Designing Playful Intelligent Tutoring Software to Support Engaging and Effective Algebra Learning

Tomohiro Nagashima[1][0000-0003-2489-5016], John Britti[2], Xiran Wang[1], Bin Zheng[1], Violet Turri[1], Stephanie Tseng[1], and Vincent Aleven[1][0000-0002-1581-6657]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] Georgia Institute of Technology, Atlanta GA 30332, USA
tnagashi@cs.cmu.edu, jbritti3@gatech.edu, xiranw@andrew.cmu.edu, binzheng@andrew.cmu.edu, vturri@andrew.cmu.edu, stseng2@andrew.cmu.edu, aleven@cs.cmu.edu

Abstract. In designing learning technology, it is critical that the technology supports both learning and engagement of students. However, achieving both aspects in a single technology design is challenging. We report on the design and evaluation of Gwynnette, intelligent tutoring software for early algebra. Gwynnette was deliberately designed to enhance students' algebra learning and engagement, integrating several playful interaction and gamification features such as drag-and-drop interactions, an alien character, and sound effects. A virtual classroom experiment with 60 students showed that the system significantly enhanced both engagement and conceptual learning in early algebra, compared to the older version of the same software. Log data analyses gave insights into how the design might have affected the outcomes. This study demonstrates that a deliberate design of learning technology can help students learn and engage well in an unpopular subject such as algebra, a challenging dual goal in designing learning technologies.

Keywords: Intelligent Tutoring System, Engagement, Algebra

1 Introduction

When designing learning technology, it is critical that the technology is designed to support both learning and engagement of students [1]. An engaging technology with no learning support might entertain students but would not result in meaningful learning. On the other hand, a learning technology with no engaging features would not fully engage students in the learning activity, even if the activity is well designed to support learning. Designing learning technology for enhancing both learning and engagement is critical particularly in disciplines in which many students have a hard time succeeding, such as early algebra. Early algebra is considered as a "gatekeeper" course to advanced learning in Science, Technology, Engineering, and Mathematics (STEM) domains [2]. Many students, especially in the United States, fail to gain important knowledge and skills in algebra, including conceptual understanding and procedural skills [3, 4]. The difficulty in learning algebra may be partly attributed to the complexity of the symbolic notation system and prior practices in arithmetic problem solving [5] but may also come from a lack of student engagement. Indeed, it is very typical of students to perceive algebra as not being enjoyable [6].

How might one design learning technology that enhances both learning and engagement in algebra? Even though a number of systems have been shown to enhance students' algebra learning, such as Intelligent Tutoring Systems (ITSs) [7–9], the important question of how these effective technologies can also be designed for student engagement has received rather little attention [10]. Of a few attempts that have been made, learning environments with playful interactions and gamification have been developed and tested in the domain of early algebra. For example, *DragonBox* (https://dragonbox.com/products/algebra-12) is an algebra game in which learners drag and drop cards with "dragons" and other associated monsters that represent numbers and variables in equations, with the goal of isolating the variable [7, 11]. DragonBox has also a number of other entertaining elements in the game, including the "dragons" and sounds for various interactions in the system [11]. Another gamified learning environment, *From Here to There!* (FH2T*,* https://graspablemath.com/fh2t.html), also employs drag-and-drop interactions to help students understand the structure of algebraic expressions and how to change them using formal algebraic strategies [12, 13].

Empirical studies show, however, that these technologies have not yet achieved the goal of supporting both effective learning and engagement. For instance, researchers experimentally compared DragonBox against *Lynnette*, an effective ITS for algebra with no enjoyable elements [7]. Although secondary-school students found DragonBox more engaging than Lynnette (as self-reported by students), students who had used DragonBox performed poorer on a posttest, suggesting a poor learning effect. FH2T has also been evaluated in experimental studies, compared against a non-gamified condition involving problem sets with hints and feedback in *ASSISTments* (https://new.assistments.org) [12]. The findings show that FH2T helps students gain conceptual understanding of algebra compared to ASSISTments. Despite its success, prior studies on FH2T have not directly measured how engaging it is compared to other software [14], making it difficult to understand whether and how it may influence student engagement when learning in FH2T. To sum, although some environments for learning algebra have been shown engaging and others have been shown effective in algebra, no studies have rigorously measured both student learning *and* engagement how it compares with learning and engagement in other learning environments.

