Evaluation of the

Lure of the Labyrinth™

Challenge

Lindsay Demers, Ph.D.

Jonathan Christiansen

TERC

The Evaluation of the Lure of the Labyrinth Challenge

2067 Massachusetts Avenue, Cambridge, MA 02140
(617) 873-9600

Table of Contents

Executive Summary

The *Lure of the Labyrinth* is a math game developed by MIT's Education Arcade, designed to improve Common Core proficiency and *deeper learning* skills in middle-school-aged students. With regard to Common Core math skills, the fictional world in which the game takes place is divided into three sections. Each of these three sections is associated with a different pre-algebra strand including Proportions, Variables & Equations, and Number & Operations. Within each of these sections there are multiple levels, each of which contains three increasingly difficult puzzles that students must complete in order to progress through the game. The puzzles are presented to students without any explicit instructions and students must experiment to figure out how to solve them. Thus, while students are never explicitly told that they are doing "math" during game play, in order to progress through the *Lure of the Labyrinth*, they must use math skills aligned with the Common Core standards.

In addition to improving Common Core math skills, the *Lure of the Labyrinth* also seeks to encourage *deeper learning* skills through the use of a messaging system, which is designed to encourage teammates to correspond with one another and strategize about the game. The *deeper learning* skills that the messaging system is meant to facilitate include 1.) *Problem solving,* specifically the ability to identify problems, and identify different strategies for coming to a solution; 2.) *Effective communication*, which includes the ability to clearly articulate problems, issues, and ideas to a wide audience; 3.) *Collaboration*, which can be seen when students work together to solve problems that they could not solve on their own, or when one student shares knowledge or skills with another; and 4.) *Learning to learn*, which includes the willingness to ask for help and the ability to identify problems as well the ability to ask the appropriate types of questions and to accept help and constructive feedback

To introduce the *Lure of the Labyrinth* to a variety of teachers and students across the country, the *Lure of the Labyrinth* Challenge was launched on April 4th and lasted until June 15th, 2012. Middle-school teachers from a variety of settings (including public schools, private and charter schools, afterschool programs, and library programs) were eligible to register teams comprised of 4 to 6 students for the Challenge. Throughout the duration of the Challenge, data were gathered from participating teachers and students to assess teacher attitudes toward technology in the mathematics classroom, student and teacher use of the in-game message board system and student performance within the game. Using these data, the evaluation of the *Lure of the Labyrinth* Challenge was conducted by the STEM Education Evaluation Center (SEEC) at TERC, a non-profit research and development organization in Cambridge, MA. This evaluation provides summative findings regarding how teachers implemented the *Lure of the Labyrinth* Challenge in the classroom and how students played game. Although teachers from a variety of settings could register teams for the Challenge, in order to understand how the *Lure of the Labyrinth* worked in more traditional classrooms, only data from formal educational settings (i.e., public, private, and charter schools) were collected and analyzed for the current report. Below, the primary research questions that drove the evaluation are listed, along with the conclusions drawn from the data.

Did participation in the *Lure of the Labyrinth* Challenge change teachers' attitudes toward technology in the classroom?

To examine change in teachers' attitudes toward technology in the math classroom an optional survey was available to participating teachers before and after their participation in the

Challenge. Approximately 213 teachers completed the pre- survey, while just under 60 teachers completed the post- survey. Of those teachers, only 39 responded to both the pre- and post- survey. Given this limited response rate, we were unable to conduct inferential tests to look for statistical change in teachers' survey responses over time. However, it is clear from descriptively looking at the pre- and post- data that the teachers who participated in the *Lure of the Labyrinth* Challenge were very interested in and comfortable with incorporating technology into their classrooms. In fact, several of the teachers who were interviewed indicated that they were computer lab teachers. It is clear that games such as the *Lure of the Labyrinth* are a welcome complement to ongoing mathematics classroom instruction.

How did teachers facilitate game play via the message boards? In what ways did teachers primarily use the message boards?

Of the 46,424 message board posts extracted, approximately 1% (n = 502) were posts authored by teachers. Of the teachers who did use the message boards, the average number of posts was 1.58 (SD = 1.48). All of these teacher posts were coded descriptively to provide information regarding how teachers used the message boards, based on emergent themes from the data.

The most common theme observed in the teacher posts was that of teachers using the message board to keep students on task. This included telling students to behave, reminding them of impending deadlines or tasks, and to encourage competitiveness with regard to the competition. This theme encompassed just over half (50.7%) of the codable teacher posts (i.e., excluding those posts that were un-interpretable). Second to reminding students to stay on task, the next most common theme observed in the data was that of teachers scaffolding or offering advice for students, which accounted for 38.8% of the codable teacher posts. This category included posts that asked students to consider how to solve a problem, and posts in which teachers explained the math required to solve the puzzle. The third most common theme that emerged from the teacher posts was that of teachers explaining how to do things in the game to students, which comprised 8.2% of the codable teacher posts. In these posts, teachers explained the "nuts- and-bolts" of how to play the game, but did not contain any math-related content. Similar to those posts were the posts in which teachers provided technical support to the students, which made up 2.3% of the codable teacher posts.

Overall, these data suggest that just under half (~46%) of teacher posts were attempts to facilitate student gameplay by either explaining how to accomplish something in the game or by posing questions to students in order to encourage them to think in ways that might help them make progress in the game. However, primarily, when teachers used the message boards, it was to tell students to post appropriately or simply to remind them to continue to make progress in the game. It is worth noting that although keeping students on task was the most common purpose for teachers' use of the message boards, that this activity was still very infrequent within the grand scheme of the game, as teachers posts comprised only 1% of the messages posted throughout the duration of the Challenge.

How did students use the message boards? Was there evidence of collaboration among teams? How did team message board use relate to in-game performance?

A sub-set of student teams was randomly selected to have their message board posts examined in detail (n = 361 teams, a total of 4,228 posts). In the first pass of coding, each of the student posts was coded as completed unrelated to game play, related to game play, but incidentally (e.g., comments about characters, attempts to identify team members) or as substantively related to game play (e.g., asking for help with some aspect of the game). A majority of the student posts were not related to game play and in general, posting on the message boards was not correlated with team performance in the game. However, there was a positive association between the quality of substantive posts (as indicated by post length) and team performance in the game, suggesting that while the message boards could be used for effective collaboration by students, that this use was not a necessary criterion for in-game success.

To gain a better understanding of how the message boards were used by students under ideal circumstances, all of the posts that were centrally related to the game were coded for finer detail. Each of these substantive posts (n = 1224) was coded for collaborative behaviors such as asking team members for help, sharing game progress, inquiring about others game progress, and explaining aspects of the game to other team members.

The most frequently observed substantive student post consisted of students asking each other for help (35% of posts). The substance of these requests for help varied greatly with some simply being the word "help" and others being much more detailed and elaborate. However, regardless of the detail included in the post, the underlying sentiment collaboration was present. The second most common use of the message boards for issues that were centrally related to game play was students sharing their progress in the game with their teammates (29% of posts). For many students, sharing their progress was an important part of the game, and a way in which they could interact with their friends or classmates. Sharing progress helped the students share their accomplishments, but it also allowed for updates, and allowed students to know who to ask if they needed help. Next most prevalent in the message boards, was students explaining something about the game to their teammates (14% of posts). These posts were typically (though not always) made in response to someone asking for help. Some of these posts were highly descriptive posts of steps or processes while others were shorter, with students simply listing a word or item to be used to overcome a particular obstacle. The least frequent use of the message boards was students asking their teammates about how far they had progressed in the game (12% of posts).

Overall, there was some evidence of *deeper learning* skills in the message boards, including evidence of students using the message boards to collaborate in various ways such as showing a willingness to ask questions and share progress, hints and strategies with teammates. However, this substantive use of the message boards was very infrequent within the larger scheme of the game, with teams only posting just over 3 substantive messages on average. Along that vein, although the quality (i.e., length) of student posts that were substantively related to game play was positively associated with in-game performance at the team level, message board use generally was not a significant predictor of in-game performance. However, because students had myriad ways to communicate with one another (e.g., face-to-face, cell phone text message, Skype) it is likely that examination of the message board posts provided only a minimal depiction of the extent of collaboration that went on among teams throughout the duration of the Challenge.

How engaged were the students with the game over time? On which levels did students persist the most?

To examine students' persistence and engagement with the game, we looked at the amount of time students' spent playing each puzzle, the average number of puzzles students completed, the number of students who beat each puzzle, and the final puzzle that students beat before they stopped progressing through the game. With regard to time spent with the puzzles, on average, students spent 3.85 minutes (*SD* = 5.56) on attempts in which the puzzle was not beaten and 4.21 minutes (*SD* = 5.62) on attempts in which the puzzle was beaten. Using time spent on successful attempts as a proxy measure for puzzle difficulty, these data suggest that the Executive Lounge was the most challenging, with an average completion time of 11.95 minutes (*SD* = 12.11) while the Employee Lounge was the easiest, with an average completion time of 2.35 minutes (*SD* = 3.05). Further, the data indicated a linear increase in the amount of time taken to complete each puzzle as the puzzles became more difficult within a given level. This shows that students were willing and able to put in additional effort when the game required it and also serves to show that the puzzles did increase in difficulty within a level, as intended by game developers. With regard to the engagement with the game, on average, students completed 7.48 puzzles (*SD* = 11.94). Although there was a great deal of individual difference in the amount that students progressed through the game, the patterns observed in the data suggest that engagement may be bi-modally distributed, with some students dropping off primarily during levels that are unlocked earlier in the game and other students persisting through to levels unlocked later in the game.

Did students show evidence of learning across consecutive puzzle attempts? What patterns of learning emerged from the data?

Student learning within the game was examined for three levels: The Cafeteria (comprised of the Employee Cafeteria, the Manager's Cafeteria, and the Executive Cafeteria), The Recipe Room (comprised of Testing Lab 1, Testing Lab 2, and Testing Lab 3) and the Vending Machine (comprised of the Employee Lounge, the Manager's Lounge, and the Executive Lounge). In order to gauge learning over time, we looked at the proportion of correct moves each student made on students' first five attempts at the three puzzles contained within each level. Overall, results showed that when presented with a puzzle of minimal difficulty, students were able to make consistent progress as they made repeated attempts. On the puzzles of medium difficulty, student performance tended to be fairly consistent over subsequent attempts. Finally, on the most difficult puzzles, student performance tended to decline with subsequent attempts, perhaps due to frustration (e.g., in cases where students were unable to solve the puzzle). Although these data cannot be used to draw general conclusions regarding student learning, their intention was to be used by game developers to identify areas of relative difficulty within specific puzzles.

What characteristics did the top-performing teams possess? How did the teachers who supervised those teams implement the game?

