Gamification and Deadending: Unpacking Performance Impacts in Algebraic Learning.

Siddhartha Pradhan
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
sppradhan@wpi.edu

Ashish Gurung
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
agurung@andrew.cmu.edu

Erin Ottmar
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
erottmar@wpi.edu

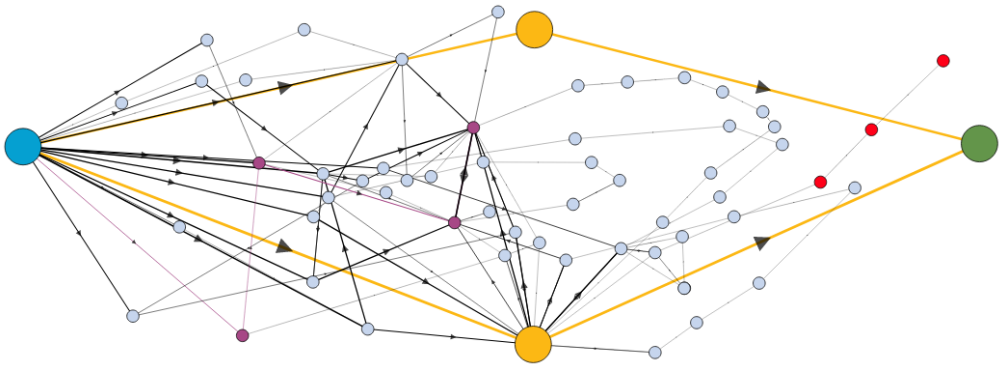


Figure 1: Classifications for all student attempts for Problem 65 in FH2T

ABSTRACT

This study explores the effects of varying problem-solving strategies on students' future performance within the gamified algebraic learning platform From Here To There! (FH2T). The study focuses on the procedural pathways students adopted, transitioning from a start state to a goal state in solving algebraic problems. By dissecting the nature of these pathways—optimal, sub-optimal, incomplete, and dead-end—we sought correlations with post-test outcomes. A striking observation was that students who frequently engaged in what we term 'regular dead-ending behavior', were significantly correlated with higher post-test performance. This finding underscores the potential of exploratory learner behavior within a low-stakes gamified framework in bolstering algebraic comprehension. The implications of our findings are twofold: they accentuate the significance of tailoring gamified platforms to student behaviors and highlight the potential benefits of fostering an environment that promotes exploration without retribution. Moreover, our insights hint at the notion that fostering exploratory behavior could be instrumental in cultivating mathematical flexibility.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636929

KEYWORDS

procedural pathways, algebraic learning, gamification, math flexibility, networks

ACM Reference Format:
Siddhartha Pradhan, Ashish Gurung, and Erin Ottmar. 2024. Gamification and Deadending: Unpacking Performance Impacts in Algebraic Learning.. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3636555.3636929

1 INTRODUCTION

Solving algebraic problems require students to utilize a broad spectrum of problem-solving techniques. These techniques enhance students' ability to synthesize solutions, shape their mathematical intuition, and reinforce their methodological approaches to problem-solving. As students progress to more advanced mathematical domains, mastering these foundational strategies becomes paramount. Indeed, proficiency in algebraic concepts is intimately linked with the acquisition of a wide array of problem-solving techniques [24]. Especially in K-12 mathematics education, efficiency and flexibility in problem-solving strategies are prioritized [15], and efficient students often employ fewer steps or transformations [32]. This is supported by various studies that suggest that strategic efficiency is a significant indicator of a student's understanding of the inherent mathematical structures [26, 30]. Yet, despite the widespread acknowledgment of the correlation between strong algebraic knowledge and enhanced performance in future advanced topics, a disconcerting number of middle school and high school students struggle with fundamental algebraic concepts. Difficulties making such as valid transformations and decomposition [24], and challenges in converting simple story problems into mathematically

equivalent equations [17], are indicative of the potential struggles these students might face as they encounter more advanced topics typically expressed in algebraic form.