In the current study, we deliberately designed an ITS (called *Gwynnette*) based on an existing ITS (i.e., Lynnette) with the goal of enhancing both learning and engagement. Gwynnette embeds several playful features to make student learning effective and engaging. We present findings from a randomized controlled experiment in a virtual classroom environment with 60 secondary-school students, in which we compared Gwynnette against an older version of the same software (i.e., Lynnette) with no playful features. The results showed that Gwynnette enhanced students' engagement and learning; students spent considerably more time using the system and gained more conceptual understanding of algebra than those who used Lynnette. We describe the design principles, findings of the experiment and a detailed analysis of the log data to examine how students interacted with the software and how the interaction and the design of the software might have contributed to student learning and engagement.

2 Gwynnette

2.1 Design Principles

To develop a playful learning environment that can be both engaging and effective in supporting algebra learning, we designed an ITS for early algebra named Gwynnette (Fig. 1). Gwynnette was designed based on Lynnette, an existing ITS for algebra learning which has no common entertaining features [7]. Based on literature review and evaluation of designs in existing software (e.g., design evaluation of DragonBox by [7]), we added the following features in Gwynnette: playful drag-and-drop interactions to enact equation transformations in solving algebra problems, and enjoyable gamification elements including a game theme, sound effects, and a character guide. In this section, we briefly describe these design features. These features were designed following a user-centered design approach; we iteratively prototyped, tested, and improved ideas and artifacts with school teachers and students through virtual interview sessions and pilot use in a secondary-school classroom before the final implementation.

Fig. 1. The main interactions in Gwynnette are drag-and-drop manipulations of equations. Users would drag operators and numbers to manipulate given equations (left). Once an operator (e.g., a "–" symbol) has been dropped onto the appropriate area (e.g., the "3" in the equation), a square box appears (middle). Users can type in a number in the box to fill in the box (right). Users can also request hints anytime (here the hint says, "On the left side, you have the terms 3 and -3. These terms cancel each other out").

Focused Practice of Algebraic Manipulations with Drag-and-Drop Interactions. Using Gwynnette, students can enact equation transformations through its drag-and-drop interactions. As shown in Fig. 2, users can drag operators (e.g., the "+" sign) to an equation to transform the state of the equation. They can also drag a constant term (a number) or a variable term onto another like term to simplify equations. When learners make an error, an "Undo" button appears so that leaners can move back to the previous state. Like other learning environments introduced above [12], such dynamic interactions may be effective by focusing students' practice on equation transformations, rather than also having to take care of arithmetic calculations. This may help students learn "conceptual knowledge that underlies procedures" [4], including the concept of *doing the same thing to both sides* (i.e., when subtracting 3 from the left-hand side of $2x + 3 = 7$, students also need to subtract 3 from the other side of the equation) and the concept of equivalence [9]. Practice on equation transformations may help students focus their attention and cognitive effort to important algebraic thinking rather than arithmetic errors [9].

Fig. 2. Drag-and-drop interactions for adding an operator to a given equation (left) and simplifying an equation (the interface takes care of the arithmetic, right). Arrows show the location where the element being dragged will be dropped.

In designing the drag-and-drop interactions for manipulating equations, we made several design decisions aimed at helping learners effectively acquire algebra concepts, based on the findings from [7] and honed through design iterations with teachers and students. For instance, when learners drag a correct operator to one side of an equation, the system requires that the learner's next action is to drag the same operator to the other side to help learners understand the principle of *doing the same thing to both sides of an equation* [9]. This deliberate design was added to help students focus on important conceptual moves while avoiding "over-scaffolding" learners [7]. Also, to further emphasize the principle of doing the same thing to both sides, the system also requires learners to wait to type in a number in the added square box until they first add a square box to each side of the equation (Fig. 3). This deliberate interaction design may help learners focus their cognitive effort and attention to the important concept. Prior work has not considered this design. For example, in FH2T, when a learner drags a number from one side to the other side of an equation, the system automatically presents subtraction of that number on both sides of the equation (and then automatically flips the sign attached to the dragged number as learners progress in the game). It seems possible that doing so may remove a critical difficulty from the task, limiting students' opportunity to learn to handle it through practice, a form of over-scaffolding.

