35%
Hispanic	54.86%	54.44%	63.13%
White	11.90%	13.19%	4.47%
Other races	10.14%	10.58%	6.15%
	Na: 5.72%	Na: 4.44%	Na: 5.31%
English language learner	10.72%	9.33%	24.86%
	Na: 5.72%	Na: 4.44%	Na: 5.31%
Special education student	10.72%	9.58%	22.91%
	Na: 5.72%	Na: 4.44%	Na: 5.31%

Table 2: Descriptions of Interruptive ER Features

ER features	Description of ER features at different granularities	
	Student granularity	Game-level granularity
Interruptive ER%	For each student, the percentage of interruptive ER sessions in all ER sessions.	For each game-level (e.g., game1-level1), the percentage of interruptive ER sessions in all ER sessions.
Failure%	For each student, the percentage of failures that followed by ER sessions.	For each game-level, the percentage of failures that followed by ER sessions.
Level before interruptive ER	For each student, the average normalized level (0-1) of failed games before all interruptive ER sessions.	Na
Game similarity	For each student, the average similarity between the failed games immediately prior to ER sessions and interruptive ER games within interruption ER sessions.	For each game level, the similarity between interruptive ER game and the current game.
Consecutive ER times	For each student, the average consecutive ER times over all ER sessions.	Na
ER level passed%	For each student, the average percentage of passed levels during interruptive ER.	Na

codes described above are the same, 0.5 if their codes are partially the same, 0 if their codes are totally different (see Table 3).

At the student granularity, we calculated average similarity scores between their failed games and ERed games for each student. At the game-level granularity, we compared ERed games and failed games for each game-level in objective 11 and qualitatively coded ERed games into four categories: the same game (i.e., game-pair score equals 7), a similar game in the same objective, a similar game in different objectives, irrelevant game (i.e., game-pair score equals 0).

4 RESULTS

RQ1: What do interruptive ER features reveal about students' ER process of when to start an interruptive ER session, what to ER, and how to ER? First, to describe students' ER process of when to start an interruptive ER session, we examined the percentage of failures that followed by ER (%Failure) at the student granularity and the game granularity. Table 4 shows that students chose to ER after 4.4% of their failures on average and 33.33% of their failures at most. This suggests that students did not chose to ER most of the time after they failed a level. Nevertheless, we found a statistically significant positive correlation between the number of failures and the number of interruptive ER sessions, r=0.43, p<.001 (at the student granularity), r = 0.54, p < .001 (at the game-level granularity), suggesting that the more failures students had, the more interruptive ER they engaged in. However, Figure 3 shows that there is a higher percentage of failures followed by an

Table 3: Fine-grained Measure of Similarity

Code		Game-pair score
Symbolic representations	1. Fractions 2. Decimals	
Visual representations	1. Grids/squares 2. Circles 3. The number line	1 – same
Tasks	1. Translate symbols to visuals 2. Translate visuals to symbols 3. Translate symbols to symbols 4. Translate visuals to visuals	
Fraction conceptualizations	1. The part-whole model 2. The measurement model	0.5 – partially same
Knowledge required	1. Visual-symbol translation 2. Symbol-symbol translation 3. Visual-visual translation 4. Fraction addition 5. Fraction-decimal addition 6. Decimal addition 7. Digital position	0 – different
Game	Game number	
Objective	Objective number	

interruptive ER session at some game-levels than others. For example, although students experienced a similar number of failures in game5-level4 and game7-level2, students chose to interruptive ER much more frequently in game7-level2 than in game5-level4. This suggests that different game designs might have influenced students' interruptive ER decisions. Moreover, the distribution of Level before interruptive ER revealed two modes on each side of the median (Table 4), suggesting that almost half of the students had ERs when they failed at lower levels of a game (0-0.58) whereas the other half had ERs when they failed at higher levels (0.58-1), M = 0.58, SD = 0.30. Besides, the average level of students' failures followed by ER was not associated with the average level of all their failures, (r = -0.09, p = 0.10).