Over the years, both researchers and developers have created a diverse set of tools and systems to bolster algebraic learning, especially in facilitating the acquisition of problem-solving strategies. Notably, rule-based approaches have stood out in the development of Intelligent Tutoring Systems (ITS), with the primary goal of strengthening the acquisition of procedural knowledge that's pivotal to algebraic comprehension. These approaches find their foundation largely in the cognitive theories presented in ACT-R [2]. A variety of cognitive tutors have emerged over time, all aiming to aid learners in achieving mastery across various subjects [3, 8, 25]. While these cognitive tutors are designed to provide learners with adaptive feedback and personalized guidance [1], the procedural pathways available to students are often constrained by what the problem creator considers essential for mastering the core concepts. In contrast, some ITS have adopted an alternative approach, developing systems that merely require students to enter their answers to the problems, without demonstrating the procedural comprehension necessary to solve them [14]. Despite the ambiguity surrounding the procedural pathways chosen by the learner, the use of such systems has led to better learning outcomes [21], and the availability of feedback has been observed to be beneficial in enhancing learning [16, 20]. Though both approaches effectively facilitate the acquisition of mathematical knowledge [21, 23, 28], little is known regarding the various procedural pathways learners might potentially employ in formulating a solution.

With the rapid development in technology and ensuing innovations in Intelligent Tutoring Systems (ITS), researchers and developers have been investigating the efficacy of implementing novel methodologies to aid learners in acquiring algebraic knowledge. A subset of these educational technology, such as Graspable Math [31] and 'From Here To There!' (FH2T)[6], have embraced dynamic procedural pathways for teaching algebra. Specifically, FH2T adopts a distinctive dynamic procedural approach: learners are presented with an algebraic expression as the starting state and a transformed version of that expression as the goal state. Students can traverse any procedural pathway they prefer, with all mathematically valid transformations being permissible. This architecture inherently grants learners the autonomy to explore various procedural avenues in their exploration of the transformations necessary to attain the goal state. Such a modality can shed light on the diversity of procedural pathways chosen by learners.

FH2T utilizes a gamification model to enhance student participation. Various prior studies have reported on their exploration of the efficacy of FH2T in improving students' algebraic knowledge [6, 9] and the different aspects of the in-game behavior that can predict better learning outcomes [7, 29]. However, to the best of our knowledge, very little exploration regarding the variation in the procedural pathways adopted by students in their attempt to reach the goal state has been studied. As illustrated in Figure 1, the transformations executed by students can be harnessed to construct a network representing their approach, from the starting state to the goal state. This network can reveal procedural pathways that are optimal, sub-optimal, incomplete, and on occasion, paths that culminate in dead-ends. An incomplete path arises when students stop working while other students have pursued this path to the goal state. Distinctly, a dead-end path represents a trajectory chosen by one or more students who ceased progress before reaching the goal state. These paths stand out from incomplete approaches because no student has ever traversed them successfully from start to goal state. Hence, it's unclear if these are genuine dead-ends or trajectories that future attempts might lead to successful completion. The underlying mechanisms that prompt students to discontinue their current procedural approach, leading to dead-end and sub-optimal pathways, remain unclear. However, various factors, both positive and negative, can sway a learner's decision to discontinue. Positive triggers might include realizing that a path will only yield a sub-optimal outcome or foreseeing a challenging state ahead. Conversely, negative factors could include frustration from an inability to solve a problem or reaching a genuine impasse where the student is unable to identify the next state.

As such, this paper aims to explore the implications of encountering dead-ends within the network of strategic pathways generated using procedural approaches adopted by the students while working on algebraic problems. Accordingly, we explore the following research questions:

RQ 1 Does the choice of procedural pathways in algebraic problem-solving lead to differentiated learning outcomes?

RQ 2 In what ways do dead-end attempts within a gamified environment impact algebraic learning?

2 METHODOLOGY

2.1 Data

The data used in this study was collected as part of a large Randomized Control Trial (RCT) conducted from September 2020 to April 2021. The study (c.f. [9]), explored the impact of three different educational technology tools on students' algebraic learning. A total of 4092 7th-grade students were recruited from 11 middle schools within a large suburban district in the United States. The students were randomly assigned to one of four conditions: 2 gamified conditions (FH2T and DragonBox [27]), ASSISTments [14] instant feedback problem sets, and an active control delayed feedback condition. This data from this study is publicly available for researchers through OSF[1] (c.f. [22]). The dataset includes information regarding assessments, demographics, logged student actions, and aggregated data for each condition. Assessment data includes pre-test and post-test scores that measure students' algebraic knowledge using ten items adapted from a previously validated measure (ranging from 0 to 10) [22]. This paper only utilized the data associated with students assigned to the FH2T condition, as our objective is to explore the dynamic procedural pathways students took to reach the goal state. In cases where students attempt mathematically impossible transformations, these invalid attempts are recorded as errors, but the transformations are not executed. This approach aims to enhance students' algebraic understanding through practice by allowing them to identify the infeasibility of certain transformations.