In order to provide a more comprehensive picture than the game play data could reveal, we decided to interview a sub-set of teachers who supervised highly successful teams in order to better understand what these teams may have had in common. The first thing that became clear from the interviews was that there was not a model set of characteristics or behaviors that emerged. The teachers all seemed to be quite different in their approach to using the *Lure of the Labyrinth*: some teachers were computer lab teachers rather than math teachers; one of the schools was a private school for gifted students, one was a math and science academy, while others were public schools; some teachers chose to emphasize the competitive aspect of the game more than others. There was also substantial variability in the way that teachers decided to introduce the game to their students: some teachers interviewed were very thorough in their preparation for the game, while others took a more *laissez-faire* approach; some were quite familiar with the game to begin with, and had in fact played it before, while others were new to the game and unclear about the Challenge.

Despite the many different ways the game was integrated into the classroom by the teachers, there were some consistencies across those varied settings. Foremost of these similarities, was that all of the teachers were enthusiastic about the game, and many had played before or plan to continue to use it. Further, these teachers seemed to foster a classroom culture in which there was a fairly open relationship between students and teachers, with teachers being willing to explain things to the students, to work alongside the students to solve puzzles, and to ask them questions about how to solve the puzzles, perhaps providing students with feelings of mastery and expertise that increased their interest in the game. This suggests that regardless of the way that teachers implement the game, that the more likely key to success is having teachers who are enthusiastic about the game and willing to use it as an opportunity for learning and collaboration *with* their students.

Overview of the Lure of the Labyrinth

Rationale

The *Lure of the Labyrinth* is a math game developed by MIT's Education Arcade designed to improve Common Core proficiency and *deeper learning* skills in middle-school-aged students. The learning theory behind the game is "preparation for future learning" which states that hands-on experience prior to classroom instruction leads to better learning outcomes in students[1]. Thus, under ideal conditions, teachers would encourage students to play the game and then follow-up with classroom instruction.

With regard to Common Core math skills, the fictional world in which the game takes place is divided into three sections. Each of these three sections is associated with a different pre-algebra strand including Proportions, Variables & Equations, and Number & Operations. Within each of these sections there are multiple levels, each of which contains three increasingly difficult puzzles that students must complete in order to progress through the game. The puzzles are presented to students without any explicit instructions and students must experiment to figure out how to solve them. Thus, while students are never explicitly told that they are doing "math" during game play, in order to progress through the *Lure of the Labyrinth*, they must use math skills aligned with the Common Core standards.

In addition to improving Common Core math skills, the *Lure of the Labyrinth* also seeks to encourage *deeper learning* skills through the use of a messaging system, which is designed to encourage teammates to correspond with one another and strategize about the game. The *deeper learning* skills that the messaging system is meant to facilitate include 1.) *Problem solving*, specifically the ability to identify problems, and identify different strategies for coming to a solution; 2.) *Effective communication*, which includes the ability to clearly articulate problems, issues, and ideas to a wide audience; 3.) *Collaboration*, which can be seen when students work together to solve problems that they could not solve on their own, or when one student shares knowledge or skills with another; and 4.) *Learning to learn*, which includes the willingness to ask for help and the ability to identify problems as well the ability to ask the appropriate types of questions and to accept help and constructive feedback.[2]

Game Play

The Lure of the Labyrinth unfolds as a comic-style story in which math- and logic-based puzzles are embedded. The story follows a character whose pet has been kidnapped. In order to retrieve the pet, the character must follow the presumed kidnapper down into the sewer system. From there, the character is taken on a wild adventure through the sewers, which incidentally also house a pet food factory that is entirely staffed by monsters. In order to infiltrate this strange world and save his pet, the character is transformed into a monster himself. Playing as the character, students must work their way through a series of jobs (puzzles) at the monster pet food factory in order to find the missing pet. In exchange for correctly solving puzzles, students are rewarded with coins that they may use to purchase devices from the shop. Using these devices, students can free other pets being kept in the factory. If a puzzle is not solved correctly, students are not rewarded with the maximum amount of coins for the puzzle and may not be able

[1] Martin, T., & Schwartz, D. L. (2004). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts.
[2] Reid, J. (1992) has also shown that cooperative learning styles such as those that would be expected from the team play used in *Lure of the Labyrinth Challenge* can lead to higher math achievement as well as increased self-esteem and social acceptance.

to free as many pets. After all of the puzzles have been completed, the pet food factory is blown up and the character's pet is freed.

Figure 1. Screenshot from *The Lure of the Labyrinth*

<u>Overview of the Lure of the Labyrinth Challenge</u>

The *Lure of the Labyrinth* Challenge was launched on April 4th and lasted until June 15th, 2012. The challenge was designed to introduce *The Lure of the Labyrinth* to a wide range of schools and classrooms by offering teachers and students the chance to win prizes for playing the game. Middle-school teachers from a variety of settings (including public schools, private and charter schools, afterschool programs, and library programs) were eligible to register teams comprised of 4 to 6 students for the Challenge. Students' correspondence with one another through the messaging system, as well as the other student gameplay information, was visible to the overseeing teacher throughout the duration of the Challenge. Partners in the Challenge included Next Generation Learning Challenges, Maryland Public Television, Fablevision Learning, Lenovo and Brain Pop. Research on the *Lure of the Labyrinth* was funded by Educause through the Next Generation Learning Challenges.

Overview of Evaluation

Evaluation of the *Lure of the Labyrinth* Challenge was conducted by the STEM Education Evaluation Center (SEEC) at TERC, a non-profit research and development organization in Cambridge, MA. This evaluation provides summative findings regarding how teachers implemented the *Lure of the Labyrinth* Challenge in the classroom and how students played game. Although teachers from a variety of settings could register teams for the Challenge, in order to understand how the *Lure of the Labyrinth* worked in more traditional classrooms, only data from formal educational settings (i.e., public, private, and charter schools) were collected and analyzed for the current report. The following research questions guided this work:

1.) Did participation in the *Lure of the Labyrinth* Challenge change teachers' attitudes toward technology in the classroom?

2.) How did teachers facilitate game play via the message boards? In what ways did teachers primarily use the message boards?

3.) How did students use the message boards? Was there evidence of collaboration among teams? How did team message board use relate to in-game performance?

4.) How engaged were the students with the game over time? On which levels did students persist the most?

5.) Did students show evidence of learning across consecutive puzzle attempts? What patterns of learning emerged from the data?

6.) What characteristics did the top-performing teams possess? How did the teachers who supervised those teams implement the game?

Description of Data

In order to address the research questions above, a variety of data was gathered. These data came from 13,410 students, across 3,804 teams and 701 teachers. The average size of each team was 3.53 students (*SD*[3] = 1.72) with a range of 1 to 6. A brief overview of each measure collected is provided below in Table 1. Additional detail regarding each measure (including coding procedures) can be found in the Findings section.

Table 1. Sources of Data for the Evaluation of the Lure of the Labyrinth Challenge

Data Source	Measure
Teachers	• Pre-/Post- Survey: All participating teachers were offered the opportunity to complete a survey regarding their attitudes toward technology in the classroom. This survey was adapted from an existing instrument that was developed and validated by Race (2003). The survey was administered electronically before and after participation in the Lure of the Labyrinth Challenge. Two-hundred-fourteen teachers (31% of the sample) completed the pre- survey, while only 59 completed the post- survey. Only 35 teachers completed both the pre- and the post- survey. • Message Board Posts: Of the 46,424 message board posts extracted, approximately 1% (*n* = 502) were posts authored by teachers. All of these posts were coded descriptively to provide information about how teachers used the message boards.

[3] The standard deviation (*SD*) is a measure of variability around the mean. The larger the standard deviation, the greater the spread of values in the data about the mean.

	• Interviews: Interviews were conducted with 8 teachers approximately 5 months after they completed the Lure of the Labyrinth Challenge. These teachers were selected purposefully based on the highest average number of pets freed per student team member (a measure of in-game success) and also so that an array of economic statuses was represented. The interviews were audiotaped and transcribed to allow for detailed analysis.
Students	• Message Board Posts: A sub-set of student teams (~10%) was randomly selected to have their message board posts examined in detail (*n* = 361 teams, a total of 4,228 posts). These posts were coded descriptively to provide information about how students used the message boards, with a specific eye toward identifying posts that exhibited collaboration within teams. • Game Play Data: Individual student game play data were tracked throughout the duration of the Lure of the Labyrinth Challenge. These data were sent to the research team at TERC from Fablevision (the game developers). In order to be included in this data file, students needed to have played at least one puzzle in the game and have been part of a team registered by a teacher in their school.

Findings

Below, each research question is reiterated followed by a brief summary of what the results of our coding and analysis suggested. Additional detail regarding how we arrived at those conclusions is provided in the body of the text.

Q: Did participation in the Lure of the Labyrinth Challenge change teachers' attitudes toward technology in the classroom?

A: *Because so few teachers completed both the pre- and post- survey, we cannot say for certain how participation in the Lure of the Labyrinth Challenge impacted teachers' views on technology in the classroom. However, it is clear even prior to participating in the Challenge that teachers were open to and saw the value of technology in the classroom.*

Only 39 teachers responded to both the pre- and post- surveys. As such, we were unable to conduct inferential tests to look for statistical change in teachers' survey responses over time. However, it is clear from looking at the descriptive statistics (Table 2), that the teachers who participated in the *Lure of the Labyrinth* Challenge were very interested in and comfortable with incorporating technology into their classrooms. In fact, several of the teachers who were interviewed indicated that they were computer lab teachers. Note that in Table 2, survey questions are organized such that the first half address issues regarding teaching philosophy whereas the second half focuses on the use of technology in the classroom.

Table 2. Descriptive Statistics for Teacher Attitude Survey (Response scale: 1 = Strongly disagree; 7 = Strongly agree)

Question	Pre- Mean (*SD*) (*n* = 210 to 213)	Post- Mean (*SD*) (*n* = 56 to 59)

1. Students learn best in mathematics when they are allowed to explore problems and test ideas about possible solutions.	6.32 (0.98)	6.36 (0.96)
2. Worthwhile mathematical tasks foster a connection between application and understanding.	6.50 (0.90)	6.64 (0.64)
3. A primary objective of mathematics is to develop the ability to identify and solve problems generated from real-life situations.	6.42 (0.96)	6.52 (0.82)
4. Problem solving can be facilitated by students working in groups.	6.21 (1.07)	6.28 (1.00)
5. All students are capable of understanding mathematics.	6.25 (1.15)	6.26 (0.97)
6. Instructional strategies I use as a teacher can increase a child's interest in a particular subject area**.**	6.42 (0.94)	6.56 (0.68)
7. Students learn best in mathematics through teacher explanations.	3.81 (1.42)	3.95 (1.46)
8. Students cannot be successful in mathematics until they have mastered computation skills.	4.28 (1.69)	3.72 (1.82)
9. Technology makes my professional work more difficult.*	1.90 (1.36)	1.75 (1.31)
10. Using computers for learning takes students away from important instructional time.*	1.97 (1.42)	1.95 (1.38)
11. Computers should be as important and available to students as pencils and books.	5.96 (1.57)	6.23 (1.02)
12. I am confident using technology as a learning resource.	6.03 (1.44)	6.39 (0.82)
13. I do not believe the quality of mathematics education is improved by the use of technology.*	2.01 (1.52)	1.74 (1.33)
14. There is not enough time to incorporate technology into the subjects I teach.*	2.68 (1.84)	2.61 (1.81)
15. I really enjoy using computers and the Internet instructionally.	6.13 (1.37)	5.95 (1.62)
16. Students should be able to use computers to help them solve math problems.	5.95 (1.39)	5.95 (1.14)

*indicates that a survey item was written such that a lower score is indicative of more positive feelings toward technology in the classroom.