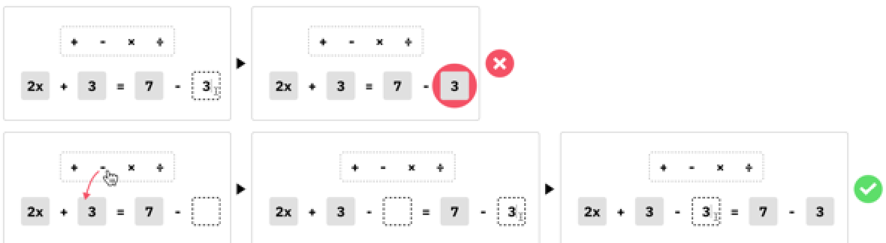

Fig. 3. Gwynnette facilitates the use of an algebra principle of *doing the same thing to both sides* before typing in a number in the added square. The system does not let users, after dragging one operator to one side, type in a number in the added square before dragging the same operator to the other side of the equation (top). Users need to first drag the same operator to both sides (creating two blank squares, bottom). The arrow shows that the negative sign is being dragged onto the "3." The green feedback indicates that the action is correct while the red feedback shows that the system rejects the input as an error.

The drag-and-drop interactions might also be perceived as a playful form of solving equations [16]. Solving equations using a drag-and-drop interaction is considerably different from a typical form of instruction and practice that involves writing out equations and their solution steps (as transformed equations). Using such a new way of solving problems might bring a new, engaging, and pleasant experience to students. However, even with the potential effectiveness, playful drag-and-drop interactions could also be unhelpful in enhancing algebra learning. For example, in the drag-and-drop interactions we implemented, users never get an experience of typing in or writing out equations and performing arithmetic operations. Without practicing those skills, users who use the drag-and-drop interactions to solve problems might not be able to apply their learned knowledge to typical algebra problems that ask students to write out equations and their solution steps in a paper or on a computer screen.

Space Travel Theme. Many gamified learning environments employ a fantasy theme (e.g., a monster adventure theme [17]) for their game context. Such a fantasy context helps create an immersive and engaging environment in which other elements (e.g., narratives) can also be integrated [18].

In designing our tutoring software, we adopted a theme of space travel, in which the learner becomes a space traveler exploring planets by solving problems. We conducted multiple design sessions with eight secondary-school students, exploring what feelings they have regarding the use of the space theme. An informal theme analysis found that students, regardless of their age or gender, have a positive view towards the theme.

Alien Guide. Use of avatars and characters is a common strategy employed widely across playful, gamified learning environments. By interacting with such characters, learners may gain a sense of autonomy [18] and therefore engage in their learning. We designed an alien character that guides the learner's space travel by helping them solve algebra problems. This alien gives feedback and hint messages in response to students' problem-solving performance and requests (Fig. 1).

Sound Effects. We also implemented sounds in the system, another popular element in many tools that are designed to enhance student engagement [19]. Specifically, we added voice sounds for the alien and sounds for the drag-and-drop interactions. For example, when a learner finishes a problem, the alien celebrates it with the sound, "ta da!" while it says "hmm" when a learner makes an error. After user study sessions with students and math teachers, to accommodate their preferences and potential simultaneous use in classroom settings, we added the option for learners to turn the sounds on and off at any time (Fig. 1, right).

3 Classroom Experiment

To experimentally examine the effect Gwynnette on student learning and engagement, we conducted a classroom experiment at two public secondary schools in the U.S. Our research questions asked if Gwynnette, which was deliberately designed to enhance student learning and engagement in algebra, would (RQ1) improve students' conceptual and procedural learning in algebra (i.e., effect on learning) and (RQ2) help students engage with the software (i.e., effect on engagement). We also investigated (RQ3) how the design elements in the software affect students' interactions and learning processes (i.e., effect on learning processes).

3.1 Method

Participants. Twenty 6th, 55 7th, and 19 8th grade students across five classes in two public schools from two different school districts in the U.S. participated (total $n = 94$). Six of the 55 7th graders were also enrolled in the school's "Math Support" class, where students received additional instruction towards the goal of meeting state standards. The study happened in 2020, during which both schools were operating under a remote synchronous instructional mode due to the COVID-19 pandemic (i.e., teachers and students had synchronous classes via a videoconferencing system [20]). Students joined from their home using their own devices or school-provided devices (laptop or tablet).

Materials. To measure student learning of domain knowledge and skills in early algebra, we developed a web-based pretest and posttest based on items in the literature [9]. Each test included seven conceptual knowledge items (CK) and five procedural knowledge items (PK). CK items asked students to provide conceptual reasoning in solving algebra problems through multiple-choice or open-ended questions. PK items asked students to solve equations in an open-ended format (i.e., students typed in their answer and solution steps in a blank box). Two isomorphic versions were created and assigned to students in a counterbalanced way as the pretest and posttest.