2.1.1 Sample. 1,649 students were assigned to the FH2T condition, of which 52.6% were male and 47.4% were female. 49.8% of

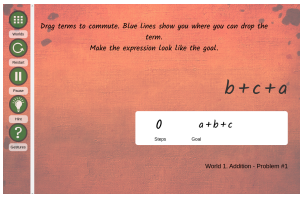
[1] https://osf.io/r3nf2/

the students identified as White, 24.8% as Asian, and 16.4% as Hispanic/Latino. The remaining 9% identified with other racial categories or reported multiple racial affiliations. It's important to note that students who did not complete both the pre- and post-tests were omitted from our analysis, resulting in a final sample of 774 students.

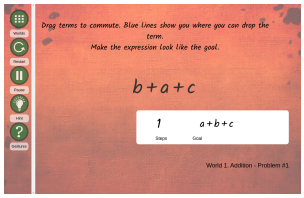
2.2 Classifying student attempts

We used the students' action level interaction log data to generate the dynamic procedural pathways for each student's attempt to reach the goal state. Once created, these pathways formed a network comprising optimal, sub-optimal, incomplete, and dead-end pathways. The subsequent section details the procedure we employed to capture the various attempts made by students, which cumulatively resulted in the formation of this network across multiple attempts by various students.

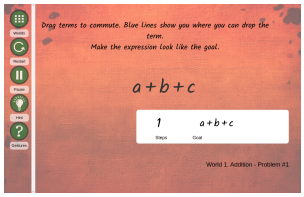
2.2.1 Solution steps as a directed graph. The sequence of actions taken by a student to reach the goal state can be visualized as a series of nodes in a directed graph or network. In algebraic problem-solving, each transformation acts as an edge between two nodes, where each node signifies a mathematical expression either before or after a transformation. For instance, in the sequence of transformations depicted in Figure2, expressions like "b+c+a" (start state, Figure2a), "b+a+c" (first step, Figure2b), and "a+b+c" (goal state, Figure2c) each constitute a node in the directed graph. The transformations ["b+c+a" → "b+a+c"] and ["b+a+c" → "a+b+c"] illustrate the edges connecting the respective nodes.



(a) Start State



(b) First Step



(c) Goal State

Figure 2: An example of an inefficient student attempt in FH2T

In the context of Figure2, identifying the most efficient strategy is straightforward since there exists a singular optimal path from the start state to the goal state (i.e. ["b+c+a" → "a+b+c"]). When students complete the task as presented in Figure2, any attempt that encompasses a singular step can be designated as efficient; any deviation from this is deemed sub-optimal. However, it's worth noting that certain problems may present a suite of equally efficient strategies. To systematically uncover all efficient paths or transformations for a given problem, we begin by populating a graph with all pertinent transformations, sourced from the log data. Subsequently, through the application of efficient graph traversal methodologies, such as A* or Dijkstra's shortest path algorithm, we can pinpoint all viable efficient solutions within a specific student sample or cohort. For this study, we favored Dijkstra's algorithm due to its inherent capability of attributing weights to particular edges – in this case, transformations. This flexibility grants us the latitude to either penalize or amplify the significance of specific transformations. Nonetheless, in this preliminary study, we have assigned a uniform weight to all transformations.

Figure 3 is an example of identified multiple equally efficient strategies or solution steps. The blue node is the start state (i.e. 4*(2+3)*(-100+1+45+55)/(2+3)*3), and the green node is the goal state (i.e. 6+6). The golden nodes represent nodes that are in the best path (i.e. are steps of an efficient solution). In this case, there are 3 distinct efficient strategies, that all require 7 steps or transformations. Additionally, the thickness of each edge and arrow represents the number of students who made that transformation. We can also deduce that the path at the bottom of the figure is more common in comparison to other paths.

2.2.2 Identified classifications. As described above, by utilizing Dijkstra's shortest path algorithm, we identified all the efficient steps for any given problem in FH2T. Students' completed attempts were classified as optimal or sub-optimal paths by referencing the observed best path between the start and the goal node in the network. Additionally, we also noticed two distinct types of incomplete procedural paths in the network: 'incomplete path' and 'dead-end path'. Incomplete paths are incomplete attempts that are part of the observed optimal or sub-optimal paths indicating that the same student or other students took the same path to reach the goal state in a different attempt. Dead-ends, on the other hand, are paths that have never led to a goal state across attempts, i.e., no student has successfully taken the path to reach the goal state. As such, dead-ends are a unique type of incomplete path within the generated networks.