Q: How did teachers facilitate game play via the message boards? In what ways did teachers primarily use the message boards?

A: *Very few teachers used the message boards. Those who did use them, did so sparingly. When teachers used the message boards, it was typically to tell students to post appropriately or to remind them to continue to make progress in the game. The second most common use of message boards by teachers was to facilitate student gameplay by either explaining how to accomplish something in the game or by posing questions to students in order to encourage them to think in ways that might help them make progress in the game.*

Of the 46,424 message board posts extracted, approximately 1% (*n* = 502) were posts authored by teachers. Of the teachers who did use the message boards, the average number of posts was 1.58 (*SD* = 1.48). All of these posts were coded descriptively to provide information regarding how teachers used the message boards. The following categories were developed, based on the following emergent themes from the data:

1.) Encourages questioning/Scaffolds
2.) Provides a "how-to" for game play
3.) Provides technical support for students
4.) Attempts to keep students on task (e.g., reminders about the game and/or the competition or admonishment for misuse of message boards)
5.) Other: post did not fit into any of the above-described categories, nor was there an emergent theme into which the posts could be categorized (many of these appeared to be test messages).

Approximately 20% of the teachers' posts were coded by two researchers to establish inter-rater reliability. Inter-rater reliability for the coding rubric was in the highly acceptable range (Cohen's kappa > 0.80).

Of the 502 teacher posts examined 40% (*n* = 198) were coded as "Other" because they were either gibberish or not substantive enough to categorize (e.g., a single word). The remaining 60% of posts (*n* = 304) were coded based on the emergent themes from the data outlined above. Given the dearth of teacher posts, no inferential statistics could be run to relate these data with student performance.

The most common theme that emerged from the teacher posts was that of teachers using the message board to keep students on task (#4, above). This included telling students to behave, to remind them of impending deadlines or tasks, and encourage competitiveness with regard to the competition. This theme encompassed just over half (50.7%) of the codable teacher posts (i.e., excluding posts that were coded as 'other'). The tone of these posts varied extensively, with some providing encouragement and positive feedback to students:

> *Be sure to communicate with your team members when you figure out how to win points and get keys!! Communication is part of the game!! Have fun!!!*

> *hey guys, it is gr8 2 message socially here, but remember 2 share ur progress & help ur team, 2 :)*

While others were just more straightforward reminders (e.g., to use their code names when addressing each other):

hey 40k...just read ur bio. Gr8 idea! Do me a fave & change the names 2 codenames...add in the letter v the 1st name 2 help jog ur memory ir u want..thnx! :)

Students sometimes used the message boards inappropriately and as a result, teachers occasionally interjected. There seemed to be some variability in what teachers found to be appropriate and inappropriate, though, with some teachers allowing a certain degree of social posts and others preferring that students only use them for game-related tasks.

I am glad you guys have learned how to use the messaging system. Remember to use this to help each other solve puzzles and uncover secrets in the game. I also want to remind you to always make sure you are posting appropriately. I can and will revoke your ability to post upon the first infraction.

DO NOT WRITE INAPPROPRIATE COMMENTS PLEASE! DO WORK!

Although a high percentage of student posts were unrelated to game play (described below), teachers infrequently messaged students to remind them that the message boards were for game play. However, when asked about this in the interviews, teachers who reported having problems with inappropriate message board use said that those students were addressed face-to-face, rather than through the message board system.

This code also encompassed teachers reminding students about the competition (e.g., mentioning prizes or remarking how well the team was doing in the competition). Examples of this type of post included:

You are getting behind. You need to work faster!

Hello red team this is Mr. Garrison. You are number 2 for the 6th grade for puzzle completions with 141 beans. Blue team has 153, and a 7th grade team earned 183. Your team has earned 14 entries into the drawing. The drawing will be held June 16th and if anyone from our school wins i will contact that team later that week. GREAT JOB i really appreciate all your hard work, have a great summer

This is Mrs. Madison, I received a report that stated 1st place for number of pets freed nationally is 1360. Our highest is 1210. Keep trying to get as many as possible.

Second to reminding students to stay on task, the next most common theme observed in the data was that of teachers scaffolding or offering advice for students (#1, above), which accounted for 38.8% of the codable teacher posts (i.e., excluding posts coded as ‘other’). This category included posts that asked students to consider how to solve a problem, and posts in which teachers explained the math required to solve the puzzle. These included things such as,

The Shipping Manifest is basically a case where you use equations to figure out what the symbols mean. You have a hint pallet that initially comes up. It's label *is an equation like 2+2=4. You use that label to begin to decode the meaning of the symbols.*

The map for the employee lounge is a position on a coordinate grid.

In level 2 of the testing lab, they up the difficulty by making you use fractions. You need to determine the correct amounts based on the total volume of the vat. Are you accounting for that? It's like having to modify a recipe to change the number that it serves.

The cafeteria is all about figuring out ratios (like what we are doing in Comparing and Scaling).

u r right!...but they offer cups of other sizes. If u filled a larger cup, is there a way 2 get the amount u need by playing with what they give u?

Hey 57e & 61. testing lab 2 is a \vamped up\" testing lab 1. What was absolutely true in testing lab1 ... each & every time? jot this down in ur Lure notebook pages. Bring ur notes to class...Let's look at that more closely tomorrow :)

These posts show evidence of engagement between teachers and students around the game. Many of them demonstrate teachers explaining how the game relates to math or lessons those students may have learned in class. Others are simply hints to get students on the right path. Importantly, the teachers provided context and asked students to think about the problems, without just telling them how to do the puzzle.

The third most common theme that emerged from the teacher posts was that of teachers explaining how to do things in the game to students (#2, above), which comprised 8.2% of the codable teacher posts. In these posts, teachers explained the “nuts- and-bolts” of how to play the game, but did not contain any math-related content. Examples of these posts include:

Did you complete the puzzle? When you do, the puzzle shuts you out with vines.

In order to go forward after entering 10 rooms, you have to have left a bean in each room. To leave a bean, you have to play the puzzle multiple times. Each time it gets harder. When you have played it enough, you leave a bean and you are done with that room. So if you can't get any new assignments, you need to go back to your rooms and play some more.

I'm the teacher, I can change my appearance, but not students.

As can be seen from the above examples, these posts were purely explanatory of how the game or aspects of the game functioned. Similar to these posts were the posts in which teachers provided technical support to the students (#3, above), which made up 2.3% of the codable teacher posts. A typical example of this type of post is,

….*If you are stuck and it won't let you go forward and it won't let you play, try the following things. First, try restarting your computer and browser (what you* *use to get to the internet). If it happens again, try taking a screen snapshot of your computer screen when this happens. Send that to my school email along with the name of the room you are in when it happens. Also, if you are using Internet Explorer as your browser, you can try switching to using Firefox (ask your parents if it is OK to download the browser before you do that). So far, we have hit this issue only with Internet Explorer, not with Firefox.*

It should be noted that technical support posts were very infrequent and a majority of them seemed to be dealing with the technical difficulties of a single student.

Summary

These data suggest that just under half (~46%) of teacher posts were attempts to facilitate student gameplay by either explaining how to accomplish something in the game or by posing questions to students in order to encourage them to think in ways that might help them make progress in the game. However, primarily, when teachers used the message boards, it was to tell students to post appropriately or simply to remind them to continue to make progress in the game. It is worth noting that although keeping students on task was the most common purpose for teachers' use of the message boards, that this activity was still very infrequent within the grand scheme of the game.

Q: How did students use the message boards? Was there evidence of collaboration among teams? How did team message board use relate to in-game performance?

A: *The most prevalent use of the message boards by students was to socialize and generally post about things that were not related to the game. When students were posting about the game, they generally were doing so to solicit help from their teammates. Generally, message board use was not associated with increased success in the game. However, the quality of posts that were about the game was associated with increased success in game. Thus, while overall message board use was not a necessary criterion for in-game success, there was a positive association between substantive message board use and in-game performance.*

A sub-set of student teams was randomly selected to have their message board posts examined in detail (n = 361 teams, a total of 4,228 posts). Of the 361 teams who were selected, approximately 1/3 (n = 121) did not post at all. Out of necessity, these 121 teams were excluded from the inferential analyses reported below. However, descriptive statistics regarding the quantity and quality of student posts are reported both with and without these 121 teams. The purpose of reporting the data in this way was to provide a sense of how frequently the message boards were used within the larger scheme of the game, while also providing information regarding how the message boards were used by the teams who opted to use them. Although the posts were coded at the individual level, these data were aggregated across team to provide a more accurate sense of the extent of collaboration and also so that we could more precisely address whether team-level message board use (i.e., collaboration) was associated with success within the game.

In the first pass of coding, each of the student posts was coded into one of three categories:

1.) Completely unrelated to game play or non-substantive

2.) Related to the game, but incidentally so (e.g., comments about characters, attempts to identify team members)
3.) Substantively related to the game (e.g., asking for help with some aspect of the game, providing assistance to a teammate, inquiries about game progress).

Approximately 20% of the students' posts were coded by two researchers so that inter-rater reliability could be established. Inter-rater reliability for this coding rubric was in the highly acceptable range (Cohen's kappa > 0.85). As a proxy measure for post quality, the length of each of the posts (in number of words) was also calculated.

The descriptive statistics regarding the quantity and quality of student posts is provided in Table 3. As noted above, the data are presented with all of the teams' data (to provide a sense of the extent to which the message boards were used within the game at large) and also only with the data of teams that used the message boards.