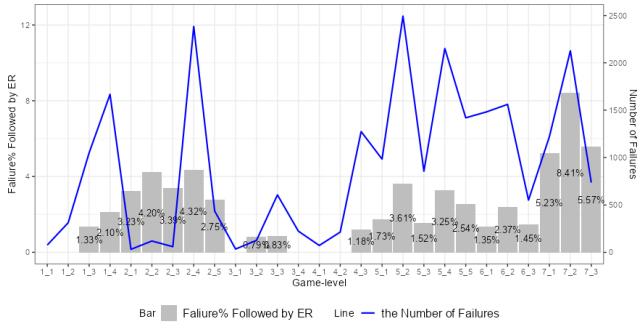

Figure 3: Faliure% Followed by ER and the Number of Failures by Game-level

Next, to describe students' decisions of what to ER, Figure 5 shows that most students chose to ER games within objective 11. However, game similarities differ across specific games. For example, students chose to replay irrelevant games most frequently when they failed in game5-level2, game5-level4, and game7-level2. Students chose to replay the same game most frequently in game1-level3 and game1-level4. These findings suggest that students mostly chose to replay a similar game within the same objective and that game design may have influenced their choices. To explore what exact games students chose to replay, we examined the flow of students' gameplay starting from the game they failed until the fifth consecutive ER and found that if students chose to replay a similar game within objective 11, they often chose to replay game1 no matter which game they failed (Figure 5). This suggests that many students' choices of what to ER may not be strategically based on game similarities but rather an intuition or a habit of "starting over."

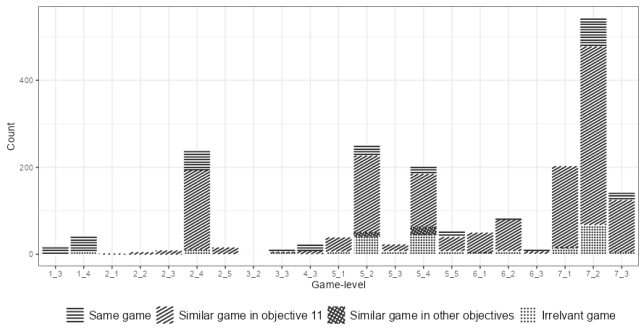

Figure 4: Similarity between Failed Games and ER Games by Game-level

Table 4: Descriptive Statistics for Interruptive ER Features at the Student Granularity (N = 358)

Interruptive ER features	Mean	SD	Median	Min	Max
Interruptive ER% (0%-100%)	96.67%	11.93%	100%	33.33%	100%
Failure% (0%-100%)	4.40%	3.96%	3.39%	0.30%	33.33%
Level before interruptive ER (0-1)	0.58	0.3	0.58	0	1
Game similarity (0-7)	4.34	2.07	4	0	7
Consecutive ER times	2.26	2.54	1.08	1	22
ER level passed% (0%-100%)	52.77%	40.24%	56.16%	0%	100%

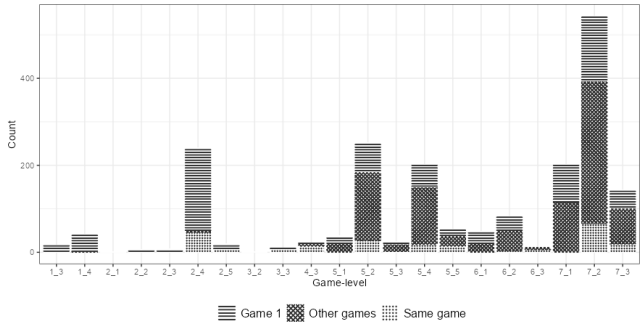

Figure 5: ER Game by Game-level

Finally, to describe students' decisions of how to ER, Table 4 shows that 50% of students only replayed an average of 1 game after a failure, M = 2.26, SD = 2,54 and on average, students passed around half of ERed levels, M = 52.77%, SD = 40.24%. More specifically, 79.89% of students had an average of fewer than three consecutive ERs per ER session/failure and only 30.45% of students passed all ERed levels.