Also, we measured students' self-reported engagement and enjoyment with the system using Intrinsic Motivation Inventory (IMI) [21]. IMI is a validated survey instrument developed to measure subjective experiences related to a psychological intervention. We only used the items related to enjoyment and engagement, which consisted of seven 7-point survey items [7]. Additionally, we also measured how long students used the system as a behavioral indicator of engagement [22].

The students in the study used two versions of intelligent algebra learning software, namely, a *playful* version (i.e., Gwynnette) and an *unplayful* version (i.e., Lynnette). These versions share the same algebra content and the core tutoring functionality; they both guide students through step-by-step problem solving and provide on-demand hints and feedback. Lynnette does not have the unique features of Gwynnette that were presented earlier. While Gwynnette allows students to solve equations through drag-and-drop manipulations, Lynnette has students type in problem-solving steps (i.e., transformed equations) into input fields (Fig. 4). Due to this difference, Lynnette software also asks students to perform the arithmetic computations involved in solving equations. In Gwynnette, computations are performed by the system. Additionally, Lynnette also allows students to skip intermediate problem-solving steps, which is not available in Gwynnette as its interaction design emphasizes the step-by-step problem solving that involves conceptual understanding. Other features were kept consistent across the two systems. In the study, we assigned the same sets of problems of 14 levels in both software versions (Table 1). Each problem level (Levels 1-10) has 2-4 problems. Level 11-14 included 8-15 problems with varying difficulty levels taken from Levels 1-10 (for advanced students). Teachers shared that most students have seen or practiced only Levels 1-4, but several advanced students had seen all levels before the study.

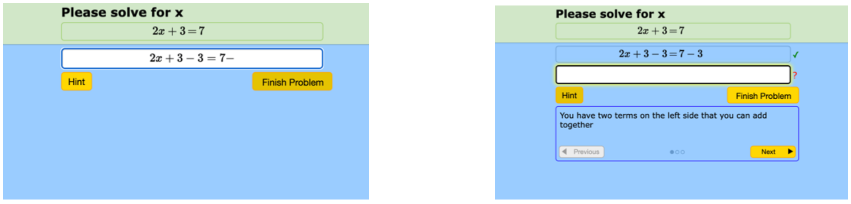

Fig. 4. To solve problems in Lynnette, users type in solution steps. (left) Users can click on the "Hint" button to request hints anytime (right).

Table 1. Problem levels, types of equation problems, and examples implemented in both Gwynnette and Lynnette.

Level	Equation Type	Example	Level	Equation Type	Example
1	One step	$x + 3 = 5$	8	Variables on both sides	$2x + 6 = 3x$
2	Two step (negative)	$6 - x = 3$	9	Variables and constants on both sides	$4x + 11 = x + 2$
3	One step (division)	$2x = 6$	10	Variables and constants on both sides (negative)	$-2x + 2 = -5x + 8$
4	Two step (division)	$2x + 3 = 7$	11	Mix	
5	Simplify first	$-3x + 5 - 3 = 14$	12	Mix	
6	With parentheses	$2(2x + 1) = 6$	13	Mix	
7	Parentheses, more steps	$1 + 2(2x - 1) = 7$	14	Mix	

Procedure. The study took place during four regular virtual mathematics class periods in each school. Experimenters joined the classes remotely through a video conferencing system. Students were also encouraged to use the assigned software version (see the assignment below) outside of the regular class periods. We allowed for such unmoderated use of the software in order to accommodate students' various needs during their remote learning and teachers' requests [20]. Students in each class were randomly assigned to either Gwynnette condition ($n = 47$) or Lynnette condition ($n = 47$). Students in each condition worked with the respective software version. Six students in the Math Support class were pre-identified and randomly assigned to the groups among them.

Students first worked on the pretest for 20 minutes. Students then watched a brief video on how to use both systems. In each of the second and third periods, students spent 20-30 minutes using the assigned software (the total study session time given in both conditions was about 50 minutes). On the final day, students took the posttest and the IMI survey. Students were given access to both versions after the study (data logging had stopped before we gave access to the tutor versions).

3.2 Results

Of the 94 students, 60 students completed all study activities (32 in the Gwynnette condition, 28 in the Lynnette condition). The high attrition rate was expected given the difficulty of conducting the study remotely. For example, teachers had to support students through a videoconferencing system and were not able to walk around the classroom to support students when necessary, which is typically done in in-person studies [20]. We included the 60 students in the final sample for the analyses. No statistically significant difference was found between the conditions on the dropout rate, X^2 (1, N = 94) = .74, p = .39.