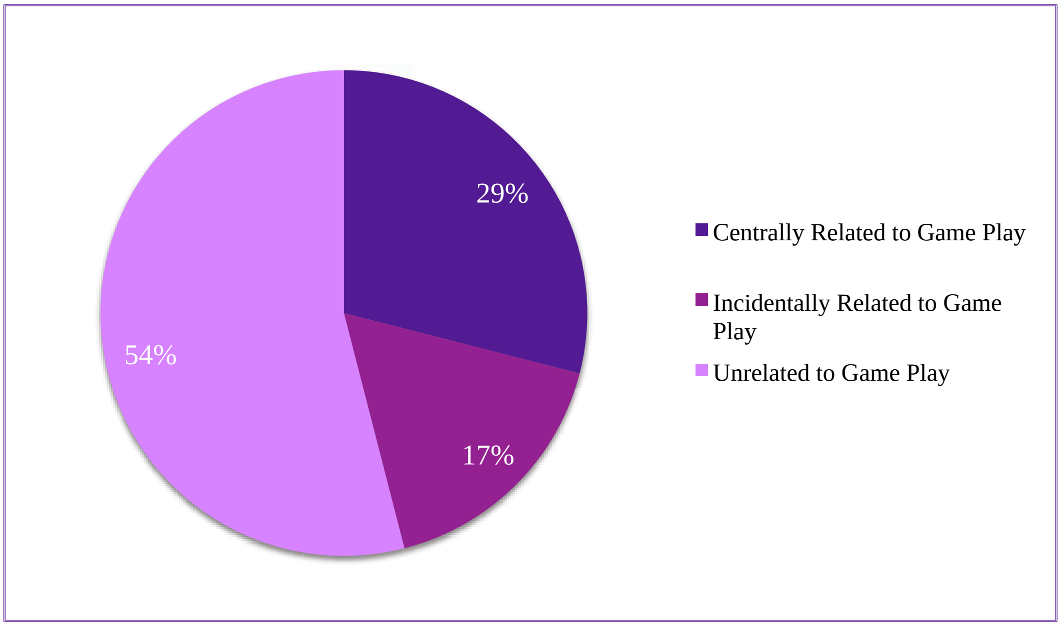
Table 3. Descriptive Statistics for Student Message Board Posts

	Average Number of Posts of this Type per Team (*SD*)		Average Length of Posts of this Type per Team (*SD*)	
	Entire Random Sample (*n* = 361)	Only Teams That Posted at Least Once (*n* = 240)	Entire Random Sample (*n* = 361)	Only Teams That Posted at Least Once (*n* = 240)
Posts Completely Unrelated to the Game	6.36 (12.36)	9.57 (14.12)	1.92 words (2.63)	2.90 words (2.75)
Posts Related to the Game, but Incidentally	2.04 (4.76)	3.07 (5.57)	2.56 words (4.14)	3.85 words (4.56)
Posts About Game Play	3.37 (8.12)	5.06 (9.53)	3.72 words (5.31)	5.59 words (5.64)
Total Posts (Across Content Types)	11.71 (19.83)	17.62 (22.08)	3.74 words (4.12)	5.63 words (3.86)

Comparing Quantity of Student Posts by Type

We conducted within-team pairwise comparisons using non-parametric sign tests to examine what post type was most common. Results of these tests showed that posts completely unrelated to game play were more prevalent than those related to game play, either incidentally or centrally. Second most prevalent were substantive posts that were centrally related to game play; least prevalent were posts that were related to game play, but incidentally (all *p*s <0.01).

Figure 2. Break Down of Posts by Type (Aggregated over Team).



Correlating Message Board Use with In-Game Performance

It seems reasonable that the quantity and quality of a team's substantive (i.e., centrally related to game play) posts would be positively associated with their performance in the game. Here, in-game success was indicated by the average number of pets freed per team member. Although this is not the sole measure of success within the game, this was used as a proxy for student engagement in these analyses. To test these hypotheses we first correlated teams' proportion of total posts that were centrally related to game play with the average pets freed per team member. This relationship was non-significant (*Spearman's Rho* = 0.17, $p>0.10$). Second, we correlated the average length of the posts that were centrally related to game play with the average pets freed per team member. This relationship was significant (*Spearman's Rho* = 0.32, $p<0.05$). Therefore, although the number of substantive posts was not associated with in-game performance, the quality and/or depth of the substantive posts was positively associated with in-game success. This pattern of results is not surprising given the way the substantive posts were coded– e.g., if a student simply typed "I need help" that was counted (within the boundaries of our coding scheme) as an attempt to elicit collaboration. There could have been multiple short posts of that nature that did not help students move forward, but a few long ones, which helped students to progress.

It is also worth pointing out that message board use generally was not associated with in-game performance. In order to investigate this, we ran a linear regression with a dummy-coded predictor that indicated whether or not the team used the message boards (0 = no message board use; 1 = message board use) and average pets freed per team member as the outcome measure. Message board use was not significantly associated with in-game performance (p>.50).

Teacher interviews indicate that part of the reason for this lack of a relationship between overall message board use and in-game performances may be because students were not using the in-game messaging system as their primary means of within-team communication. Rather, they were often able to take time in class to discuss with their fellow players and teachers how to do certain puzzles. More than one teacher explained that kids would often just get up and walk across the room to show another student how to pass a puzzle. Similarly, one teacher explained that even when students were playing at home at least one group of students used Skype to communicate about the game rather than use the in-game messaging system. Therefore, although students who used the message boards in a substantive way tended to perform well in the game, there is no general association between message board use and in-game performance. More than likely, the significant correlation found between substantive post quality and in-game performance was the result of an underlying factor of general engagement with the game, as it is clear from these data that simply posting to the message boards is an insufficient criterion to predict in-game success.

In order to gain a better understanding of how the message boards were used by students under ideal circumstances, all of the posts that were centrally related to the game were coded for finer detail. Each of these substantive posts (n = 1224) was coded into one (or more) of the following categories[4]:

1.) Asking for help
2.) Sharing game progress
3.) Inquiring about game progress
4.) Explaining an aspect of the game
5.) Other

Approximately 20% of the substantive message board posts were coded by two researchers to establish inter-rater reliability. Inter-rater reliability was in the highly acceptable range (Cohen's kappa > 0.85)

The first category, *Asking for help*, consisted of requests for help with problems , e.g., "can someone help me with the cafeteria place?" or "where is xx room?" It also included questions about how to solve puzzles, questions about where things were in the game, and sometimes included questions about mathematical formulas or how to apply a math concept to solve a problem. The second category *Sharing game progress,* included things like "I'm stuck" or "I have freed xxx pets!" but not explicitly asking for help. The third category, *Inquiring about game progress*, consisted of students asking where other players were in the game, which levels they had passed, how many pets they had freed or if another player needed help. The final substantive category, *Explaining an aspect of the game*, included students helping each other to understand the underlying mathematical concept in order to solve the puzzle (e.g., "you need to add the fractions"); describing what the goal of a puzzle is without providing any steps to get there (e.g., "you need to free all the pets" or "you're supposed to pour the right amounts of each ingredient"); explaining the steps taken to solve the puzzle without necessarily focusing on the underlying mathematical concepts; or sharing a formula or equation with another student to help them with some aspect of the game.

The coding rubric applied to these posts was meant to capture the degree to which students engaged in *deeper learning* skills, which include *problem solving*, *effective communication*, *collaboration*, and *learning to learn.* In our conceptualization, each of our codes captured some aspect of these four skills.

Problem Solving and *Effective Communication* were meant to be captured through 1) explanations of the math needed to progress in the game, 2) explanations of the required steps to solve puzzles, and 3) strategies and tactics to be used in the game. Unfortunately, there were so

[4] These codes were developed based on the research of Steinkuehler, C., & Duncan, S. (2008), which focuses on the development of scientific habits of mind in video gaming, and Swing, S & Peterson, P. (1982), which explore how students learn math in small group settings.

few posts that illustrated these, that they were collapsed into one all-encompassing code of "Explaining." While at times posts in this category did illustrate effective communication (especially within the realm of online messaging, which is often in shorthand), there were few that were truly exemplars of effective communication.

Collaboration was captured by all four of the codes as they all illustrated, by their nature, a degree of collaboration. That said, some of the posts were clearly more "collaborative" in nature than others. For example, asking for help could be something as simple as "help." In those cases, it was not clear what specific element of the game students were asking for help with, but the mere fact that students were willing to ask for help from fellow students was seen as significant, as it represented an attempt to collaborate on the part of a team member.

Learning to Learn was captured by the "Asking for Help" and "Explaining" codes. Because *Learning to Learn* is a concept that demonstrates students' ability to frame a problem, ask the appropriate questions, and apply the correct tools and concepts to solve a problem, we saw students' willingness to ask for help when it is needed as significant. Further, students who responded to these requests for help (or who offered unsolicited assistance) demonstrated an ability to identify which aspects of the puzzle were likely to be problematic and share with their teammates what the appropriate course of action is, also a component of *Learning to Learn.*

As noted above, all of the student posts that were centrally related to game play (n = 1224) were coded into one (or more) of these five categories:

1.) Asking for help (*Collaboration, Learning to Learn*)
2.) Sharing game progress (*Collaboration*)
3.) Inquiring about game progress (*Collaboration*)
4.) Explaining an aspect of the game (*Problem Solving, Effective Communication, Learning to Learn, Collaboration*)
5.) Other

Note that because a single post could include more than one type of content, the total percentages provided below surpasses 100%.

Approximately 14% of the posts did not fit into one of the emergent categories we identified from the data. Those posts aside, the most frequently observed substantive student post consisted of students asking each other for help (35% of posts). The substance of these requests for help varied greatly with some simply being the word "help" and others being much more detailed and elaborate, but the underlying sentiment was the same. Some of theses posts were implied requests for help such as,

> *i cant find the right room/ house/ triangle? The map doesnt make any sense!!!!!!!!!!!!!!!!!!!!*
>
> *i cant finish the test tube thingy grrr its confusing*
>
> *I dont get it does any1 else get it?*
>
> *AHHHH i am stuck!*

While these students aren't explicitly asking for help, we interpreted these as "Asking for help" because they have posted on the messaging system places where they are having trouble which implied the need for help. Others were more explicit in their request for help but not clear about what they needed help with. For example,

help me

I need help with a problem :/

I need help

All of the above examples were posted in the “general” discussion room, so it was not clear what the students are asking for help with. These generalized requests for help sometimes elicited responses from team members asking for additional detail. At other times, these simple help messages were posted within particular rooms’ chat boards such as the Employee Cafeteria. In those cases it was clearer what the student was requesting help with, but specific details regarding their problem were still lacking. It was very rare that a student made a request for help that specifically outlined the problem. In the context of how it seems many actually played the game in the classroom, these short messages may make more sense: students may not have needed to explain in detail what they were having trouble with because they were able to talk to their classmates about it face-to-face. The simple “help” messages may have served as a quick method for getting a teammate to come to their aid in the classroom. That said, there were still some more specific requests for help that did appear in the message boards, though it was not clear that these always elicited more descriptive responses. These more specific requests included,

What Pleases the monster guy in Minecraft 1??