RQ2: Which interruptive ER features predict students' post-test performance after taking into account their prior knowledge and number of failures during the game? After merging interruptive ER features data and students' pretest and post-test data, we ended up with a sample size of 312, because some students' pretest or post-test scores were missing. We calculated descriptive statistics about students' pretest scores, the number of failures they experienced in games and their post-test scores (Table 5). To answer the second research question, we did a Bayesian multi-model linear regression with JASP [10]. In the context of the current study, the Bayesian method has three major advantages over traditional linear regression analysis under the frequentist framework (for a full discussion of the benefits of Bayesian inference, see [24]). First, the Bayesian hypothesis testing can provide relative evidence of different hypothesized models based on the observed data via the Bayes factor. Specifically, the Bayes factor quantifies both evidence for the presence or absence of an effect and a model's relative predictive power. Second, the Bayesian parameter estimation could summarize the location and uncertainty of parameters based on the posterior distribution of all possible parameter values, taking into account prior information about these parameters. Particularly, the Bayesian method uses a credible interval to show the uncertainty about parameters which can tell the probability of the true parameter value lying within the credible interval. Third, the Bayesian

Table 5: Descriptive Statistics for Pretest, Failures in Game, and Post-test (N = 312)

Variables	Mean	SD	Median	Min	Max
Pretest (0-1)	0.36	0.24	0.4	0	1
Failures in game	69.45	62.41	56	3	670
Post-test (0-1)	0.59	0.25	0.6	0	1

multi-model analysis can estimate the parameter through model averaging that accounts for not only the uncertainty about the parameter in any one model but also uncertainty about the model per se. Such an unconditional parameter estimation method will not entirely rule out the possibility of any candidate models to avoid overconfident parameter estimates and biased inference [2].

We used a default prior option to assign prior distributions to the model parameters because we have little prior knowledge about the topic. The default prior option for model parameters in JASP applies a Cauchy distribution with spread r set to $\frac{1}{\sqrt{2}}$. Moreover, we used a uniform model prior to indicate before observing the data we believe all candidate models are equally likely given an absence of prior knowledge. For the sampling method to generate posterior distributions, we used Bayesian Adaptive Sampling (BAS) method. We entered students' pretest performance and their in-game performance (i.e., the number of failures) as nuisance predictors so that we can assess the contribution of interruptive ER features above and over the contribution of students' prior knowledge and in-game performance. Model comparison results (Table 6) showed that nine of the ten best models included Failure%, indicating the strong predictive power of this interruptive ER feature. The best model also selected level before interruptive ER and ER level passed% (BF_m = 10.29). The posterior summaries of coefficients (Table 7) confirmed the importance of the three ER features via posterior inclusion probabilities for Failure%, P (incl|data) = .81, level before interruptive ER, P (incl|data) = .67, and ER level passed%, P (incl|data) = .71. The Bayes factors provided moderate evidence for the importance of Failure%, $BF_{inclusion}$ = 4.17, but weak evidence for the importance of level before interruptive ER, $BF_{inclusion}$ = 2.04, and ER level passed%, $BF_{inclusion}$ = 2.43. The posterior summaries of coefficients (Table 7) also showed a positive influence of Failure% (Mean = 0.70, 95% Credible Interval [0,1.41]), level before interruptive ER (Mean = 0.06, 95% Credible Interval [0, 0.16]), and ER level passed% (Mean = 0.04, 95% Credible Interval [0, 0.12]) on students' post-test performance. These findings suggested that students who chose to ER more frequently after failures, chose to ER when failing at a higher, more difficult level in a game, and

had a higher percentage of passing ERed levels were more likely to perform better in the post-test.