Results on Learning. All student responses for the open-ended questions were coded for whether student answers were correct or incorrect by two researchers (*Cohen's kappa* = .86) and disagreements were resolved through discussions. Table 2 shows students' mean pretest and posttest scores. To test RQ1 (i.e., effect on learning), we conducted two separate linear regressions, with conceptual knowledge posttest score (CK) and procedural knowledge posttest score (PK) as the dependent variable, respectively. In both models, condition (Gwynnette or Lynnette, coded as 1 or 0) served as a predictor, and pretest score was added as a covariate to control for students' incoming knowledge. Results showed that students in the Gwynnette condition significantly outperformed those in the Lynnette condition on CK (β = .78, t(57) = 2.10, p = .04) but no difference was found on PK (β = .45, t(57) = 1.08, p = .28) (Fig. 5, left). Across the conditions, there was a significant pretest-to-posttest gain on CK (β = 2.02, t(57) = 5.43, p < .01) but not on PK (β = .59, t(57) = 1.42, p = .16).

Table 2. Mean pretest and posttest scores (standard deviations) in each condition.

Condition	CK (*max* = 7)		PK (*max* = 5)	
	Pretest	Posttest	Pretest	Posttest
Gwynnette	2.94 (1.44)	3.56 (1.74)	1.25 (1.52)	2.12 (1.84)
Lynnette	2.61 (1.81)	2.71 (1.80)	1.39 (1.83)	1.61 (1.97)

Results on Engagement. For RQ2 (i.e., effect on engagement), we analyzed students' ratings from the IMI survey. We took the mean score from the seven items for each student (*range:* 1-7). The mean score in the Gwynnette condition was 5.20 (SD = 1.14), whereas it was 4.80 (SD = 1.26) in the Lynnette condition. A Welch two-sample t-test showed that this difference is not statistically significant, t(55.12) = 1.28, p = .21.

To further understand student engagement, we also measured students' total time spent working on the system, a behavioral indicator for engagement [22]. In this study, the time spent working on the software was not controlled; students were encouraged to use the software outside the class periods. Therefore, a longer period of system use indicates higher engagement (i.e., the student used the system outside the class periods)

On average, students spent 54.03 minutes (*SD* = 34.47) with Gwynnette while those with Lynnette spent 27.98 minutes (*SD* = 22.69), about a half the amount of time as the Gwynnette condition (Fig. 5, right). A Welch two-sample t-test revealed that this difference was statistically significant, $t(53.98) = 3.50$, $p < .01$.

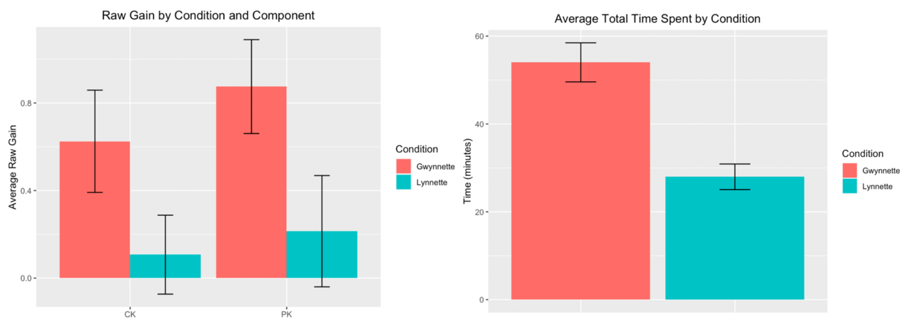

Fig. 5. Students' raw gain scores on the two test components (conceptual and procedural knowledge, left) and total time spent on the system (right). Error bars show standard errors.

System Log Data Analysis. To examine RQ3 (i.e., effect on learning processes), we analyzed the log data gathered from the software. Specifically, we investigated students' learning process measures, including average time spent per problem, average number of hints used per problem, and error rate (i.e., proportion of incorrect attempts for each problem-solving attempt in the system). Table 3 shows descriptive statistics for these measures. We conducted three separate linear regressions, with one of the three process measures as the dependent variable in each model, and condition and prior knowledge as predictors for all the models. These models revealed that students with Gwynnette had a significantly lower error rate ($\beta = -0.17$, $t(57) = -3.87$, $p < .01$) but spent more time on each problem ($\beta = 39.15$, $t(57) = 2.60$, $p = .01$), compared to those with Lynnette. No significant difference was found on the frequency of hint use. Also, students with Gwynnette solved significantly more problems ($M = 25.8$ problems, $SD = 14.4$) than those with Lynnette ($M = 18.1$ problems, $SD = 10.7$) ($\beta = 7.48$, $t(57) = 2.32$, $p = .02$). This difference is likely due to the significantly more time spent by students who learned with Gwynnette. Indeed, the number of problems solved and total time spent were strongly correlated ($r = .81$, $p < .01$).