How do you get to the Employee Lounge?

How do you know who to give the food to?

ok... i have inventory and i am in a cave how am I supposed to use them?????????

And from the Employee lounge,

what do the colors mean?

Thus, although the intent of coding requests for help was to show evidence of *Collaboration* and *Learning to Learn*, given that a great deal of communication likely happened outside of the message boards, we cannot say for certain the extent to which the Lure of the Labyrinth fostered these skills in students.

The second most common use of the message boards for issues that were centrally related to game play was students sharing their progress in the game with their teammates (29% of posts). One of the more common ways in which students shared their progress was by revealing to their teammates how many pets they had freed, indicating that this may have been a good motivator for the students. Examples of these posts include,

i saved 465 pets!!

Thats great. I've saved 850!

I just let out some animals. Now i have 110 pet point things

Others kept a running tally for their teammates, for example,

i have 540 pets

690 pets

800 pets

920 pets

1035 pets ftw

1210 pets ftw... again

Clearly, at least some of the students were motivated by this aspect of the game.

Students also shared about the puzzles they had passed, or where they were in the game:

I BEAT THE LEVEL WOOHOO!!!!!!!!

im still stuck on number 10 its super hard

I have 17 rooms unlocked and 15 completed…

For many students, sharing their progress was an important part of the game, and a way in which they could interact with their friends or classmates. Sharing progress helped the students share their accomplishments, but it also allowed for updates, and allowed students to know who to ask if they needed help. Thus, having a reward system built into the game seemed to foster some degree of collaboration among student teams.

Next most prevalent in the message boards, was students explaining something about the game to their teammates (14% of posts). These posts were typically (though not always) made in response to someone asking for help. Some of these posts were highly descriptive posts of steps or processes such as,

Ok so it's basically numbers as you can see. So, first I just started to randomly select things and it kinda worked, but not really. So, at the right of that calculator thing you've got there there's math problems-really easy. So when you guess one of the number symbol things it's gonna show up what the number is. What can *also help is when you look at the tray thing with the symbols written on the bottom of it and you can count how many of the green things there are to find that answer. If they're just stacks they'll be multiplication or division and if they're two seperate things they'll be addition or subtraction.*

When you get to the flowers and shapes, you have to count the # of sides on the shape and then check where the flowers are. You cant land on any flowers so basically the multiples of the shapes # of sides.

For the testing lab you mess around with the values of the cups. For example, in order for me to get 1 beetle eye with the cups 6, 3, and 2; I have to grab the beetle eye with the 3 and pour it in 2 (I think)

guys if you look at all the doors that say like any letter and number find a cave and you have to buy stuff from the store and you give it to them and you get to free animals its the same with every room like the houses on pink lava the wells thats the same and the bomb thingys the holes you do the same and you can free animals if anyone needs any help let me know

Though the players sometimes struggled to articulate themselves, these posts show an attempt to explain how to pass puzzles and some of them even make use of math terms in order to explain the puzzles. Many of these posts were students sharing which items to use in each of the pet cages, for example,

watergun for electro fence

termites for boxes

banana for monster guy

Other times they were sharing specific directions for how to navigate the game or hints about the game.

go to the end - the left

go to a3

go in the door

iris gives u a letter dont talk to her twin

nevermind i hate to play it again for some reason....i kinda figured it out, you have to get the guy to land where the big purple thing is.

When used in this way, the message board system seemed to afford students opportunities to communicate with each other, to help each other problem solve, and to help each other identify and articulate what components of the game were most problematic or challenging. However, it is not clear from these posts that students were engaging with each other in the rich way that was intended to be captured by the coding rubric. That being said, given that this was one of several modalities through which team members communicated with one another, it is hard to say with certainty the extent to which the game fostered collaboration skills.

The least frequent use of the message boards was students asking their teammates about how far they had progressed in the game (12% of posts). Examples of these posts included things like:

> *anyone on warhouse yet?*
>
> *[student name] where are you?*
>
> *how many pet do u have*
>
> *have u saved any pets yet*
>
> *Hey can everyone like write down the number of rooms they've completed (as of right NOW)??? I think it's a good idea to add them up!!!!*

Interestingly, one of the more common types of posts was students asking if others needed help, for example,

> *need any help?*
>
> *Anyone need any help?*
>
> *This is easy , & yeah i'm suree (: , do you need help thuddd ? , whoever that is ?*
>
> *Serria , do you need help , that one is so eassy , i can help you ?*
>
> *i cant remember where the can goes but i can help with the cafeteria*

This was an interesting theme that emerged out of these messages. It seems that part of the reason for asking and sharing about progress was not just to create a sense of competition, but also to open up opportunities for help. In talking with teachers, at least one pointed out that she encouraged this sort of behavior, so this could be a result of teacher facilitation. However, it could also be because teammates were motivated to help each other in order to progress further in the competition.

<u>Summary</u>

In sum, the data do suggest that students used the message boards to collaborate in various ways. There was some evidence of *Deeper Learning* skills in the message boards as well, including showing a willingness to ask questions and share progress, hints and strategies with teammates. Some of that sharing included students explaining the math concepts used in solving the puzzles, albeit in a somewhat limited way. However, this substantive use of the message boards was very infrequent within the larger scheme of the game, with the entire sub-sample of teams only posting just over 3 substantive messages on average. As noted before, this lack of use is likely due to the fact that students had numerous other means for communicating with one another while they were playing the game. Along that vein, although the quality (i.e., length) of student posts that were substantively related to game play was positively associated with in-game performance at the team level, message board use generally was not a significant predictor of in-game performance. As will be discussed in greater detail in the Limitations section, there are many reasons why students did not use the message boards. These possible reasons come from both practical consideration of the circumstances around game play, as well as from the additional context provided by the teacher interviews.

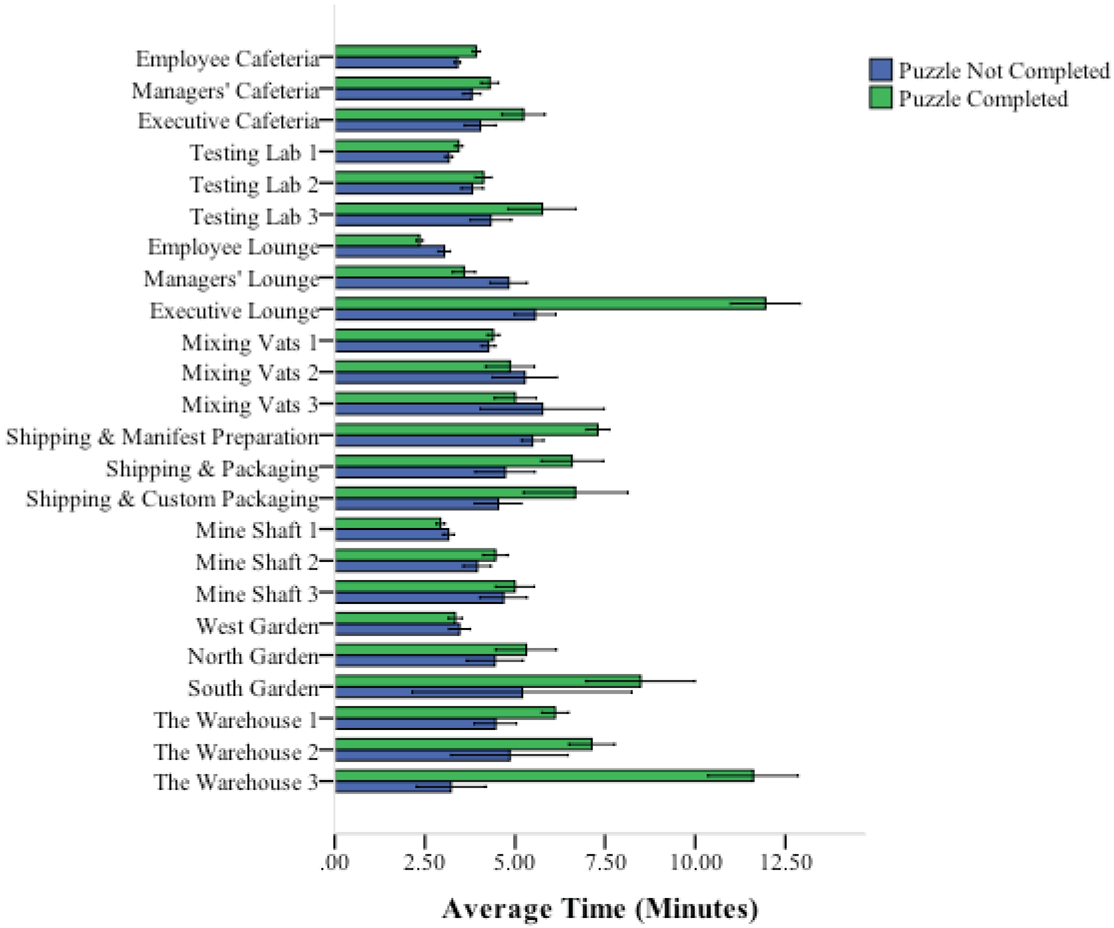
Q: How engaged are the students with the game over time? On which levels did students persist the most?

A: *In terms of persistence, students seemed to put in the most effort on puzzles that were the most difficult within a given level. This demonstrates a capability and willingness on the part of the students to put in additional effort when the game required it, and also serves to corroborate the intended structure of the puzzles within levels (i.e., that they get more difficult). Student engagement was fairly variable, however the data suggest that there may be two game play patterns among students, with one group of students dropping off primarily during levels that are unlocked earlier in the game and a second group of students persisting through to levels unlocked later in the game.*

To examine students' persistence and engagement with the game, we looked at the amount of time students' spent playing each puzzle, the average number of puzzles students completed, the number of students who beat each puzzle, and the final puzzle that students beat before they stopped progressing through the game. Note that the data analyzed here included all eligible student participants, not just a random sub-sample.

With regard to time spent with the puzzles, on average, students spent 3.85 minutes (*SD* = 5.56) on attempts in which the puzzle was not beaten and 4.21 minutes (*SD* = 5.62) on attempts in which the puzzle was beaten. As shown in Figure 3, with minimal exception, students spent more time on puzzle attempts in which they ended up beating the puzzle. This could be due to things like students restarting a puzzle as soon as they realized they made an error or to students being more cavalier in their actions within the game on unsuccessful attempts. Using time spent on successful attempts as a proxy measure for puzzle difficulty, these data suggest that the Executive Lounge was the most challenging, with an average completion time of 11.95 minutes (*SD* = 12.11) while the Employee Lounge was the easiest, with an average completion time of 2.35 minutes (*SD* = 3.05). As shown by the green bars in Figure 3 there was, generally, a linear increase in the amount of time taken to complete each puzzle as the puzzles became more difficult within a given level (e.g., within The Cafeteria, students' solve time for the Employee Cafeteria was shorter than for the Managers' Cafeteria, which in turn was shorter than the solve time for the Executive Cafeteria). This shows that students were willing and able to put in additional effort when the game required it and also serves to show that the puzzles did increase in difficulty within a level, as intended by game developers.