5 DISCUSSION

RQ1: What do interruptive ER features reveal about students' ER process of when to start an interruptive ER session, what to ER, and how to ER? We constructed ER features by following students' sequential behaviors of ER process after they fail a level in a game. To investigate students' ER process, we analyzed descriptive statistics about ER features. Regarding when to ER, results showed that students in the current study who had interruptive ER behaviors in objective 11 did not often choose to ER after they failed, suggested by a low percentage of failures that followed by ER. Nevertheless, interruptive ER, by its definition, is a failure-driven behavior: Students who had more failures tended to have more interruptive ERs, and the more students failed at a game-level, the more likely they would have interruptive ERs. However, different game designs might also influence students' decisions whether to start an interruptive ER or not: For games that have a similar number of failures, there was a much higher percentage of failures followed by ER in some games than others. For instance, there were a similar number of failures in game1-level4 and game7-level2 but students ERed much more often in game7(Bubble Fraction Trap)-level2 than in game1(JiJi Cycle Basket)-level4 probably because game1-level4 used the simpler area model as visual representations whereas game7-level2 used the more difficult number line model as visual representations. Also, the design of game1-level4 included more attractive animations than that of game7-level2. Our findings that game designs influenced students' replay decisions are not only consistent with past research measuring students' intention of replay via subjective self-report [4, 20] but also provide evidence of objective observations via log data of students' realtime replay behaviors. Moreover, there are individual differences in the level of failed games that were followed by ER, with almost half of the students choosing ERs when they failed at lower levels of a game whereas the other half had ERs when they failed at higher levels. Regarding what to ER, results showed that most students chose to ER different games within the same objective (i.e., objective 11) but different game designs might have influenced their decisions to ER irrelevant games. Among students who chose to ER different games in objective 11, they mostly chose to replay game 1 no matter which game they previously failed, suggesting that students' choices of what to ER may not be strategic. Regarding how to ER, results showed that students typically had fewer than three consecutive ERs each time they started an ER session after failure, and, on average, students only passed half of the ER levels–this latter fact is of particular interest given that, by definition, all students had previously passed levels they chose to ER.

RQ2: Which interruptive ER features predict students' post-test performance after taking into account their prior knowledge and number of failures during the game? We did a Bayesian multi-model linear regression to investigate the second research question. Three ER features were selected in the best model that has the most predictive power: The percentage of failures followed by ER, the average level of the failed games before ER, and the percentage of passed ERs. Moreover, the Bayes factors provided moderate evidence for the importance of the percentage of failures followed by ER and weak evidence for the importance of the average level of the failed games before ER, and the percentage of passed ER levels. Parameter estimation indicated that the three features had positive influences on students' post-test performance. Specifically, students who chose to ER more frequently after failures, chose to ER when failing at a higher, more difficult level in a game, and had a higher percentage of passing ERed levels tended to perform better in the post-test. This model implies that interruptive ER after failures might be beneficial for students, and it would be even better if students chose to engage in interruptive ER when failing at a higher, more difficult level in a game. In addition, it might be important that they pass the ER level. It makes sense that if students keep failing at higher levels in a game, it might suggest a lack of sufficient knowledge or skills to pass the levels, and replaying a previously passed game (i.e., interruptive ER) might help students obtain the required knowledge and skills through more practice. However, if students fail at low levels of a game that are usually easy and basic, interruptive ER might not help much. Meanwhile, students might benefit from interruptive ER only if they pass the ER levels to have sufficient practice required to improve knowledge and skills.

It should be noted that the percentage of interruptive ER in all ER was not an important predictor in the current study. This finding did not necessarily contradict with past research that found a high percentage of interruptive ER negatively influenced students' learning [12], because most students who engaged in ER in the current study had a high percentage of interruptive ER (M = 96.67%, SD = 11.93%), which resulted in low variations in this predictor. Additionally, it seems counterintuitive at first glance that the similarity between failed games and ER games was not an important predictor. However, a closer examination of the data indicated that students who played irrelevant games in the current study happened to perform better in the game (i.e., fewer failures in the game). Besides, results from the first research question suggested that students may not be strategic in their choice of ER based on game similarities, because most students chose to start from game 1 after they failed. Taken together, our finding that game similarity was not an important predictor may be due to the quality of our data and the fact that students' choices of what to ER might not have been based on game similarity. One way to investigate the value of ER similarity to the failed game may be to assign students to ERs of varying similarities. In our study, although there is value in investigating actual student choice, we were constrained by the choices students made.