An example classification for problem attempts can be seen in Figure 4. This figure is similar to Figure 3, however, it contains all the student attempts from the log data for that particular problem rather than just the best paths. Examples of specific classifications have been given in Figure5, which are isolated attempts from Figure 4. The grey nodes represent sub-optimal steps (see Figure 5b), any attempt that contains these nodes is inefficient. The red nodes represent dead-end nodes (see Figure 5c), there are no paths or edges that lead to the goal state. Any attempt containing one of the red nodes is classified as a dead-end attempt. The incomplete attempts are difficult to distinguish visually, as by definition, those paths have been taken by other students to reach the goal state. To visually represent a sample incomplete attempt we highlighted the associated nodes and edges in purple (see Figure 5d). In the highlighted attempt, the student initially used a sub-optimal strategy but did not reach the goal state, i.e., decided to reset.

2.2.3 Classifying attempts for all problems. In order to compare what paths led to better learning outcomes, we classified all attempts across all problems into their respective categories. The

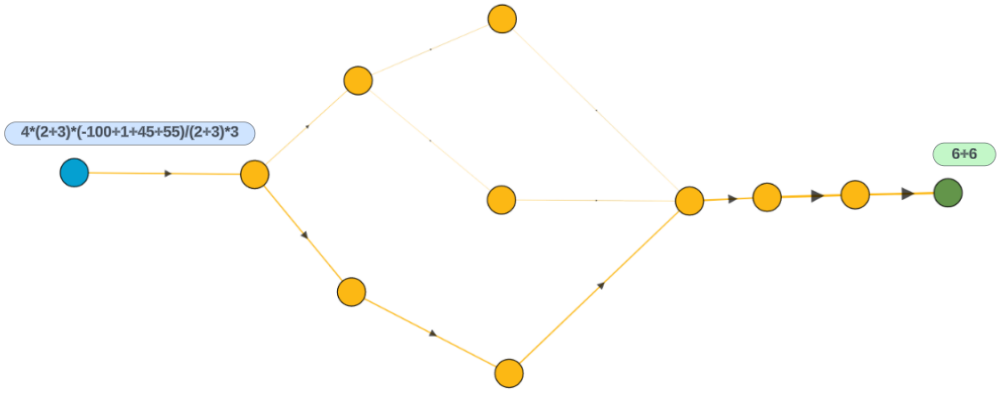


Figure 3: The identified best paths in a graph for problem 252 in FH2T

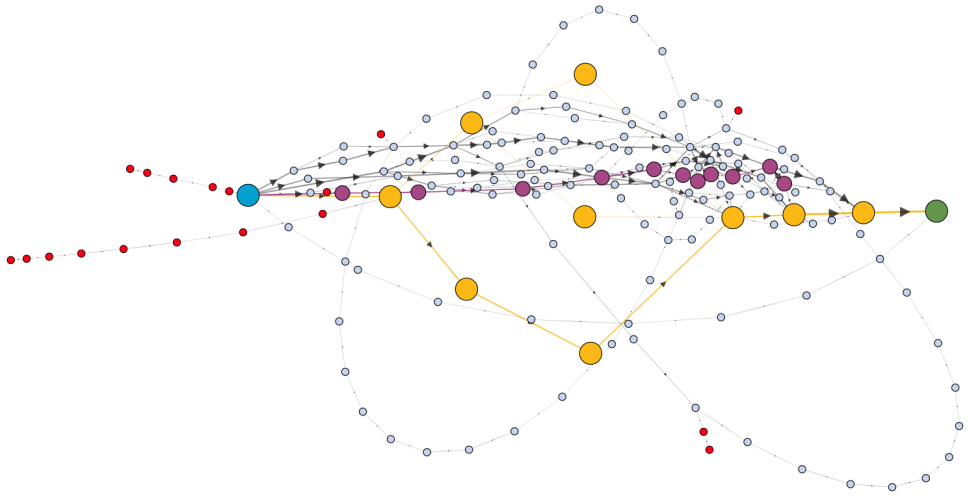


Figure 4: All attempt classifications for Problem 252 in FH2T

classification was performed using pandas [19], and NetworkX in Python 3.11.