Figure 3. Average Length of Puzzle Attempt in Minutes



Note. Error bars represent +/- two standard errors of the mean.
Note. All of these data are presented with outliers (i.e., data points more than 2 standard deviations above the mean) removed. The removal of outliers was particularly important given the nature of the data – for example, if a student was playing a puzzle but neglected to log out of the game, this would be logged as a very long puzzle attempt but is not representative of how the game was actually played.

With regard to the engagement with the game, on average, students completed 7.48 puzzles (*SD* = 11.94). However, there was a great deal of individual difference in the amount that students progressed through the game. Table 4 provides a list of each puzzle and the percentage of students for whom that puzzle was the apex of their game play. In this table, a small percentage indicates that most students who completed that level continued on to play other puzzles, whereas a large percentage indicates that many students who completed that level did not go on to play other puzzles. This is not to say that these students did not *go back* and play previously unlocked puzzles, but rather that after completing the particular puzzle they did not unlock any new puzzles.

Table 4. Student Engagement with the Game

Puzzle	Number of Students Who Beat Puzzle (of the 13,410 students who played the game)	Number of Students for Whom this was the Final Puzzle Beat	Percentage of Students Who Stopped Progressing After Beating This Puzzle
Employee Cafeteria	6298	1223	19%
Manager's Cafeteria	1186	39	3%
Executive Cafeteria	463	5	1%
Testing Lab 1	6004	1559	26%
Testing Lab 2	725	37	5%
Testing Lab 3	355	5	1%
Employee Lounge	4211	711	17%
Manager's Lounge	1345	131	10%
Executive Lounge	414	17	4%
Mixing Vats 1	2034	232	11%
Mixing Vats 2	294	2	1%
Mixing Vats 3	168	0	0%
Shipping Manifest & Preparation	908	25	3%
Shipping & Packaging	306	0	0%
Shipping & Custom Packaging	282	0	0%
Mine Shaft 1	4949	2171	44%
Mine Shaft 2	1528	354	23%
Mine Shaft 3	706	137	19%
West Garden	1664	545	33%
North Garden	657	158	24%
South Garden	264	51	19%
The Warehouse 1	2355	1634	69%
The Warehouse 2	726	393	54%
The Warehouse 3	336	336	100%

As shown above, the most common stopping points for students were the puzzles in The Warehouse. Clearly, this is due to the fact that these were among the last puzzles that could be unlocked within the game. However beyond that, these data suggest that student engagement may be bi-modally distributed, with one group of students dropping off primarily during levels that are unlocked earlier in the game and a second group of students persisting through to levels unlocked later in the game.

Summary

In sum, these data show a great deal of variability in students' persistence with individual puzzles and also the extent they remained engaged with the game. However, from these data it is clear that the intended pattern of difficult of the three puzzles within each level corresponded to the amount of persistence shown in students' attempts to beat the puzzle. Further, by considering the patterns of engagement shown in Table 4 as they relate to specific puzzle content and difficulty, you may be able to gain a sense of what puzzle characteristics may be associated with an increased likelihood in students' continued progress through the game.

Q: Did students show evidence of learning across consecutive puzzle attempts? What patterns of learning emerged from the data?

A: *The extent to which students exhibited learning across consecutive attempts at the same puzzle seemed to vary as a function of the difficulty of the puzzle. On the easier puzzles, students made clear improvements in their performance across subsequent attempts to complete the puzzle. However, on the more difficult puzzles, students' seemed either to perform consistently or to subtly decline over subsequent attempts.*

Student learning within the game was examined for three levels: The Cafeteria (comprised of the Employee Cafeteria, the Manager's Cafeteria, and the Executive Cafeteria), The Recipe Room (comprised of Testing Lab 1, Testing Lab 2, and Testing Lab 3) and the Vending Machine (comprised of the Employee Lounge, the Manager's Lounge, and the Executive Lounge). In order to gauge learning over time, we looked at the proportion of correct moves each student made on subsequent attempts at the three puzzles contained within each level. We chose to focus on the first five attempts students made on each of these puzzles. This decision was made based on examination of the raw data. On average, students attempted each puzzle 2.35 times (*SD* = 2.42) irrespective of whether or not the puzzle was beat. The number of puzzle attempts was very positively skewed: 94% of the data fell within one *SD* of the mean. Although it is atypical for so many data points to fall within one *SD* of the mean, given that it was the case here, five attempts seemed a reasonable cut-off point, as it was representative of students' general gameplay tendencies.

To look at student learning over time, the data were analyzed using hierarchical linear modeling (HLM).[5] Hierarchical linear modeling is an appropriate analytic method for these data because there is nesting (i.e., multiple puzzle attempts nested within students, students nested within teams, and teams nested within teachers). This nesting needs to be taken into account statistically in order to get accurate results. Not properly taking this shared variance into account can lead to statistical tests that are too liberal (i.e., an increased likelihood of a false positive). Another advantage of this analytic method is that missing data are not a problem (e.g., in a case where a student attempted a puzzle level fewer than five times).

In each of the nine hierarchical linear models (one for each puzzle), we tested for linear and non-linear patterns of learning. However, only the pattern that most closely fit the data (i.e., that was statistically significant) is presented here. Across all nine puzzles, there was substantial individual variability in students' baseline performance (i.e., how well they fared on their first attempt to complete the puzzle). However, regardless of how students performed at baseline, the pattern of learning tended to be uniform across students. That is, whether a student made correct moves 20% of the time or 80% of the time at baseline, the overall pattern of their performance

[5] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications, Thousand Oaks, CA.

on subsequent attempts was the same. For more detailed information regarding the statistical analyses, please see the Technical Appendix.

Employee Cafeteria

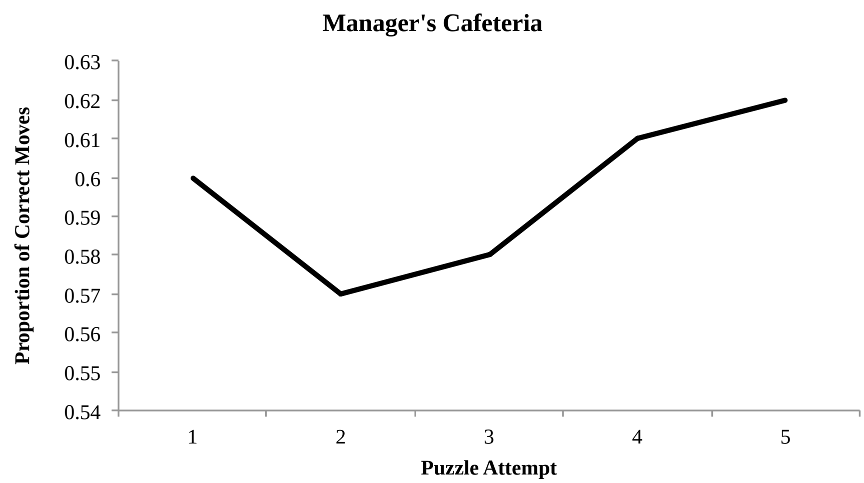
On average, students made correct moves about 50% of the time on their first attempt at the puzzle in the Employee Cafeteria. Across subsequent attempts at this puzzle, student learning was non-linear. As shown in Figure 4, this pattern was the result of students making steep learning gains across the first three attempts at the puzzle, but showing slightly less evidence of learning between their third and fifth attempts. With regard to success in this puzzle, 52% of students beat it at least once in their first five attempts.

Figure 4. Student Performance Over Time in the Employee Cafeteria Puzzle

Manager’s Cafeteria

On average, students made correct moves about 60% of the time on their first attempt at the puzzle in the Manager’s Cafeteria. Across subsequent attempts at this puzzle, student learning was also non-linear. As shown in Figure 5, this pattern was the result of students’ performance declining between attempts 1 and 2, but then improving incrementally on each subsequent attempt. With regard to success in this puzzle, 29% of students beat it at least once in their first five attempts.

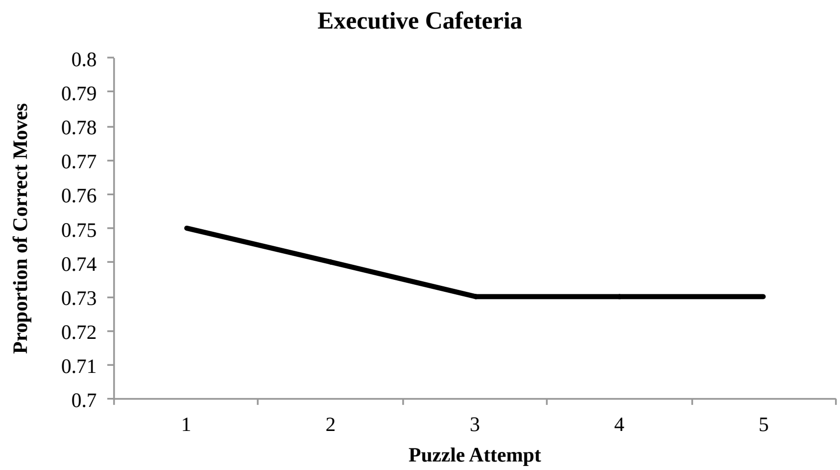
Figure 5. Student Performance Over Time in the Manager’s Cafeteria Puzzle



Executive Cafeteria

On average, students made correct moves about 75% of the time on their first attempt at the puzzle in the Executive Cafeteria. Across subsequent attempts at this puzzle, student learning was also non-linear. As shown in Figure 6, this pattern was the result of students' performance declining steadily between attempts 1 and 3 and then plateauing through attempt 5. With regard to success in this puzzle, 35% of students beat it at least once in their first five attempts.