There is a wide literature about "replay" in research on educational games (e.g., [9, 19]). However, there is limited literature and research about "elective replay." Prior research has tested whether ER is associated with students' learning outcomes [12]. The current study dug deeper to ask which interruptive ER features are associated with students' learning outcomes. Our findings suggest that interrruptive ER after failing a game could positively influence learning. Such a positive influence also depends on students' decisions of when to ER and how to ER. However, 5th graders may not make these decisions strategically and need instructional support to benefit from interruptive ER. Evidence from our study could be used to design alert systems in educational games to guide students' interruptive ER behaviors. Moreover, ER features developed in the

Table 6: Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Failure% + Level before interruptive ER + ER level passed%	0.016	0.140	10.288	1.000	0.121
Failure% + interERpercent + Level before interruptive ER + ER level passed%	0.016	0.081	5.579	0.580	0.127
Failure% + Level before interruptive ER	0.016	0.079	5.399	0.562	0.109
Failure% + ER level passed%	0.016	0.076	5.204	0.544	0.109
Failure% + Consecutive ER times + Level before interruptive ER + ER level passed%	0.016	0.040	2.642	0.287	0.122
Failure% + Level before interruptive ER + Game similarity + ER level passed%	0.016	0.037	2.447	0.266	0.122
Failure% + interERpercent + Level before interruptive ER	0.016	0.034	2.228	0.243	0.113
Failure% + interERpercent + ER level passed%	0.016	0.033	2.165	0.237	0.113
Failure%	0.016	0.031	2.032	0.223	0.094
Level before interruptive ER + ER level passed%	0.016	0.029	1.914	0.210	0.103

Note. All models include Pretest, Failures in game. Table displays only the 10 best models.

Table 7: Posterior Summaries of Coefficients

								95% Credible Interval	
Coefficient	P(incl)	P(excl)	P(incl|data)	P(excl|data)	$BF_{inclusion}$	Mean	SD	Lower	Upper
Intercept	1.000	0.000	1.000	0.000	1.000	0.590	0.014	0.562	0.619
Pretest	1.000	0.000	1.000	1.110e-16	1.000	0.098	0.058	-0.020	0.220
Failures in game	1.000	0.000	1.000	1.110e-16	1.000	-6.514e-4	2.320e-4	-0.001	-1.392e-4
Failure%	0.500	0.500	0.806	0.194	4.168	0.701	0.475	0.000	1.524
interERpercent	0.500	0.500	0.354	0.646	0.548	-0.049	0.093	-0.297	0.044
Consecutive ER times	0.500	0.500	0.229	0.771	0.296	-6.451e-4	0.003	-0.009	0.003
Level before interruptive ER	0.500	0.500	0.671	0.329	2.041	0.062	0.057	-4.865e-4	0.171
Game similarity	0.500	0.500	0.217	0.783	0.278	-3.704e-4	0.003	-0.007	0.010
ER level passed%	0.500	0.500	0.708	0.292	2.429	0.050	0.043	-3.327e-4	0.129

current study could serve as a starting point for future research to investigate causal effects of elective replay by experimentally manipulating different ER features.

