Moving beyond Test Scores: Analyzing the Effectiveness of a Digital Learning Game through Learning Analytics

Huy Anh Nguyen
Carnegie Mellon University
hn1@cs.cmu.edu

Xinying Hou
Carnegie Mellon University
xhou@cs.cmu.edu

John Stamper
Carnegie Mellon University
jstamper@cs.cmu.edu

Bruce M. McLaren
Carnegie Mellon University
bmclaren@cs.cmu.edu

ABSTRACT

A challenge in digital learning games is assessing students' learning behaviors, which are often intertwined with game behaviors. How do we know whether students have learned enough or needed more practice at the end of their game play? To answer this question, we performed post hoc analyses on a prior study of the game *Decimal Point*, which teaches decimal numbers and decimal operations to middle school students. Using Bayesian Knowledge Tracing, we found that students had the most difficulty with mastering the number line and sorting skills, but also tended to over-practice the skills they had previously mastered. In addition, using students' survey responses and in-game measurements, we identified the best feature sets to predict test scores and self-reported enjoyment. Analyzing these features and their connections with learning outcomes and enjoyment yielded useful insights into areas of improvement for the game. We conclude by highlighting the need for combining traditional test measures with rigorous learning analytics to critically evaluate the effectiveness of learning games.

Keywords

Decimal, Digital Learning Game, Bayesian Knowledge Tracing, Over-practice

1. INTRODUCTION

Digital learning games are typically regarded as a powerful tool to promote learning by engaging students with a novel and interactive game environment. While there have been concerns about the lack of empirical results on learning games' effectiveness [21, 32], recently we have seen more research that addresses this issue by showing students' learning gains from pretest to posttest in rigorous randomized experiments [9, 41, 52]. More generally, a meta-analysis of 69 studies by [10] showed that game conditions promoted significantly more learning than non-game conditions with equivalent knowledge content, and that augmented game designs with more learning-oriented features were more instructionally effective than standard designs.

While this prior research has demonstrated that digital learning games can enhance learning, the next step is to examine how they do so. In particular, even though the common measures of pretest and posttest scores are necessary to evaluate students' transferable learning, they are inadequate to address many questions about how learning takes place during the game. For example, did students get just enough practice from the game, or more practice than necessary? How does in-game learning correlate with test performance? These questions have been explored in great detail in Intelligent Tutoring Systems (ITS), but not as much in digital learning games, primarily because of the differences in design approaches between these two platforms. ITS are typically very structured environments where students are frequently evaluated on their knowledge and, in the mastery learning settings [28], move to a new skill as soon as the system determines they have mastered the current skill. In contrast, digital learning games emphasize students' freedom in shaping their own learning experience without concern about the consequences of failure [15]; as a result, the game's learning objectives are not always obvious to the students [4]. The question, then, is how can we combine the traditional pretest and posttest measures in learning game studies with learning analytics methods from ITS to paint a better picture of students' learning, both inside and outside of the game context? Furthermore, given the game's dual goal of promoting both learning and enjoyment, do in-game learning metrics also relate to students' enjoyment in any meaningful way?

Our work explores these questions in the context of *Decimal Point*, a game that teaches decimal numbers and operations to middle-school students. Here we present a post hoc analysis of the data from a prior study [22]. First, we investigated how well students mastered the in-game skills, how long it took them to master each skill, and whether students continued practicing after mastery. Next, we used student data from before and during game play to predict their learning outcomes and enjoyment after the game. Based on this re-

sult, we derived lessons for improving learning support in *Decimal Point* as well as in a more general learning game context.

2. RELATED WORK

2.1 Learning Analytics in Games

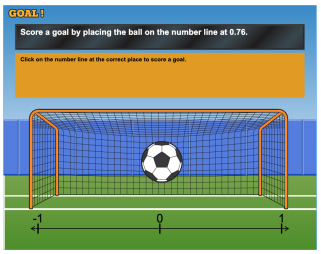
In-game formative assessment can be a powerful complementary tool for capturing students' learning progress [59]. Traditional formative measures typically make use of game-based metrics, such as the number of completed levels or the highest level beaten [2, 11], but these metrics may not always align with actual learning. Prior studies on *Decimal Point*, for instance, reported that students who played more mini-game rounds did not learn more than those who played fewer [18, 39]. An alternative approach is to employ learning analytics methods from ITS studies. For example, learning curve analysis, which visualizes students' error rates over time, has been applied in several learning games and yielded valuable insights that range from instructional redesign lessons to discovery of unforeseen strategy by students [17, 29, 42].