Table 3. Learning process measures (standard deviations in parentheses) in each condition.

Condition	Error rate	Time spent on each problem (sec.)	Hint requests per problem
Gwynnette	0.29 (0.20)	136.0 (61.4)	1.83 (2.52)
Lynnette	0.47 (0.22)	98.5 (66.2)	1.89 (2.74)

What Design Features Might Explain the Observed Differences? The results above raise a question, "what made Gwynnette more effective and engaging?" Although it is not possible to tease apart causal effects of each design feature, we investigated two additional questions to get further insights into the potential mechanism involved: 1)

How do students' interaction patterns in each system relate to their learning outcomes? and 2) *Are there any pain points in the system that students struggle with?*

First, we took a closer look at the dataset in each condition separately 1) to examine behaviors that are related to the distinctive features in each system. For Gwynnette, one of the deliberate design choices was to encourage the learning of *doing the same thing to both sides of an equation*, important conceptual knowledge [9] by not allowing students to type in a number in the blank box before creating another blank box on the other side of the equation (Fig. 3). To examine if this interaction design might have something to do with students' conceptual learning, we calculated the number of instances where students tried to type in a number in the added box before doing the same thing to both sides of an equation (to which the system gave feedback saying, e.g., "You need to drag the plus sign to the other side of the equation before choosing what to add"). We found that, of the 32 students in the Gwynnette condition, all but two at least once tried to type in a number before dragging an operator to both sides of an equation. On average, this action was performed 4.86 times (SD = 4.46) per student. The number of times this action was performed was strongly, negatively correlated with students' pretest score on CK items (r = -.35, p = .048); however, there was no significant relations between this action and their conceptual knowledge posttests score nor the gain from pretest to posttest on students' conceptual knowledge. In other words, students with high prior knowledge of algebra concepts were more likely to avoid such a behavior than those with lower prior knowledge, but their performance with this interaction in the system does not predict their learning.

For the dataset from the Lynnette condition, we examined if students showed a "guess-and-check" behavior [23] by calculating the number of instances where students, for their first attempt, typed in "x = [a number]" (e.g., "x = 3") without showing any intermediate steps (Gwynnette requires step-by-step solutions and preempts guess-and-check strategies). The "guess-and-check" behavior is considered an informal, unideal strategy in solving algebra problems [23]. We investigated this behavior because it could indicate lower conceptual learning and lower engagement with the system (i.e., "gaming the system" behavior [24]). Of the 28 students in the Lynnette condition, 20 students showed the behavior at least once, with the average number of times being 15.8 (SD = 18.5) per student. Correlational analyses showed no relationships between the number of times students used guess-and-check and their conceptual pretest score, conceptual posttest score, and pretest-posttest gain score on conceptual knowledge. However, we found that students who used this strategy tended to spend less time with the system (r = -.41, p = .03), suggesting that students who used this strategy tended not to engage with the system.

Next, we explored 2) how the interaction patterns changed over time, to examine any pain points in the system that students encountered. Fig. 6 (left) shows a visualization of distributions in error rate for each problem set as side-by-side boxplots. One can expect that, within each problem set, learners would normally show some errors but subsequently errors would decline over time as they make progress in the system. As can be observed, students in the Gwynnette condition show smaller variance overall, especially after the first two levels. However, the error rate in the Lynnette condition generally shows a greater variance, indicating that some students in the Lynnette condition solved problems with very few errors while others in the same condition made many incorrect attempts. This may be an indication that, while most of the students who learned with Gwynnette were able to quickly learn how to use the drag-and-drop interactions, a new way to solve equations, after some practice, some students in the Lynnette condition did not become fluent in using the "type-in" interaction to solve equations, even after practice. Furthermore, students in the Lynnette condition had greater difficulty compared to those with Gwynnette especially in levels 5, 8, and 10 (Fig. 6, right). These levels are where the *simplify-before-transform* problems (i.e., students first need to subtract/add constant or variable terms before starting to use operators to transform equations), problems with variable terms on both sides of an equation, and problems with negative numbers in a complex equation format were introduced, respectively (Table 1). These factors (e.g., negative numbers and variable terms in "unusual places") are reported to have a strong influence on students' problem-solving performance [25].