6 CONCLUSION

In this study we analyzed 5th graders' log data in a game-based mathematics tutorial, ST Math, to systematically describe features of students' interruptive ER process in learning fractions and to investigate the relationships between these features and students' learning outcomes. Regarding students' ER process, we found that students' decision to start an interruptive ER session may depend on the design of the particular level they failed. Some students were more likely to start an ER when they failed at lower levels of a game, whereas other students were more likely to start an ER when they failed at higher, more difficult levels of a game. Features of students' choices of what to ER showed that most students chose to replay different games in the same objective, and they most often chose game 1 no matter which game they failed. Once students started an ER session, they typically had fewer than three ERs in a row and seldom passed all levels in ER games. Regarding the relationships between interruptive ER features and students' learning outcomes, we found that, after considering their prior knowledge and the number of failures during the game, students who had a higher percentage of failures followed by ER, chose to ER when failing at a higher level in a game, and had a higher percentage of passing ER levels tended to perform better in the post-test. These findings imply that interruptive ER after failures might be beneficial for students under certain circumstances: namely after they have already persisted for a time on difficult content and when they pass the ERed level.

ACKNOWLEDGEMENTS: This work was supported by NSF grant #1845584 "CAREER: The Measurement and Influence of Mathematics Motivation in a Digital Context" and NSF grant #1544273 "Evaluation for Actionable Change: A Data-Driven Approach" Teomara Rutherford PI.

REFERENCES

[1] Merlyn J. Behr, Richard Lesh, Thomas Post, and Edward A. Silver. 1983. Rational number concepts. *Acquisition of mathematics concepts and processes* 91 (1983), 126.

[2] Don van den Bergh, Merlise A. Clyde, Akash R. Komarlu Narendra Gupta, Tim de Jong, Quentin F. Gronau, Maarten Marsman, Alexander Ly, and Eric-Jan Wagenmakers. 2021. A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods* (April 2021). https://doi.org/10.3758/s13428-021-01552-2

[3] Acey Boyce, Katelyn Doran, Antoine Campbell, Shaun Pickford, Dustin Culler, and Tiffany Barnes. 2011. BeadLoom Game: Adding competitive, user generated, and social features to increase motivation. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG '11)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/2159365.2159384

[4] Christian Burgers, Allison Eden, Mélisande D. van Engelenburg, and Sander Buningh. 2015. How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior* 48 (July 2015), 94–103. https://doi.org/10.1016/j.chb.2015.01.038

[5] Douglas B. Clark, Brian C. Nelson, Hsin-Yi Chang, Mario Martinez-Garza, Kent Slack, and Cynthia M. D'Angelo. 2011. 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 (Nov. 2011), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

[6] Diana I. Cordova and Mark R. Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology* 88, 4 (Dec. 1996), 715–730. https://doi.org/10.1037/0022-0663.88.4.715 Publisher: American Psychological Association.

[7] Melissa Sommerfeld Gresalfi, Bethany Rittle-Johnson, Abbey Loehr, and Isaac Nichols. 2018. Design matters: Explorations of content and design in fraction games. *Educational Technology Research and Development* 66, 3 (June 2018), 579–596. https://doi.org/10.1007/s11423-017-9557-7

[8] Glenda A. Gunter, Robert F. Kenny, and Erik H. Vick. 2008. Taking educational games seriously: Using the RETAIN model to design endogenous fantasy into standalone educational games. *Educational Technology Research and Development* 56, 5 (Dec. 2008), 511–537. https://doi.org/10.1007/s11423-007-9073-2

[9] Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, and Brad A. Myers. 2015. Replay analysis in open-ended educational games. In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler (Eds.). Springer International Publishing, Cham, 381–399. https://doi.org/10.1007/978-3-319-05834-4_17

[10] JASP Team. 2021. JASP (Version 0.16)[Computer software]. https://jasp-stats.org/

[11] Thomas E. Kieren. 1976. On the mathematical, cognitive and instructional. In *Number and measurement. Papers from a research workshop*, Vol. 7418491. Citeseer, 101.

[12] Zhongxiu Liu, Christa Cody, Tiffany Barnes, Collin Lynch, and Teomara Rutherford. 2017. The antecedents of and associations with elective replay in an educational game: Is replay worth it? International Educational Data Mining Society. https://eric.ed.gov/?id=ED596614 Publication Title: International Educational Data Mining Society.