Learning analytics techniques can also connect formative assessment with external performance. For example, Bayesian networks have been applied to predict posttest responses from students' in-game data in several learning games [30, 48, 54]. Similarly, [27] employed feature engineering and gradient boosted random forest algorithm to identify struggling students in real-time in a physics learning game. Recently we have also seen more usage of deep learning for this prediction task [24, 51]. In general, research work in this direction can illustrate how well students' learning aligns with the game's learning objectives, while also guiding the development of adaptive support game features.

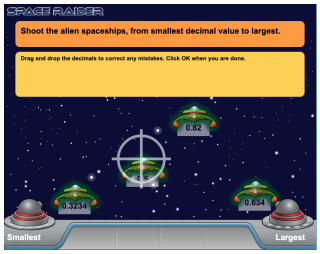
2.2 Decimal Point

Decimal Point is a web-based single-player digital learning game that helps middle-school students learn about decimal numbers and their operations (e.g., adding and comparing). The game features an amusement park metaphor, with a map of the park used to guide students (Figure 1). There are 8 theme areas with 24 mini-games, connected by a line that is designed to interleave skill types and theme areas. Each mini-game is aimed at helping students solve one of the common decimal misconceptions: `Megz` (longer decimals are larger), `Segz` (shorter decimals are larger), `Pegz` (the two sides of a decimal number are separate and independent) and `Negz` (decimals smaller than 1 are treated as negative numbers) [25]. Also, each mini-game calls for one of the following skills:

1. `Addition`: add two decimals by entering the carry digits and the sum.
2. `Bucket`: compare given decimals to a threshold number and place each decimal in a "less than" or "greater than" bucket.
3. `Number Line`: locate the position of a decimal number on the number line.
4. `Sequence`: fill in the next two numbers of a sequence of decimal numbers.
5. `Sorting`: sort a list of decimal numbers in ascending or descending order.



(a) Goal



(b) Space Raider

Figure 1: Screenshots of the main map screen and two example mini-games. Goal is a `Number Line` game and Space Raider is a `Sorting` game.

In each mini-game, students solve a number of decimal problems related to the game's targeted skill and receive immediate feedback about the correctness of their answers. Students don't face penalty on incorrect responses and can resubmit answers as many times as needed; however, they are not allowed to move forward without solving all the problems in the mini-game. More details about the instructional content of the mini-game problems can be found in [35].

The original study of *Decimal Point* showed that the game led to more learning and enjoyment than a conventional tutor with the same instructional content [35]. Subsequent studies have integrated the element of agency into the game, by endorsing students to select their preferred mini-games to play and stopping time [18, 39]. Based on their findings, students who were provided agency acquired equivalent learning gains in less time than those who were not. Most recently, a study by [22] compared two versions of the game, one that encourages students to play to learn, and one that encourages them to play for fun. Their results indicated that the learning-oriented group focused on re-practicing the same mini-games, while the enjoyment-oriented group did more exploration of different mini-games. In general, while all of these previous works reported that students learned from the game across all study conditions, it is not yet clear which game factors contributed to these findings. Furthermore, no connection between students' learning and their enjoyment has been identified. Our work aims at acquiring more insights into these areas.

Table 1: Survey items before and after game play.

Pre-intervention surveys		
Dimension (item count)	**Example statement**	Cronbach's α
Decimal efficacy (3) [44]	I can do an excellent job on decimal number math assignments.	.83
Computer efficacy (3) [31]	I know how to find information on a computer.	.71
Identification agency (2) [50]	I work on my classwork because I want to learn new things.	.60
Intrinsic agency (2) [50]	I work on my classwork because I enjoy doing it.	.86
External agency (3) [50]	I work on my classwork so the teacher won't be upset with me.	.61
Perseverance (3) [12]	Setbacks don't discourage me. I don't give up easily.	.79
Math utility (3) [13]	Math is useful in everyday life.	.63
Math interest (2) [14]	I find working on math to be very interesting.	.75
Expectancy (1) [23]	I plan to take the highest level of math available in high school.	-
Post-intervention surveys		
Dimension (item count)	**Example statement**	
Affective engagement (3) [5]	I felt frustrated or annoyed.	.78
Cognitive engagement (3) [5]	I tried out my ideas to see what would happen.	.54
Game engagement (5) [7]	I lost track of time.	.74
Achievement emotion (6) [43]	Reflecting on my progress in the game made me happy.	.89

3. DATASET

Our work uses data from 159 fifth and sixth grade students in our prior study [22], where students could select and play the mini-games from the map in Figure 1 in any order, and were allowed to stop playing at any time after finishing 24 mini-game rounds. They could also play more rounds of the completed mini-games, with the same game mechanics but different question content. For example, the first round of the mini-game `Goal` asks students to locate 0.76 on the number line, while the second round features the same game interactions but involves locating 0.431. Before playing, students did a pretest and answered demographic survey questions. After game play, they completed another survey to evaluate their experience and did a posttest, followed by a delayed posttest one week later. Here we outline the measures which are relevant to our analyses. A more detailed description of the experimental design can be found in [22].