3 RESULTS

3.1 RQ 1: Exploring what types of attempts lead to better learning outcomes

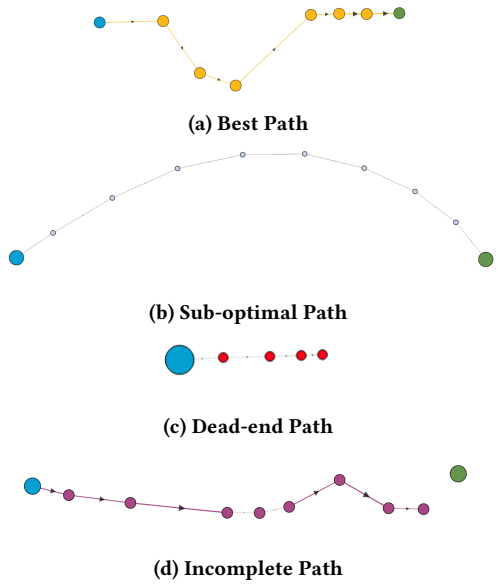
To address **RQ1** and identify problem-solving strategies that led to better learning outcomes, we estimate two linear models. Model 1 predicts the post-test scores of students based on the identified pathways or classifications, while the second model accounts for prior algebraic knowledge (mean-centered) in addition to the classifications. Table 1 contains the results of running the linear models. The results of model one suggest that at the student level, neither classification of best (β = -0.38, p = 0.407) nor sub-optimal (β = 1.46, p = 0.118) was a significant predictor of post-test scores. Surprisingly, the classification of incomplete (β = -1.69, $\boldsymbol{p}$ **<0.001**) and classification of dead-end (β = 5.17, $\boldsymbol{p}$ **<0.001**) were significant predictors of post-test scores. These results were surprisingly counter-intuitive, as we originally hypothesized that dead-ends indicate poor procedural knowledge and would consequently lead to lower post-test scores.

In model two, we observed that higher prior knowledge was correlated with higher post-test performance (β = 0.72, $\boldsymbol{p}$ **<0.001**). The best path (β = -0.76, $\boldsymbol{p}$ = **0.028**) was also a significant predictor of post-test scores. Additionally, while the effect decreased, dead-end (β = 2.28, $\boldsymbol{p}$ = **0.004**) was still a significant predictor of higher post-test performance.

Overall, the results of these models suggest that after accounting for prior algebraic knowledge, the average student exhibiting dead-ending behavior is more likely to succeed. On the other hand, students who exhibit efficient problem-solving behavior, tend to perform worse on the post-test. The positive correlation between dead-ending behavior and student post-test performance indicates the likelihood that the underlying mechanism that results in dead-ending behavior is likely positive in nature. Such learners are able to

Table 1: Student level linear regression results predicting post-test score.

	Model 1 post test math score			Model 2 post test math score		
Predictors	*Estimates*	*CI (α=0.05)*	*p*	*Estimates*	*CI (α=0.05)*	*p*
(Intercept)	4.79	[4.02, 5.56]	**<0.001**	4.70	[4.12, 5.29]	**<0.001**
Class Best	-0.38	[-1.27, 0.52]	0.407	-0.76	[-1.44, -0.08]	**0.028**
Class Incomplete	-1.69	[-2.27, -1.11]	**<0.001**	-0.35	[-0.80, 0.10]	0.129
Class Sub-optimal	1.46	[-0.37, 3.29]	0.118	1.08	[-0.31, 2.46]	0.129
Class Dead-end	5.17	[3.17, 7.16]	**<0.001**	2.28	[0.75, 3.81]	**0.004**
Pre Total Math Score				0.72	[0.66, 0.77]	**<0.001**
Observations	774			774		
R^2/R^2 *adjusted*	0.057/0.052			0.459/0.456		



(a) Best Path

(b) Sub-optimal Path

(c) Dead-end Path

(d) Incomplete Path

Figure 5: Examples of the 4 different attempt types in Problem 252

identify that the path will only yield a sub-optimal path or foresee a challenging state ahead. As such, we posit that a dead-end attempt may, in fact, be an indicator of 'exploratory play', an in-game behavior that potentially leads to a more nuanced understanding of the transformations to avoid or the ability to identify problematic states when solving algebraic problems to reach the total state. Consequently, resulting in a better post-test performance.