[13] Hugues Lortie-Forgues, Jing Tian, and Robert S. Siegler. 2015. Why is learning fraction and decimal arithmetic so difficult? *Developmental Review* 38 (Dec. 2015), 201–221. https://doi.org/10.1016/j.dr.2015.07.008

[14] Ofer Marmur, Xiaoheng Yan, and Rina Zazkis. 2020. Fraction images: The case of six and a half. *Research in Mathematics Education* 22, 1 (Jan. 2020), 22–47. https://doi.org/10.1080/14794802.2019.1627239 Publisher: Routledge _eprint: https://doi.org/10.1080/14794802.2019.1627239.

[15] Jack Mostow, Greg Aist, Joseph Beck, Raghuvee Chalasani, Andrew Cuneo, Peng Jia, and Krishna Kadaru. 2002. A la recherche du temps perdu, or as time goes by: Where does the time go in a reading tutor that listens?. In *Intelligent Tutoring Systems (Lecture Notes in Computer Science)*, Stefano A. Cerri, Guy Gouardères, and Fàbio Paraguaçu (Eds.). Springer, Berlin, Heidelberg, 320–329. https://doi.org/10.1007/3-540-47987-2_36

[16] Manuel Ninaus, Kristian Kiili, Jake McMullen, and Korbinian Moeller. 2017. Assessing fraction knowledge by a digital game. *Computers in Human Behavior* 70 (May 2017), 197–206. https://doi.org/10.1016/j.chb.2017.01.004

[17] Marilena Pantziara and George Philippou. 2012. Levels of students' "conception" of fractions. *Educational Studies in Mathematics* 79, 1 (Jan. 2012), 61–83. https://doi.org/10.1007/s10649-011-9338-x

[18] Zhongxiu Peddycord-Liu, Rachel Harred, Sarah Karamarkovich, Tiffany Barnes, Collin Lynch, and Teomara Rutherford. 2018. Learning curve analysis in a large-scale, drill-and-practice serious math game: Where is learning support needed?. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 436–449. https://doi.org/10.1007/978-3-319-93843-1_32

[19] Dorothy Ann Phoenix. 2014. How to add replay value to your educational game. *Journal of Applied Learning Technology* 4, 1 (2014), 20–23. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=ehh&AN=97479993&site=ehost-live&scope=site&custid=uga1 Publisher: Society for Applied Learning Technology.

[20] Jan L. Plass, Paul A. O'Keefe, Bruce D. Homer, Jennifer Case, Elizabeth O. Hayward, Murphy Stein, and Ken Perlin. 2013. The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology* 105, 4 (Nov. 2013), 1050–1066. https://doi.org/10.1037/a0032688 Publisher: American Psychological Association.

[21] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 54–67. https://doi.org/10.1006/ceps.1999.1020

[22] Jennifer L. Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. 2013. Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *Journal of Educational Data Mining* 5, 1 (2013), 9–38. https://eric.ed.gov/?id=EJ1115354 Publisher: International Educational Data Mining.

[23] Robert S. Siegler, Lisa K. Fazio, Drew H. Bailey, and Xinlin Zhou. 2013. Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Sciences* 17, 1 (Jan. 2013), 13–19. https://doi.org/10.1016/j.tics.2012.11.004

[24] Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F. Gronau, Martin Šmíra, Sacha Epskamp, Dora Matzke, Jeffrey N. Rouder, and Richard D. Morey. 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review* 25, 1 (Feb. 2018), 35–57. https://doi.org/10.3758/s13423-017-1343-3

[25] Lu Zhang, Junjie Shang, Tim Pelton, and Leslee Francis Pelton. 2020. Supporting primary students' learning of fraction conceptual knowledge through digital games. *Journal of Computer Assisted Learning* 36, 4 (2020), 540–548. https://doi.org/10.1111/jcal.12422 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12422.