Pretest, Posttest, and Delayed Posttest: Each test consisted of 43 items, for a total of 52 points. The items were designed to probe for specific decimal misconceptions, and involved either the five decimal skills targeted by the game or conceptual questions (e.g., "is a longer decimal larger than a shorter decimal?"). There are three test versions (A, B and C), which are isomorphic to one another and counterbalanced across students (e.g., ABC, ACB, BAC, etc. for pre, post, and delayed). Our prior analysis showed no differences in difficulty between the three versions [22].

Questionnaires: Before game play, students reported their age and gender, as well as their ratings to survey items about their background information, from 1 ("Strongly Disagree") to 5 ("Strongly Agree"). After playing, students rated their enjoyment (also from 1 to 5) via survey questions that address four enjoyment dimensions (Table 1). If a dimension comprises several items, we compute the average ratings of all items in that dimension to derive its representative rating score. According to [16], a measure should have $\alpha \geq .60$ to be considered reliable; therefore, based on Table 1 we removed the cognitive engagement dimension (with $\alpha = .54$) from further analyses.

The full log data from the study is archived in the DataShop repository [55], in dataset number 3086. We present our analysis of this data in the following section.

4. RESULTS

4.1 Investigating in-game learning

In our prior work on Knowledge Component (KC) modeling in *Decimal Point*, based on data from a separate study, we used the correctness of the student's first attempt in answering each mini-game problem to update their mastery of the KC covered by that mini-game. With this mapping from in-game action to KC, we found that students' learning can be better captured by a KC model based on skill types (e.g., `Addition`, `Bucket`) than on decimal misconceptions (e.g., `Segz`, `Negz`) [40]. Therefore, in this work we used the five skill types as our KCs, and tracked students' learning progress of these skills by Bayesian Knowledge Tracing (BKT) [60]. The BKT parameters were set as $p(L_0) = 0.4$, $p(T) = 0.05$, $p(S) = p(G) = 0.299$ [3], and the mastery threshold is 0.9.

First, we looked at how well students mastered each of the five skills in the game. Comparing the students' final mastery probabilities in each skill and our mastery threshold,

we observed that: there were 4 students who did not master any skill, 20 students who mastered one skill, 33 students who mastered two skills, 42 students who mastered three skills, 34 students who mastered four skills, and 26 students who mastered all five skills. Next, we counted how many opportunities each student who mastered a skill took to reach mastery in that skill. An opportunity is defined as one complete decimal exercise; each mini-game round consists of one opportunity, except for those in `Sequence`, which contain three opportunities (i.e., students have to fill in three decimal sequences per round). The distributions of opportunity count until mastery are plotted in Figure 2, which shows that `Number Line` and `Sorting` took the longest to master, at around 5 opportunities on average. For `Number Line`, one student even needed 26 opportunities to reach mastery.

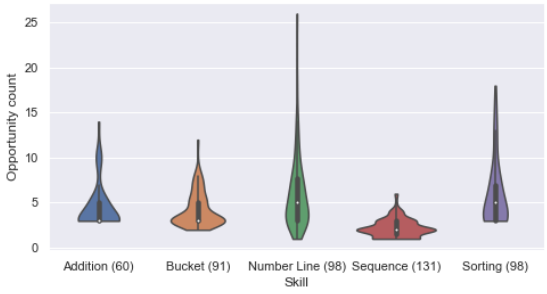


Figure 2: Opportunity counts until mastery for each skill. The number next to each skill indicates the count of students who mastered that skill and were included in the violin plot.

Next, we examined how well students regulated their learning, i.e., after mastering a skill, did they tend to continue practicing the same skill, or switch to a different skill? For each student, following [8], once they mastered a skill ($\geq$ 90% mastery probability), we considered their subsequent opportunities as over-practice. Then, for each student who mastered a particular skill, we computed the ratio between their over-practice count and total opportunity count in that skill. Plotting these ratios for all the mastered students in each skill (Figure 3), we observed that between 20-80% of a student's practice opportunities in a skill could be considered over-practice, i.e., they took place after the student had mastered the skill.