Fig. 6. Distributions of error rates across problem levels (left). Means of error rates across problem levels, plotted as a line chart (right). In both graphs, patterns seen across Levels 12-14 do not inform consistent insights as there were only 9 students in total who reached Level 12 problems (Gwynnette: $n = 7$, Lynnette: $n = 2$).

4 Discussion and Conclusion

We tested whether the deliberate design of playful features in algebra learning software can help enhance student learning and engagement through a controlled classroom experiment. Gwynnette had several playful interaction and gamification features, including drag-and-drop interactions, a space theme, an alien guide, and sounds. Lynnette had no such features; instead of transforming equations through dragging and dropping, students typed in transformed equations. The results of the experiment showed that students who used Gwynnette learned more conceptual knowledge in algebra than those with Lynnette. Students with Gwynnette also showed greater problem-solving efficiency, demonstrated by the overall lower error rate in the system. As well, students engaged significantly more with Gwynnette than with Lynnette, as measured by the total time spent working with the system. There was no statistically significant difference between the conditions in students' ratings on the IMI survey, therefore we cannot fully establish the effect on engagement. However, their ratings correlated positively with the total time spent working on the system (r = .33, p <.01), and we view the behavioral measure as more compelling evidence than students' self-report. Despite the extended use of the software, however, students with Gwynnette did not outperform those with Lynnette on procedural knowledge. This result suggests less effective procedural learning, possibly due to the new format for solving equations on the interface (i.e., their practice did not transfer to the performance on the posttest).

It is interesting to ask how the design features may have contributed to improved engagement and conceptual learning. Although our study design does not allow us to attribute any outcomes to specific design elements, we can make somewhat speculative inferences based on the design of the systems and findings. For instance, the fact that students with Gwynnette had a lower error rate suggests that the playful drag-and-drop interactions brought about a smooth learning experience. It may also be that the playful features might have led to greater enjoyment, resulting in longer use of the system, hence greater learning gains. As well, the drag-and-drop interactions in Gwynnette perhaps allowed students to focus on the transformations, rather than dealing with arithmetic calculations. This focus may have led to greater conceptual learning. Also, we found that students' explicit action of *doing the same thing to both sides of an equation* had a positive association with students' conceptual knowledge on the pretest, which disappeared on the posttest. This might imply that students' interactions with (and learning from) this deliberate aspect of the drag-and-drop design, rather than whether they performed well or not with the specific action, had a positive influence on conceptual learning. Students with Lynnette, on the other hand, had to perform arithmetic calculations when solving equations, which may have contributed to the greater difficulty that students in the Lynnette condition experienced for new problem types. Also, some students with Lynnette tended to skip problem-solving steps using the "guess-and-check" strategy. Although those who used "guess-and-check" frequently spent less time on each problem (r = -.63, p < .01), they also did not finish all problem and rather tended to stop using it early (r = -.41, p = .03), indicating lower engagement.

The current study makes a design contribution that the deliberate design of playful features in a learning technology can help achieve the challenging goal of supporting both learning and engagement in algebra. The features helped students engage with algebra, a notoriously unpopular subject among students [6], with double the time spent. Practically, the study offers an example that designing for a playful learning experience with learning technology can support students' remote learning during difficult times (e.g., a pandemic). Studies report that it is highly challenging for students to engage with school work during remote learning [20, 26]. By allowing for unmoderated system use outside of the class time, we found that students with Gwynnette spent twice as much time (and learned more).

We acknowledge several limitations of the study. First, the study tested very specific design features in the domain of early algebra. We do not know if the findings will generalize across domains. Some of the added features were domain-independent (e.g., alien guide, space theme, and sounds), but at least one of them (drag-and-drop interactions for solving equations) may not be. Second, the study was conducted during the COVID-19 pandemic where students, teachers, and experimenters were all connected virtually. It is possible that the findings would have looked different if the study had been conducted in the in-person classroom where teachers were able to help students as they would have done *normally*. Finally, we cannot attribute the results to specific features in the system. Future studies could experimentally test the question.

5 Acknowledgements

This research was supported by NSF Award #1760922. We thank Martha W. Alibali, Max Benson, Jenny Yun-Chen Chan, Octav Popescu, Jonathan Sewall, and all the participating teachers and students.