3.2 RQ 2: Exploring the effect of regular dead-ending on algebraic learning outcomes

Building on the surprising results of **RQ1**, we further explored the relationship between dead-ending (or exploratory behavior) and higher post-test scores. We examined potential variance in the dead-end states across students by constructing individual networks per student per problem. Such student-level networks were generated to identify dead-end paths of students that were potentially masked by their peers' attempts. For example, if a student had an exploratory attempt ('start state' → 'a'), and another student used the same path to reach the goal state sub-optimally ('start state' → 'a' → 'b' → 'goal state'), the student's exploratory attempt would be masked and classified as incomplete. By identifying dead-end paths on student-level networks, we localize the definition of dead-end paths to individual students' attempts. It is important to note that this modification does not change the classification for optimal or sub-optimal attempts, as the best paths found from the entire sample are used for this classification.

Next, we examined the frequency of dead-ending behavior per student by examining the total number of problems in which the student had at least one dead-end attempt. Similarly, we calculate the percentage of problems with at least one dead-end attempt. These results can be found in Table 2. Since the dead-end count and percentages were not normally distributed, and certain students were regularly utilizing the dead-end pathways in comparison to their peers, we classified the students into *'regular dead-enders'* and *'occasional dead-enders'* by utilizing a cutoff point at the 5th percentile of the dead-ending behavior distribution. We ran a mixed-effects model at the attempt level, predicting post-test scores while accounting for prior knowledge (mean-centered), using the student-level network classification and an indicator for the students' regular usage of dead-end paths. As the data is at the attempt level, we introduce random intercepts for the problem ID, attempt number, and pre-test scores.

Table 3 suggests that for a student with an average score on the pretest, the use of optimal or best paths correlates significantly with higher scores on the post-test (β = 0.33, $p <$ **0.001**), especially when compared to the reference category of incomplete paths. This trend is also seen with sub-optimal paths (β = 0.07, $p <$ **0.001**) and dead-end paths (β = 0.05, $p =$ **0.006**), both showing a positive correlation with the students' post-test scores. Similar to the results of **RQ1**, the pre-test score remains a significant predictor of the post-test scores (β = 0.65, $p <$ **0.001**). Interestingly, students who regularly adopt dead-ending strategies in their problem-solving tend to perform better than those who use such strategies less frequently (β = 0.24, $p <$ **0.001**).

Table 2: Summary Statistics of Dead-end Count and Percentage

Statistic	Mean	St. Dev.	5%	Median	95%
Dead-end Count	27.1	16.1	4	24.5	56
Dead-end Percentage	23.4	7.5	12.05	23.2	35.5

Table 3: Exploring the correlation between different types of procedural pathways taken by individual students and their post-test performance.

	post test math score		
Predictors	*Estimates*	*CI (α=0.05)*	*p*
(Intercept)	4.90	[4.50, 5.29]	**<0.001**
Attempt Best	0.34	[0.31, 0.38]	**<0.001**
Attempt Deadend	0.05	[0.01, 0.08]	**0.006**
Attempt Sub-optimal	0.07	[0.04, 0.11]	**<0.001**
Pre Total Math Score	0.65	[0.53, 0.76]	**<0.001**
Regular Deadending	0.24	[0.18, 0.31]	**<0.001**
Random Effects			
σ^2	4.07		
$\tau_{00 problem\ id}$	0.54		
$\tau_{00 attempt\ number}$	0.08		
$\tau_{00 pre\ total\ math\ score}$	0.38		
ICC	0.20		
$N_{attempt\ number}$	94		
$N_{pre\ total\ math\ score}$	11		
$N_{problem\ id}$	252		
Observations	179575		
Marginal R^2*/Conditional* R^2	0.391/0.512		

4 DISCUSSION

In this study, we find that students who exhibit regular dead-ending behavior have a higher post-test score (i.e. higher learning outcome), than students who are irregular dead-enders. In other words, students exhibiting regular dead-ending behavior (i.e. exploratory), gain more from the gamified system. This suggests that students who display regular exploratory (i.e. dead-ending) behavior may be learning the various algebraic rules and notations in a low-stakes gamified environment, eventually leading to better algebraic understanding.

The findings of this study have two major implications. Firstly, the positive effect of gamified systems on algebraic learning outcomes depends on the behaviors exhibited by the student. Past studies such as [7, 18, 29], have shown that different in-game behaviors are predictive of algebraic learning outcomes. In particular, studies [7, 18], showed that students who paused before answering tend to perform better in the post-test. Similarly, [29] showed that students with a higher propensity for persistence benefit more from the gamified system. In the current study, using log data, we identified an exploratory behavior that results in better learning outcomes. We provided additional evidence suggesting that the effect of gamified platforms on learning outcomes depends on the behaviors and intentions of the user.