4.2 Investigating factors related to posttest and delayed posttest performance

Having examined students' in-game learning, we then looked at how it related to test performance after the game. In order to predict posttest and delayed posttest scores, we collected features that reflected students' in-game learning and also included demographic measures that account for individual student differences. In total, we considered 19 features: pretest score, decimal efficacy, gender, computer efficacy, identification agency, intrinsic agency, external agency, perseverance, utility, math interest, expectancy, final in-game mastery probabilities of the five skills (`Addition`, `Bucket`, `Sequence`, `Number Line`, `Sorting`), total opportunity count, over-practice opportunity count and total incorrect answer counts. To identify the most important features, we (1) performed feature selection with linear regression, and (2) ran another linear regression model with the selected features on the full dataset to inspect the coefficient and significance of each feature. In step (1), we use the `mlxtend` library [45] to run a forward feature selection procedure that returns the feature subset with the best cross-validated performance, measured in terms of mean squared error (MSE).

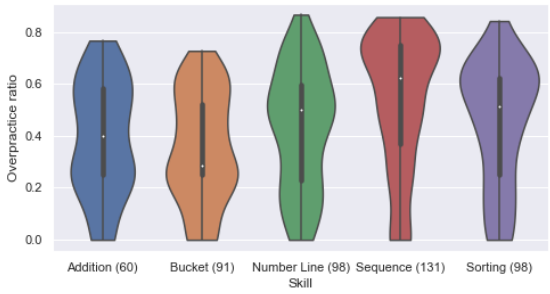


Figure 3: Over-practice ratio in each skilll. The number next to each skill indicates the count of students who mastered that skill and were included in the violin plot.

In predicting posttest scores, our feature selection identified three features: `Bucket` mastery, `Sorting` mastery and pretest score. A linear regression model with these three features, when trained and evaluated on the entire dataset, had an MSE of 26.167 and an adjusted R^2 of .735. Based on the regression table, the coefficient and significance of each feature was as follow: pretest score with $\beta = 0.734$, $p < .001$, `Bucket` mastery with $\beta = 6.833$, $p < .001$, `Sorting` mastery with $\beta = 5.100$, $p = .001$. In other words, pretest scores, `Bucket` mastery and `Sorting` mastery each had a positive and significant association with posttest scores.

The delayed posttest model incorporated two additional features – `Number Line` mastery and gender – and yielded an MSE of 24.218, as well as an adjusted R^2 of .747. Based on the regression table, the coefficient and significance of each feature was as follows: pretest score with $\beta = 0.730$, $p < .001$, `Bucket` mastery with $\beta = 4.276$, $p = .018$, `Sorting` mastery with $p = 4.270$, $p = .003$, `Number Line` with $\beta = 3.099$, $p = .029$, and gender with $\beta = 1.426$, $p = .074$. In other words, the three skill mastery values – `Bucket`, `Sorting`, `Number Line` – as well as pretest score each had a positive and significant association with delayed posttest score, while gender (male = 0, female = 1) had a positive and marginally significant association

4.3 Investigating factors related to enjoyment

For each enjoyment dimension measured in post-intervention surveys (achievement emotion, game engagement, affective engagement - see Table 1), we computed the per-student average Likert scores to the statements in that dimension. Then, we performed the same feature selection procedure as in 4.2 and reported our results in Table 2.

We observed that the adjusted R^2 values of the game engagement and affective engagement models were much lower

Table 2: Results of feature selection for predicting game enjoyment. The Overall performance row indicates the selected model's scores when trained and evaluated on the entire dataset.

	Achievement Emotion	Game Engagement	Affective Engagement
Selected features	computer efficacy, identification agency, intrinsic agency, math interest, pretest score, total opportunity count	math interest, computer efficacy, gender	decimal efficacy, gender, intrinsic agency, `Sorting` mastery, `Bucket` mastery, total incorrect attempt count, identification agency
Overall performance	MSE = 0.520 Adjusted $R^2 = 0.386$	MSE = 0.602 Adjusted $R^2 = 0.225$	MSE = 0.660 Adjusted $R^2 = 0.218$

than those of the test score models. Even when trained and evaluated on the entire dataset, Linear Regression could only explain about 20% of the variance in game engagement and affective engagement. On the other hand, the achievement emotion model did have reasonable performance (adjusted $R^2 = .386$), so we focused on analyzing the features in this model. The linear regression table showed the coefficient and significance of each feature as follows: computer efficacy with $\beta = 0.047$, $p = .063$, identification agency with $\beta = 0.099$, $p = .024$, intrinsic agency with $\beta = 0.116$, $p = .002$, math interest with $\beta = 0.114$, $p = .001$, pretest score with $\beta = -0.017$, $p = .011$, opportunity count with $\beta = 0.009$, $p = .033$. In other words, computer efficacy had a positive and marginally significant association, while pretest score had a negative and significant association; the remaining features (identification agency, intrinsic agency, math interest and opportunity count) each had a positive and significant association.