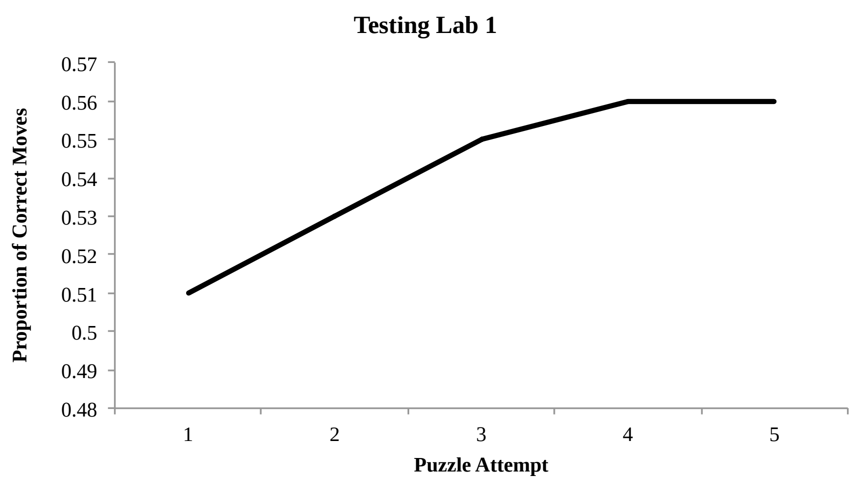
Figure 6. Student Performance Over Time in the Executive Cafeteria Puzzle



Testing Lab 1

On average, students made correct moves about 51% of the time on their first attempt at the puzzle in Testing Lab 1. Across subsequent attempts at this puzzle, student learning was non-linear. As shown in Figure 7, this pattern was the result of students' performance improving steadily from attempt 1 through attempt 3, improving somewhat between attempt 3 and 4 and then plateauing through attempt 5. With regard to success in this puzzle, 55% of students beat it at least once in their first five attempts.

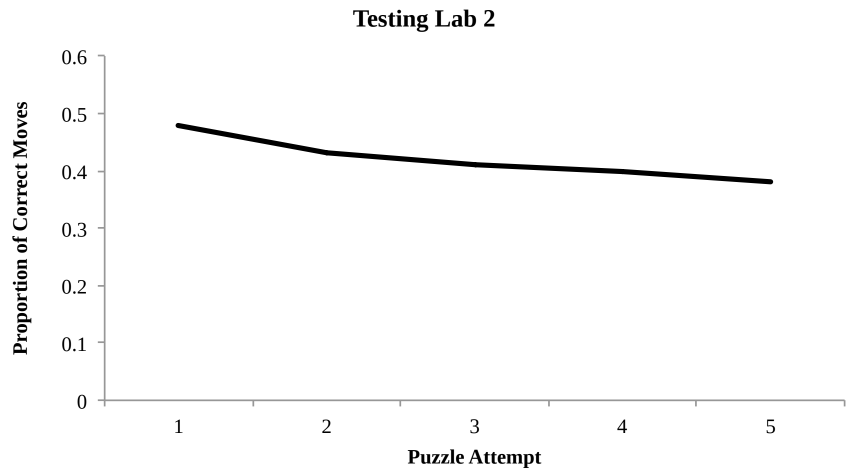
Figure 7. Student Performance Over Time in the Testing Room 1 Puzzle



Testing Lab 2

On average, students made correct moves about 50% of the time on their first attempt at the puzzle in Testing Lab 2. Across subsequent attempts at this puzzle, student learning was linear. As shown in Figure 8, this pattern was the result of students' performance declining steadily across each of their five attempts at the puzzle. With regard to success in this puzzle, 31% of students beat it at least once in their first five attempts.

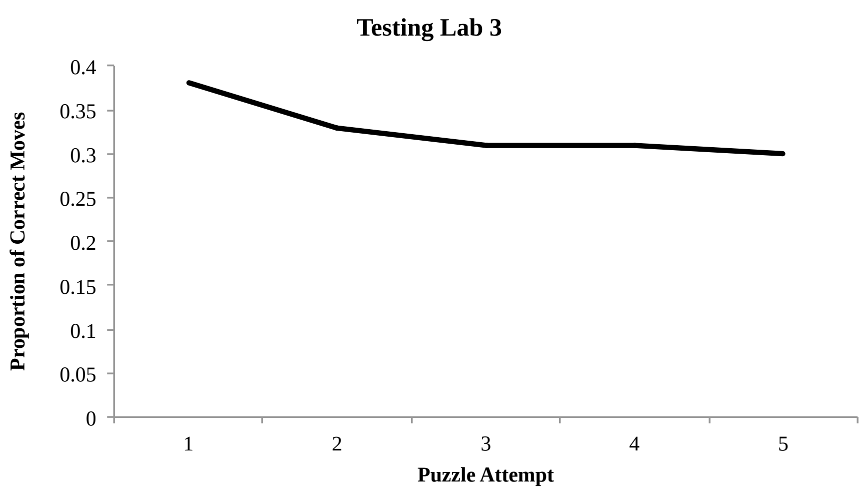
Figure 8. Student Performance Over Time in the Testing Lab 2 Puzzle



Testing Lab 3

On average, students made correct moves just under 40% of the time on their first attempt at the puzzle in Testing Lab 3. Across subsequent attempts at this puzzle, student learning was non-linear. As shown in Figure 9, this pattern was the result of students' performance declining between attempts 1 and 2, and then continuing to decline, though less severely, across subsequent puzzle attempts. With regard to success in this puzzle, 68% of students beat it at least once in their first five attempts.

Figure 9. Student Performance Over Time in the Employee Cafeteria Puzzle



Employee Lounge

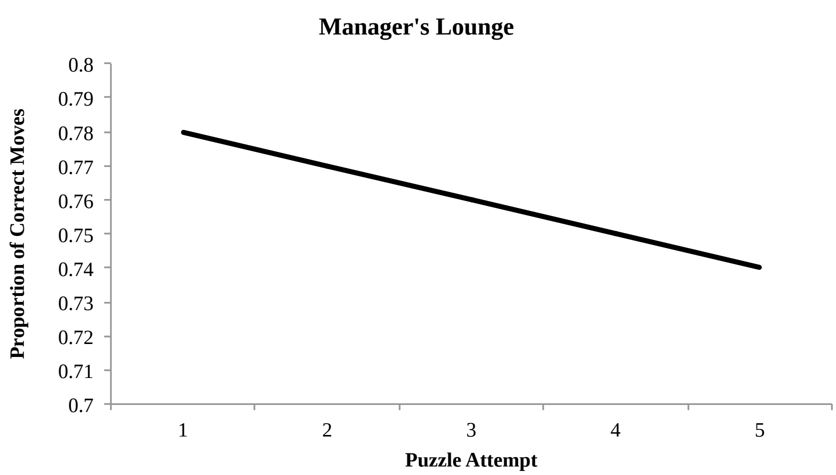
On average, students made correct moves about 68% of the time on their first attempt at the puzzle in the Employee Lounge. Across subsequent attempts at this puzzle, student learning was non-linear. As shown in Figure 10, this pattern was the result of students' performance generally improving over time, but at a greater rate across earlier attempts to complete the puzzle. With regard to success in this puzzle, 46% of students beat it at least once in their first five attempts.

Figure 10. Student Performance Over Time in the Employee Lounge Puzzle

Manager's Lounge

On average, students made correct moves about 78% of the time on their first attempt at the puzzle in the Manager's Lounge. Across subsequent attempts at this puzzle, student learning was linear. As shown in Figure 11, this pattern was the result of students' performance declining steadily across each subsequent attempt at the puzzle. With regard to success in this puzzle, 74% of students beat it at least once in their first five attempts.

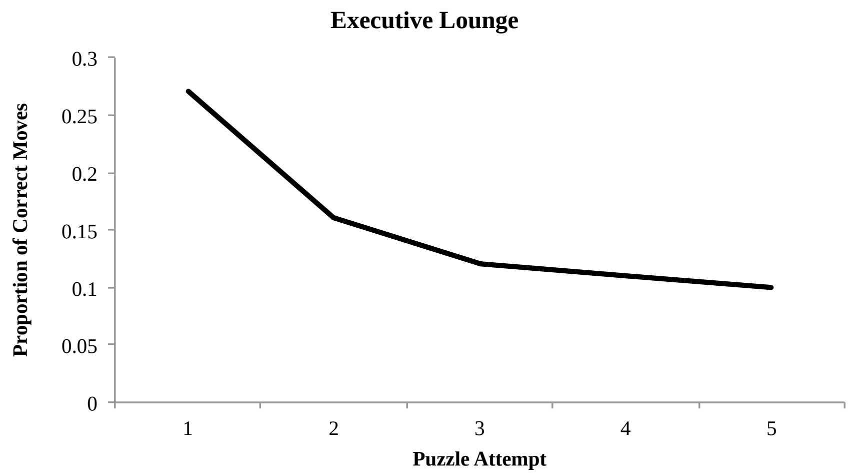
Figure 11. Student Performance Over Time in the Manager's Lounge Puzzle



Executive Lounge

On average, students made correct moves under 30% of the time on their first attempt at the puzzle in the Executive Lounge. Across subsequent attempts at this puzzle, student learning was non-linear. As shown in Figure 12, this pattern was the result of students' performance declining sharply between attempts 1 and 2, and then continuing to decline, albeit less severely, across subsequent attempts at the puzzle. With regard to success in this puzzle, 53% of students beat it at least once in their first five attempts.

Figure 12. Student Performance Over Time in the Executive Lounge Puzzle



Summary

Overall, these results show that when presented with a puzzle of minimal difficulty, students were able to make consistent progress as they made repeated attempts. On the puzzles of medium difficulty, student performance tended to be fairly consistent over subsequent attempts. Finally, on the most difficult puzzles, student performance tended to decline with subsequent attempts, perhaps due to frustration (e.g., in cases where students were unable to solve the puzzle). However, it is important to highlight that in these data, it is not clear what percentage of correct moves was necessary for a student to beat a level. This information would be useful to provide additional context when interpreting these figures. It also is not clear what the probability of making a correct move is within a given level. For example, if there are only two possible moves then the probability of making a correct one by chance is much greater than in a level where there are five possible moves. This is also something that should be considered when interpreting these figures. With regard to the percentage of students who beat the puzzle on the first five attempts, it is important to keep in mind that the number of students who attempted each puzzle became successively smaller with increasing puzzle difficulty. As such, we would not necessarily expect this percentage to decrease with puzzle difficulty because it could be that the more skilled players are more likely to persist on and complete the more advanced puzzles.

Q: What characteristics did the top-performing teams possess? How was the game implemented by the teachers who supervised these teams?

A: *There was substantial variability in the settings in which the Lure of the Labyrinth Challenge was played, how individual teachers chose to implement the Challenge in their classrooms, how teachers chose to assemble the student teams, and how (if at all) the message boards were used within these teams. A key similarity across all the teachers who led successful teams was that they were all very interested in and enthusiastic about the game. Many teachers reported having played themselves before the launch of the Challenge and seemed to really enjoy the game and looked forward to sharing it with their students.*

In order to provide a more comprehensive picture than the game play data could reveal, we decided to interview a sub-set of teachers who supervised highly successful teams in order to better understand what these teams may have had in common. Specifically, we wanted to better understand 1.) How the teachers explained the game to the students, 2.) How student teams were assembled, 3.) How the game was used in the classroom and 4.) How the teachers interacted with and used the game themselves, including whether they played it themselves, how (if at all) they used the messaging system, and how they used the administrator tool. In order to answer these questions, we conducted eight interviews with teachers who supervised the highest performing teams (as measured by the average number of pets freed in the game per team member). In addition to considering team performance when identifying potential interviewees, we also looked at the median incomes of the communities of the teachers to ensure we got a variety of perspectives. All of the interviewees happened to be women. The interviews typically lasted between 15 and 30 minutes and followed a specified protocol (see Appendix A).