The second major implication is that in-game behaviors exhibited by students may be the main driving force behind improved algebraic knowledge in gamified systems. Desirable behaviors, such as the exploratory behavior identified in this study, should not be penalized. If the results presented in this study are consistent for similar gamified systems, there are profound impacts on the design of gamified platforms to foster exploration. Additionally, our results suggest that in order to develop math flexibility, students may need to explore various procedural pathways. In the long run, this may allow students to develop the important skill of choosing efficient problem-solving strategies.

5 LIMITATIONS AND FUTURE WORK

In considering the outcomes of this study, several important caveats should be acknowledged. To begin with, our analysis was narrowly focused on data derived from the FH2T platform. This specificity introduces potential limitations on the generalizability of the results. There remains an open question about the replicability of the observed student behaviors and interactions across a wider range of platforms that employ similar dynamic procedural pathways. To strengthen the findings of this study, it would be instructive for subsequent investigations to explore the generalizability of our findings further. Additionally, the insights extrapolated here

might be more germane to gamified environments rather than to traditional tutoring platforms such as the Cognitive Tutor [3] and ASSISTments [14] mentioned earlier. These platforms, with varying affordances regarding procedural requirements, might influence student behavior differently, possibly reducing the propensity for the kind of exploratory action observed in our analysis.

While this study aimed to identify and understand the implications of various procedural pathways in solving algebraic problems within a gamified setting, the broader implications of these classifications must be acknowledged. Future research should investigate the effects of hints on the paths and explore variations in their effective utilization. Prior research has underscored the value of using the response times as a metric to infer productive hint usage [13] and the formulation of optimal solutions [7]. Additionally, several studies have highlighted the benefits of providing error-specific feedback to frequently occurring incorrect answers [10, 11]. The models established in this research can greatly enhance our understanding of the mechanisms underlying the procedural pathways that lead to these common errors. Similarly, insights into these pathways can improve the quality of automated grading and feedback generation for student responses in open-ended algebraic problems [4, 5] by helping mitigate potential biases [12] by facilitating an objective understanding of the potential mechanisms influencing the students' responses.

It would also be of academic interest for subsequent studies to investigate the interplay between these classifications and various demographic or evaluative indicators, such as levels of math anxiety. Such a focus can illuminate nuanced patterns of interaction across heterogeneous student groups. By doing so, we can better inform and adapt educational strategies, aiming to enhance both the inclusivity and efficacy of gamified instructional methodologies.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through an Efficacy and Replication Grant (R305A180401) and an NSF CAREER Grant (2142984) to Worcester Polytechnic Institute. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. Additionally, thanks to the LASER Institute for supporting this work.

REFERENCES

[1] Vincent Aleven, Bruce Mclaren, Ido Roll, and Kenneth Koedinger. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education* 16, 2 (2006), 101–128.

[2] John R Anderson. 2014. *Rules of the mind.* Psychology Press.

[3] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.

[4] Sami Baral, Anthony F Botelho, Abhishek Santhanam, Ashish Gurung, John Erickson, and Neil T Heffernan. 2023. Investigating patterns of tone and sentiment in teacher written feedback messages. In *International Conference on Artificial Intelligence in Education*. Springer, 341–346.

[5] S Baral, A Santhanam, A Botelho, A Gurung, and N Heffernan. 2023. Automated Scoring of Image-based responses to Open-ended mathematics question.. In *The Proceedings of the 16th International Conference on Educational Data Mining*.

[6] Jenny Yun-Chen Chan, Ji-Eun Lee, Craig A Mason, Katharine Sawrey, and Erin Ottmar. 2022. From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. *Journal of Educational Psychology* 114, 1 (2022), 56.

[7] Jenny Yun-Chen Chan, Erin R Ottmar, and Ji-Eun Lee. 2022. Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency. *Learning and Individual Differences* 93 (2022), 102109.

[8] Albert T Corbett, Kenneth R Koedinger, and John R Anderson. 1997. Intelligent tutoring systems. In *Handbook of human-computer interaction*. Elsevier, 849–874.

[9] Lauren E Decker-Woodrow, Craig A Mason, Ji-Eun Lee, Jenny Yun-Chen Chan, Adam Sales, Allison Liu, and Shihfen Tu. 2023. The impacts of three educational technologies on algebraic understanding in the context of COVID-19. *AERA open* 9 (2023), 23328584231165919.