5. DISCUSSION

5.1 Investigating in-game learning

Based on the opportunity count until mastery in each skill (Figure 2), we identified `Sorting` and `Number Line` as the most difficult skills in the game. Our prior learning curve analysis [40] on a different *Decimal Point* study reported a consistent finding – that the learning curves of these two skills were mostly flat and reflected small learning rates. Based on previous research in decimal learning, a plausible explanation is that there are several misconceptions which can lead to students making a mistake in `Sorting` or `Number Line` problems, including (1) treating decimals as whole numbers, (2) treating decimals as fractions, and (3) ignoring the zero in the tenths place [46]. Furthermore, even when students recognize their misconception, they may shift to a different misconception instead of arriving at the correct understanding [56]. This phenomenon likely also occurred in *Decimal Point*, as the game provides corrective feedback (whether an answer is right or wrong) but does not emphasize the underlying reasoning; consequently, as an example, a student realizing it is wrong to assume longer decimals are larger may end up concluding that shorter decimals must be larger, thereby adopting a new misconception. This highlights the need for more refined tracing of the student's dynamic learning states in a digital learning environment. While the standard KC modeling technique can track when students make an intended mistake (e.g., longer decimals are larger), it does not investigate their specific input to see whether a new misconception (e.g., shorter decimals are larger) has emerged. To address this issue, future iterations of the game should provide more instructional support that can react to various misconceptions from students, for example via explanatory feedback [19] or predefined error messages for different types of error [36].

Once students have mastered a skill, however, our analysis showed that over-practice was very common, i.e., students kept playing more mini-games in the mastered skill. At the same time, there were only 26 out of 159 students who mastered all five skills, suggesting that the majority of students still had room for improvement in the unmastered skills but chose not to practice them. One possible reason is that the game environment did not explicitly indicate when the student has reached mastery or force them to switch to practicing a different skill. Consequently, young students, who were likely to be weak at self-regulated learning [37,53], simply played the mini-games that they thought were engaging, which in this case involved the skills they had already mastered. A prior study by [29] similarly found that, in a game about locating fractions on number line, students were more engaged when the game was easier, contradicting game design theories that optimal engagement would occur at moderate difficulty level.

5.2 Investigating factors related to posttest and delayed posttest performance

We saw that our linear regression models were able to predict posttest and delayed posttest performance well, capturing about 75% of the variance in test scores with only 3-5 features. The three features present in both models are pretest score, `Sorting` mastery and `Bucket` mastery. The inclusion of pretest score is not surprising, as it is consistent with the standard practice of controlling for prior knowledge when analyzing posttest score [58]. On the other hand, both `Sorting` mastery and `Bucket` mastery suggest that the ability to compare decimal numbers plays a large role in test performance. This is likely due to the game and test materials focusing on the four most common decimal misconceptions (`Megz`, `Segz`, `Pegz`, `Negz`), three of which are related to decimal comparison [25]. Based on the distribution of practice opportunities until mastery, however, students took much more attempts to master `Sorting` problems than `Bucket` problems, which may explain why they did not achieve high scores on the posttest and delayed posttest, averaging at only around 30 out of 52 points [22]. Therefore, improving students' performance on `Sorting` problems, potentially by incorporating hints and error messages as we previously discussed, is crucial in future studies of the game.

At the same time, we saw that `Number Line` mastery had a significant positive association with delayed posttest score, but was not selected in the posttest model. An interpretation of this result is that `Number Line` tasks, which we identified as among the most difficult in the game, could be at a *desirable difficulty* level, which can promote deeper and longer-lasting learning than the more straightforward tasks [61]. For instance, a prior study on comparing erroneous examples and problem-solving decimal tasks found that erroneous examples, which are more aligned with the desirable difficulty, led to significantly higher delayed posttest scores but similar posttest scores [34]. In our case, we also saw that `Number Line` is an important feature for predicting delayed posttest but not for predicting posttest performance.

Similar to `Number Line` mastery, gender (male = 0, female = 1) was not a feature in the posttest model, but had a positive association with delayed posttest scores. In other words, with other factors being equal, females could achieve higher delayed posttest scores than males. While this association is only marginally significant ($p = .074$), similar findings about females' tendency to outperform males in retention and delayed posttest have been reported in previous mathematics intervention studies [1, 20]. Using the same dataset as in this work, [22] also found that females demonstrated significantly higher pre-post and pre-delayed learning gains than males, with a larger effect size in pre-delayed learning gains. Therefore, an important next step is to conduct future studies of *Decimal Point* on a larger sample size to draw more conclusive findings about whether the game promotes more retention in females and what could lead to this effect.