References

1. Nguyen, H., Harpstead, E., Wang, Y., McLaren, B.M.: Student agency and game-based learning: A study comparing low and high agency. In: International Conference on Artificial Intelligence in Education. pp. 338–351. Springer (2018).
2. Spielhagen, F.R.: Closing the achievement gap in math: The long-term effects of eighth-grade algebra. Journal of Advanced Academics. 18, 34–59 (2006).
3. Rittle-Johnson, B., Siegler, R.S.: The relation between conceptual and procedural knowledge in learning mathematics: A review. The Development of Mathematical Skills. 338, 75–110 (1998).
4. Crooks, N.M., Alibali, M.W.: Defining and measuring conceptual knowledge in mathematics. Dev. Rev. 34, 344–377 (2014).
5. McNeil, N.M., Alibali, M.W.: Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. Child Dev. 76, 883–899 (2005).
6. Norton, S., Irvin, J.: Developing positive attitudes towards algebra. In: Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia. pp. 561–570. (2007).
7. Long, Y., Aleven, V.: Educational game and Intelligent Tutoring System: A classroom study and comparative design analysis. ACM Trans. Comput.-Hum. Interact. 24, 1–27 (2017).
8. Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of Cognitive Tutor Algebra I at scale. Educ. Eval. Policy Anal. 36, 127–144 (2014).
9. Nagashima, T., Bartel, A., Silla, E., Vest, N., Alibali, M.W., Aleven, V.: Enhancing conceptual knowledge in early algebra through scaffolding diagrammatic self-explanation. In: Gresalfi, M. and Horn, I.S. (eds.) Proceedings of the 14th International Conference of the Learning Sciences. pp. 35–43. International Society of the Learning Sciences (2020).
10. Jackson, G.T., McNamara, D.S.: Motivation and performance in a game-based intelligent tutoring system. J. Educ. Psychol. 105, 1036–1049 (2013).
11. Siew, N.M., Geofrey, J., Lee, B.N.: 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, 66–79 (2016).

12. Chan, J.Y.-C., Lee, J.-E., Mason, C.A., Sawrey, K., Ottmar, E.: From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. J. Educ. Psychol. 114, 56–71 (2021).
13. Ottmar, E., Landy, D., Goldstone, R.: Teaching the perceptual structure of algebraic expressions: Preliminary findings from the pushing symbols intervention. In: Proceedings of the Annual Meeting of the Cognitive Science Society. (2012).
14. Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., Ottmar, E.: From Here to There! Elementary: a game-based approach to developing number sense and early algebraic understanding. Educ. Technol. Res. Dev. 67, 423–441 (2019).
15. Schneider, B., Jermann, P., Zufferey, G., Dillenbourg, P.: Benefits of a tangible interface for collaborative learning and interaction. IEEE Trans. Learn. Technol. 4, 222–232 (2011).
16. Ruan, S., He, J., Ying, R., Burkle, J., Hakim, D., Wang, A., Yin, Y., Zhou, L., Xu, Q., AbuHashem, A., Dietz, G., Murnane, E.L., Brunskill, E., Landay, J.A.: Supporting children's math learning with feedback-augmented narrative technology. In: Proceedings of the Interaction Design and Children Conference. pp. 567–580. ACM, New York, NY, USA (2020).
17. Xi, N., Hamari, J.: Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. Int. J. Inf. Manage. (2019).
18. Kim, J.T., Lee, W.H.: Dynamical model for gamification of learning (DMGL). Multimed. Tools Appl. 74, 8483–8493 (2015).
19. Nagashima, T., Yadav, G., Aleven, V.: A framework to guide educational technology studies in the evolving classroom research environment. In: Proceedings of the European Conference on Technology Enhanced Learning. pp. 207–220. Springer International Publishing (2021).
20. McAuley, E., Duncan, T., Tammen, V.V.: Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. Res. Q. Exerc. Sport. 60, 48–58 (1989).
21. Marwan, S., Price, T.W., Chi, M., Barnes, T.: Immediate data-driven positive feedback increases engagement on programming homework for novices. In: CSEDM@EDM. (2020).
22. Koedinger, K.R., Alibali, M.W., Nathan, M.J.: Trade-offs between grounded and abstract representations: evidence from algebra problem solving. Cogn. Sci. 32, 366–397 (2008).
23. Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K.: Why students engage in "gaming the system" behavior in interactive learning environments. J. Interact. Learn. Res. 19, 185–224 (2008).
24. Long, Y., Holstein, K., Aleven, V.: What exactly do students learn when they practice equation solving?: refining knowledge components with the additive factors model. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge. pp. 399–408. ACM, Sydney, New South Wales, Australia (2018).
25. Stelitano, L., Doan, S., Woo, A., Diliberti, M., Kaufman, J.H., Henry, D.: The digital divide and COVID-19: Teachers' perceptions of inequities in students' internet access and participation in remote learning. Data note: Insights from the American Educator Panels. Research Report. RAND Corporation. (2020).