[10] Ashish Gurung, Sami Baral, Morgan P Lee, Adam C Sales, Aaron Haim, Kirk P Vanacore, Andrew A McReynolds, Hilary Kreisberg, Cristina Heffernan, and Neil T Heffernan. 2023. How Common are Common Wrong Answers? Crowdsourcing Remediation at Scale. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*. 70–80.

[11] Ashish Gurung, Sami Baral, Kirk P Vanacore, Andrew A Mcreynolds, Hilary Kreisberg, Anthony F Botelho, Stacy T Shaw, and Neil T Hefferna. 2023. Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 399–410.

[12] Ashish Gurung, Anthony Botelho, Russell Thompson, Adam Sales, Sami Baral, and Neil Heffernan. 2022. Considerate, unfair, or just fatigued? examining factors that impact teacher. In *Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.*

[13] Ashish Gurung, Anthony F Botelho, and Neil T Heffernan. 2021. Examining student effort on help through response time decomposition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 292–301.

[14] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24 (2014), 470–497.

[15] Common Core State Standards Initiative et al. 2010. Common core state standards for mathematics. *http://www. corestandards. org/assets/CCSSI_Math% 20Standards. pdf* (2010).

[16] Kenneth R Koedinger, Elizabeth A McLaughlin, and Neil T Heffernan. 2010. A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research* 43, 4 (2010), 489–510.

[17] Kenneth R Koedinger and Mitchell J Nathan. 2004. The real story behind story problems: Effects of representations on quantitative reasoning. *The journal of the learning sciences* 13, 2 (2004), 129–164.

[18] Ji-Eun Lee, Jenny Yun-Chen Chan, Anthony Botelho, and Erin Ottmar. 2022. Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development* 70, 5 (2022), 1575–1599.

[19] Wes McKinney et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[20] Michael Mendicino, Leena Razzaq, and Neil T Heffernan. 2009. A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education* 41, 3 (2009), 331–359.

[21] Robert Murphy, Jeremy Roschelle, Mingyu Feng, and Craig A Mason. 2020. Investigating efficacy, moderators and mediators for an online mathematics homework intervention. *Journal of Research on Educational Effectiveness* 13, 2 (2020), 235–270.

[22] Erin Ottmar, Ji-Eun Lee, Kirk Vanacore, Siddhartha Pradhan, Lauren Decker-Woodrow, and Craig A Mason. 2023. Data from the Efficacy Study of From Here to There! A Dynamic Technology for Improving Algebraic Understanding. *Journal of Open Psychology Data* 11 (2023), 5.

[23] John F Pane, Beth Ann Griffin, Daniel F McCaffrey, and Rita Karam. 2014. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis* 36, 2 (2014), 127–144.

[24] National Mathematics Advisory Panel. 2008. *Foundations for success: The final report of the National Mathematics Advisory Panel.* US Department of Education.

[25] Steven Ritter and Stephen Fancsali. 2016. MATHia X: The Next Generation Cognitive Tutor.. In *EDM*. ERIC, 624–625.

[26] Katherine M Robinson, Jerilyn E Ninowski, and Melissa L Gray. 2006. Children's understanding of the arithmetic concepts of inversion and associativity. *Journal of experimental child psychology* 94, 4 (2006), 349–362.

[27] Nyet Moi Siew, Jolly Geofrey, and Bih Ni Lee. 2016. Students' algebraic thinking and attitudes towards algebra: the effects of game-based learning using Dragonbox 12+ App. *The Research Journal of Mathematics and Technology* 5, 1 (2016), 66–79.

[28] Saiying Steenbergen-Hu and Harris Cooper. 2013. A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of educational psychology* 105, 4 (2013), 970.

[29] Kirk Vanacore, Adam Sales, Allison Liu, and Erin Ottmar. 2023. Benefit of Gamification for Persistent Learners: Propensity to Replay Problems Moderates Algebra-Game Effectiveness. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 164–173. https://doi.org/10.1145/3573051.3593395

[30] Hamsa Venkat, Mike Askew, Anne Watson, and John Mason. 2019. Architecture of mathematical structure. *For the Learning of Mathematics* 39, 1 (2019), 13–17.

[31] Erik Weitnauer, David Landy, and Erin Ottmar. 2016. Graspable math: Towards dynamic algebra notations that support learners better than paper. In *2016 Future Technologies Conference (FTC)*. IEEE, 406–414.

[32] Le Xu, Ru-De Liu, Jon R Star, Jia Wang, Ying Liu, and Rui Zhen. 2017. Measures of potential flexibility and practical flexibility in equation solving. *Frontiers in Psychology* 8 (2017), 1368.