5.3 Investigating factors related to enjoyment

Our enjoyment prediction models did not perform as well as the learning models and could explain only about 20% of the variance in game engagement and affective engagement. These poor model fits likely result from the lack of appropriate features in our data. To track student engagement, previous work has emphasized the use of fine-grained measures such as time spent on decision making [47], social engagement profile [49] and interaction traces [6]; in contrast, our feature set consists mainly of quantitative scores (e.g., Likert responses) and aggregate data (e.g., error count). Related to this direction, a previous study of *Decimal Point* by [57] has clustered students based on their mini-game selection orders and found that the cluster which demonstrated more agency reported higher enjoyment. Adopting their method of encoding students' mini-game sequences is a good first step in building more fine-grained features for our prediction tasks. On the other hand, the lack of association between our in-game learning measures (e.g., skill mastery, over-practice opportunity count, error count) and game engagement or affective engagement implies that students' game performance, whether good or bad, were unlikely to yield any negative emotion such as confusion or frustration. This is a positive outcome, indicating that our game environment does not impose any performance pressure on students – one of the primary principles of learning games [15].

At the same time, we did find that a linear regression model was able to predict achievement emotion reasonably well from student's identification agency, intrinsic agency, math interest, computer efficacy, pretest score and opportunity count. Identification and intrinsic agency indicate that, with all other factors being equal, the more students identified their learning as coming from intrinsic motivation (rather than external pressures), the more achievement they felt after learning. Math interest and computer efficacy suggest that students' acquaintance with the learning domain or medium could also be positively associated with achievement emotion [26]. On the other hand, pretest score had a negative association, likely because students with lower prior knowledge were able to learn more from the game and therefore felt more achievement than those with high prior knowledge. Similarly, for opportunity count, a plausible reason for students choosing to play more mini-game rounds is that they felt the mini-games were helpful, which contributed to their achievement emotion after game play. Overall, the features we identified could serve as a guideline for promoting achievement emotion in learning games and in more general instructional contexts.

6. CONCLUSIONS

From our analyses, we gained several insights into students' learning outcomes and enjoyment in *Decimal Point*. First, we found that `Sorting` and `Number Line` are important skills for posttest and delayed posttest performance, but students required more instructional support to effectively master them. Second, very few students mastered all five decimal skills from the game, while the majority engaged in over-practice, likely due to their preference for playing easy mini-games, i.e., those they had already mastered. Third, expanding on prior findings about gender effect in *Decimal Point* [22,33], we identified a trend of females outperforming males in the delayed posttest, which should be investigated on a larger sample size. Fourth, we learned that students' achievement emotion can be reasonably captured by their level of computer efficacy, learning motivation, prior knowledge and number of mini-game rounds. All of these insights can be derived from log data alone and would serve as useful metrics to assist digital learning game researchers in evaluating and improving their own games. For *Decimal Point*, in particular, an important next step is to perform similar analyses in other studies of the game to see which of our findings can be replicated. Identifying consistent trends in student data could allow us to construct a more generalized model of students' game play that combines existing theories with novel exploratory analyses [38].

In a broader context, we have seen the rapid growth of digital learning games in recent years, from being conceived as a novel learning platform [15, 21] to having their effectiveness validated by rigorous studies [10]. The game *Decimal Point*, in particular, has been shown to significantly improve students' learning through several research works [18,22,35,39]. When viewing from a learning analytics perspective, however, one could identify room for improvement that would otherwise not be reflected in pretest and posttest scores alone. For instance, a game may not adequately support all of its learning objectives, or students may engage in non-optimal learning behavior due to a lack of self-regulation. At the heart of these issues is the question of how digital learning games can optimize student learning while retaining its core value as a playful environment, where players are free to exercise their agency. Addressing this question is an important step for future works in the field.

7. REFERENCES

[1] J. Ajai and B. Imoko. Gender differences in mathematics achievement and retention scores: A case of problem-based learning method. *International Journal of research in Education and Science*, 1(1):45–50, 2015.

[2] E. Andersen, E. O'Rourke, Y.-E. Liu, R. Snider, J. Lowdermilk, D. Truong, S. Cooper, and Z. Popovic. The impact of tutorials on games of varying complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 59–68, 2012.

[3] R. S. Baker. Personal correspondence, 2019.

[4] R. S. Baker, M. J. Habgood, S. E. Ainsworth, and A. T. Corbett. Modeling the acquisition of fluent skill in educational action games. In *International Conference on User Modeling*, pages 17–26. Springer, 2007.

[5] A. Ben-Eliyahu, D. Moore, R. Dorph, and C. D. Schunn. Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology*, 53:87–105, 2018.

[6] P. Bouvier, K. Sehaba, and É. Lavoué. A trace-based approach to identifying users' engagement and qualifying their engaged-behaviours in interactive systems: application to a social game. *User Modeling and User-Adapted Interaction*, 24(5):413–451, 2014.

[7] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, 2009.

[8] H. Cen, K. R. Koedinger, and B. Junker. Is over practice necessary?-improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in artificial intelligence and applications*, 158:511, 2007.

[9] C.-H. Chen, K.-C. Wang, and Y.-H. Lin. The comparison of solitary and collaborative modes of game-based learning on students' science learning and motivation. *Journal of Educational Technology & Society*, 18(2):237–248, 2015.

[10] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth. Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, 86(1):79–122, 2016.

[11] G. C. Delacruz, G. K. Chung, and E. L. Baker. Validity evidence for games as assessment environments. cresst report 773. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*, 2010.

[12] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6):1087, 2007.

[13] A. M. Durik, M. Vida, and J. S. Eccles. Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, 98(2):382, 2006.

[14] W. Fan and C. A. Wolters. School motivation and high school dropout: The mediating role of educational expectation. *British Journal of Educational Psychology*, 84(1):22–39, 2014.

[15] J. P. Gee. What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, 1(1):20–20, 2003.

[16] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham, et al. Multivariate data analysis (vol. 6), 2006.

[17] E. Harpstead and V. Aleven. Using empirical learning curve analysis to inform design in an educational game. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 197–207, 2015.

[18] E. Harpstead, J. E. Richey, H. Nguyen, and B. M. McLaren. Exploring the subtleties of agency and indirect control in digital learning games. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 121–129, 2019.

[19] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[20] L. L. Haynes and J. V. Dempsey. How and why students play computer-based mathematics games: A consideration of gender differences. *2001 Annual Proceedings-Atlanta: Volume*, page 178.

[21] M. A. Honey and M. L. Hilton. Learning science through computer games. *National Academies Press, Washington, DC*, 2011.

[22] X. Hou, H. Nguyen, J. E. Richey, and B. M. McLaren. Exploring how gender and enjoyment impact learning in a digital learning game. In *International Conference on Artificial Intelligence in Education*. Springer, 2020.

[23] C. S. Hulleman, O. Godes, B. L. Hendricks, and J. M. Harackiewicz. Enhancing interest and performance with a utility value intervention. *Journal of educational psychology*, 102(4):880, 2010.

[24] A. Illanas Vila, J. R. Calvo-Ferrer, F. J. Gallego-Durán, F. Llorens Largo, et al. Predicting student performance in foreign languages with a serious game. 2013.

[25] S. Isotani, D. Adams, R. E. Mayer, K. Durkin, B. Rittle-Johnson, and B. M. McLaren. Can erroneous examples help middle-school students learn decimals? In *European Conference on Technology Enhanced Learning*, pages 181–195. Springer, 2011.

[26] M. Jansen, O. Lüdtke, and U. Schroeders. Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology*, 46:116–127, 2016.

[27] S. Karumbaiah, R. S. Baker, and V. Shute. Predicting quitting in students playing a learning game. *International Educational Data Mining Society*, 2018.

[28] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[29] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, pages 89–98, 2013.

[30] M. Manske and C. Conati. Modelling learning in an educational game. In *AIED*, pages 411–418, 2005.

[31] G. Marakas, R. Johnson, and P. F. Clay. The evolving nature of the computer self-efficacy construct: An empirical investigation of measurement construction, validity, reliability and stability over time. *Journal of the Association for Information Systems*, 8(1):2, 2007.

[32] R. E. Mayer. *Computer games for learning: An evidence-based approach*. MIT Press, 2014.

[33] B. McLaren, R. Farzan, D. Adams, R. Mayer, and J. Forlizzi. Uncovering gender and problem difficulty effects in learning with an educational game. In *International Conference on Artificial Intelligence in Education*, pages 540–543. Springer, 2017.

[34] B. M. McLaren, D. M. Adams, and R. E. Mayer. Delayed learning effects with erroneous examples: a study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education*, 25(4):520–542, 2015.

[35] B. M. McLaren, D. M. Adams, R. E. Mayer, and J. Forlizzi. A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning (IJGBL)*, 7(1):36–56, 2017.

[36] B. M. McLaren, S.-J. Lim, D. Yaron, and K. R. Koedinger. Can a polite intelligent tutoring system lead to improved learning outside of the lab? *Frontiers in Artificial Intelligence and Applications*, 158:433, 2007.

[37] J. Metcalfe and N. Kornell. The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, 132(4):530, 2003.

[38] R. J. Mislevy, J. T. Behrens, K. E. Dicerbo, and R. Levy. Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *JEDM| Journal of Educational Data Mining*, 4(1):11–48, 2012.

[39] H. Nguyen, E. Harpstead, Y. Wang, and B. M. McLaren. Student agency and game-based learning: A study comparing low and high agency. In *International Conference on Artificial Intelligence in Education*, pages 338–351. Springer, 2018.

[40] H. Nguyen, Y. Wang, J. Stamper, and B. M. McLaren. Using knowledge component modeling to increase domain understanding in a digital learning game. In *International Conference on Educational Data Mining*, pages 139–148, 2019.

[41] M. Ninaus, K. Moeller, J. McMullen, and K. Kiili. Acceptance of game-based learning and intrinsic motivation as predictors for learning success and flow experience. 2017.

[42] Z. Peddycord-Liu, R. Harred, S. Karamarkovich, T. Barnes, C. Lynch, and T. Rutherford. Learning curve analysis in a large-scale, drill-and-practice serious math game: Where is learning support needed? In *International Conference on Artificial Intelligence in Education*, pages 436–449. Springer, 2018.

[43] R. Pekrun. Progress and open problems in educational emotion research. *Learning and Instruction*, 15(5):497–506, 2005.

[44] P. Pintrich, D. Smith, T. Garcia, and W. McKeachie. A manual for the use of the motivated strategies for learning questionnaire (mslq) ann arbor. *MI: National Center for Research to Improve Postsecondary Teaching and Learning*, pages 1–76, 1991.

[45] S. Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), Apr. 2018.

[46] L. B. Resnick, P. Nesher, F. Leonard, M. Magone, S. Omanson, and I. Peled. Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for research in mathematics education*, pages 8–27, 1989.

[47] V. Riemer and C. Schrader. Impacts of behavioral engagement and self-monitoring on the development of mental models through serious games: Inferences from in-game measures. *Computers in Human Behavior*, 64:264–273, 2016.

[48] J. P. Rowe and J. C. Lester. Modeling user knowledge with dynamic bayesian networks in interactive narrative environments. In *Sixth AI and Interactive Digital Entertainment Conference*, 2010.

[49] J. A. Ruiperez-Valiente, M. Gaydos, L. Rosenheck, Y. J. Kim, and E. Klopfer. Patterns of engagement in an educational massive multiplayer online game: A multidimensional view. *IEEE Transactions on Learning Technologies*, 2020.

[50] R. M. Ryan and J. P. Connell. Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of personality and social psychology*, 57(5):749, 1989.

[51] J. L. Sabourin, L. R. Shores, B. W. Mott, and J. C. Lester. Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education*, 23(1-4):94–114, 2013.

[52] R. Sawyer, A. Smith, J. Rowe, R. Azevedo, and J. Lester. Is more agency better? the impact of student agency on game-based learning. In *International Conference on Artificial Intelligence in Education*, pages 335–346. Springer, 2017.

[53] W. Schneider. The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, 2(3):114–121, 2008.

[54] V. J. Shute, L. Wang, S. Greiff, W. Zhao, and G. Moore. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63:106–117, 2016.

[55] J. Stamper, K. Koedinger, R. S. d Baker, A. Skogsholm, B. Leber, J. Rankin, and S. Demi. Pslc datashop: A data analysis service for the learning science community. In *International Conference on Intelligent Tutoring Systems*, pages 455–455. Springer, 2010.

[56] W. Van Dooren, D. De Bock, A. Hessels, D. Janssens, and L. Verschaffel. Remedying secondary school students' illusion of linearity: A teaching experiment aiming at conceptual change. *Learning and Instruction*, 14(5):485–501, 2004.

[57] Y. Wang, H. Nguyen, E. Harpstead, J. Stamper, and B. M. McLaren. How does order of gameplay impact learning and enjoyment in a digital learning game? In *International Conference on Artificial Intelligence in Education*, pages 518–531. Springer, 2019.

[58] B. E. Whitley and M. E. Kite. *Principles of research in behavioral science*. Routledge, 2013.

[59] J. Wiemeyer, M. Kickmeier-Rust, and C. M. Steiner. Performance assessment in serious games. In *Serious Games*, pages 273–302. Springer, 2016.

[60] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*, pages 171–180. Springer, 2013.

[61] C. L. Yue, E. L. Bjork, and R. A. Bjork. Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology*, 105(2):266, 2013.