The first thing that became clear from the interviews was that there was not a model set of characteristics or behaviors that emerged. The teachers all seemed to be quite different in their approach to using the *Lure of the Labyrinth*. There were also several different settings in which the game was used. For example, two of teachers were computer lab teachers rather than math teachers. One of the schools was a private school for gifted students; another was a math and science academy. Others were public schools (questions about the schools were not asked in the interviews so we do not have detailed information about the schools except for what came up naturally during the interviews).

Introducing the *Lure of the Labyrinth*

In addition to the heterogeneity observed across educational settings, there was also substantial variability in the way that teachers decided to introduce the game to their students. Some teachers interviewed were very thorough in their preparation for the game, while others took a more *laissez-faire* approach. Similarly, some were quite familiar with the game to begin with, and had in fact played it before, while others were new to the game and unclear about the Challenge. One teacher, a licensed math teacher who ran a computer lab and who had played the game before, explained her approach:

> *I tried not to explain it too much to them, that they could find out more about it on their own…. you know just let them learn how to go…I think gaming is a good way to learn. It wasn't foreign to them to play games. Its very comfortable.*

This is in contrast with another teacher who sat with her students for the beginning of the game, played it with them and went over what was involved with the game as well as with the Challenge.

Some teachers used the Challenge as a way to motivate students, while others downplayed the competitive element. One teacher explained that she waited a week to tell her students about the Challenge because she wanted to see how many would be motivated without the contest. Another teacher shared:

> *We didn't even think about contests or, you know, that there was going to be prizes, and I didn't even realize, that, you know about the prizes. We had heard about this, like, my grade partner and I had heard about this at the [NCTM] math convention here in Philadelphia last year and we just thought it would be, you know, such a great thing for the kids to do, especially with their poor ratio and proportion – it seems like a lot of the puzzles were dealing with that. And since they are poor in that, we thought it would be a good way to get them excited about math [laughing].*

While some de-emphasized the contest or simply did not know about it, some used the contest as a way to motivate their students. The competitiveness in some of these classrooms also led to some contention about teams. One teacher who chose to emphasize the competition had to deal with students wanting to work only with the strongest math students so that they could win prizes. Another teacher who felt badly that her students were working really hard and making good progress and not seeing any rewards decided to reward them with little prizes as they passed some significant benchmarks within the game.

The way the teachers introduced the game to students, then, varied across high-achieving classrooms with some teachers putting in significant effort to prepare themselves and their class for the Challenge. On the other hand, some teachers took a more hands-off approach, letting students figure out things on their own and encouraging them to ask their peers if they had questions. Based on this variability, it seems that success in the game was not necessarily related to the way in which teachers prepared themselves or their students; nor was it due to whether or not the students were aware of the Challenge.

<u>Forming Teams</u>

When teachers were asked how they chose to group the students together in teams it was clear that they took different approaches. Several of the teachers explained that because the students were already grouped together in high-, medium-, and low-achieving math classes, that the teams within those classes were somewhat homogenous (though as one teacher noted, there was still some variability within those streamlined groups). At least one teacher explained that she tried to group a high-achieving student with two middle-achieving students and one lower-performing student. Several of the teachers explained that they just let the students choose their own teams since it was an enrichment activity. Further, two teachers (one from an advanced math class and one from a computer lab) regrouped students in the midst of the challenge, mostly because the highly motivated students expressed an interest in working together in order to win prizes. As one teacher explained, *[the students that] really, really had a drive to somehow figure out how to, you know, get to the next level of releasing animals and so on, so um, those students I just regrouped, because those were the ones that really took it far…*

Aside from student ability, team sizes also varied. Some teachers reported teams of two students while others had teams of up to six. In fact, one teacher explained that most of her groups were five to six students and that she merged smaller groups of students together. It is clear from the interviews that there was no consistent way that high performing teams were organized. Though, at least two of the teachers did regroup students into new teams during the challenge so that those who were more interested and motivated could work together, despite receiving instructions from game developers not to do so. It is likely then, that the reason these teachers had high performing teams is because they allowed the most interested and motivated students to work together.

Game Play

Although it is clear from talking to teachers that they took many different approaches to explaining the game and grouping the students, there was some consistency in the way they described in-class gameplay. All of the teachers we spoke with allowed students to play the game during class time. In some situations, students all had their own computer while in other classes they had to share computers. It was not clear that one of these situations worked better than the other. All of the teachers allowed the students to talk amongst themselves and figure out the puzzles, and in fact most of them mentioned that they encouraged this. While in theory they could have been using the message boards to communicate with each other it was not necessary in these class settings, as one teacher explained,

Oh yeah they would walk around. They had the online conversation thing, you know the chat, they could use the chat, but basically they didn't have enough patience to communicate in writing, how to do a task, but they were able to express their frustration if they were stuck, or if they didn't know what to do.…so they would mostly vent …"ah its not working" or "I can't do this" "I don't know what to do" or "where is that?" you know, and then maybe the person would respond in writing, which is what we would want, but most often they got up out of their seat and go walk around and help each other…

The message boards then, were unnecessary for many teams because they were able to talk with one another face-to-face. This is consistent with our finding that message board use was not a necessary component of successful gameplay.

The pets freed total came up in several of the teacher interviews as well, and while teachers indicated that the students seemed motivated by this component of the game, there was some frustration on both the part of the students and teachers that teams so quickly reached the maximum number of pets freed allowed within the game. This frustration persisted even after game administrators sent out an update stating that the maximum number of pets that could be freed was 1360. While this was likely meant as a motivator for teams, it led to frustration among some who could not seem to reach that number and instead could only reach 1260. One teacher mentioned that a new team of the best players in the class formed mid-way through the challenge with the express purpose of reaching 1360 pets freed. However, even a team comprised of the highest-achieving students was never able to free all 1360 pets.

<u>Integration of the Game into Formal Education Settings</u>

Although the teachers we spoke with were generally enthusiastic about the game (several teachers mentioned how much they enjoyed it, and in fact, only one did not play the game with the students) most of them did not integrate the game into math lessons. Most of the teachers used the game as an enrichment activity, or some other more casual activity, such as during study hall time or free time. Teachers did not regularly use the game in math lessons, though some of them did explain the math concepts more informally (for example, if they knew students were working on proportions in math class, they would relate this to proportional tasks in the game). One teacher explained that she taught a few mini-lessons around the Cafeteria and Recipe Room puzzles because they were covering ratios and proportions in class. Another teacher, who was teaching science that semester, explained,

> *I did a little bit of math lessons, like in my science class, but it wasn't directly related to the curriculum I was teaching. But when a number of kids were stuck on the same puzzle I would actually go in and use, um, use that as a way of introducing, or basically trying to tie what they were doing in to their math class.*

This quote is illustrative of the way several seemed to have used it. They did not necessarily integrate it into standards-based math lessons, but were able to explain things to the students mathematically. Of course, as the same teacher later mentioned,

> *I personally love [the game]. I think it's great [laughing]. It's funny because I think a lot of the math skills that are in there, they're embedded deep enough so a lot of kids don't think they're doing math. They think they're playing games and puzzles and they actually are doing some pretty deep thinking in their math so I think it's a really nice effective tool.*

This typifies what many teachers seemed to like about the game, and why so many were so enthusiastic. They could have students playing the game and applying math concepts while the students felt like they were having fun and being rewarded. Another teacher points out that the kids appreciated the game because as they explained to her, "You're meeting us at our level!"

<u>Summary</u>

As mentioned above, teachers used the game somewhat differently in their classes and there were different types of settings in which they used it. Some of the teachers were more active than others on the messaging system, as were some of the students. Most of the teachers monitored the message boards, though. Teachers generally used the administrator tool to monitor students' progress, but it was not necessarily the case that they used what they learned to intervene and help students with the game, although some teachers did either directly or by pointing students towards other students who could help them. Despite the many different ways the game was integrated into the classroom by the teachers, there were some consistencies across those varied settings. All of the teachers were enthusiastic about the game, and many had played before or plan to continue to use it. It is important to point out, however, that this enthusiasm could be a self-selection bias, meaning that those teachers who were most enthusiastic were the ones most willing to talk about the game with their students.

Many of the teachers seemed to have good relationships with their students as well. Several teachers mentioned how good their classes were, or how much they trusted the students. The students also seemed to be willing to ask questions both in class and outside of official class time. One teacher mentioned that students would come use the computer in the classroom to play the game while they were waiting for the bus. Teachers that had successful teams also showed a willingness to "meet them at their level," by asking the kids for help when they ran into problems. This likely made kids feel, to some extent, like experts in this realm, giving them more reason to want to succeed.

The most successful teams then, while they did not necessarily share many traits across classrooms, seemed to all be within classroom settings in which the teacher showed a lot of interest and enthusiasm in the game. Further they seemed to be in classes where there was a fairly open relationship between students and teachers, with teachers being willing to explain things to the students, to work alongside the students to solve puzzles, and to ask them questions about how to solve the puzzles, perhaps providing students with feelings of mastery and expertise that increased their interest in the game. This suggests that regardless of the way that teachers implement the game, that the more likely key to success is having teachers who are enthusiastic about the game and willing to use it as an opportunity for learning and collaboration *with* their students.

Conclusions

The *Lure of the Labyrinth* presented its middle school-aged players with an opportunity to engross themselves in a fun, whimsical world, in which they had to rely on their math- and logic-based problem solving skills in order to succeed. Over 700 teachers participated in the *Lure of the Labyrinth* Challenge; through these teachers over 13,000 students were reached.

As indicated by the teacher survey data, the teachers who participated in the Challenge clearly felt comfortable incorporating technology into the math classroom and saw the value of doing so. This finding is underscored by data from the teacher interviews showing that the most successful teams were lead by teachers who were enthusiastic about the game, who were excited to share it with their students, and who were able to foster a climate of collaboration in their classrooms. Importantly, this climate of collaboration was only minimally evident in our examination of the message board posts. However, this is likely due to the fact that students had myriad ways to communicate with each other (e.g., face-to-face, cell phone text message, Skype). Aside from being more familiar to students (and perhaps more efficient) these modes of communication also had the added benefit of being kept private from teachers. Given these things, the message board posts likely provided only a small snapshot into how the teams communicated. As such, it is likely that the *Lure of the Labyrinth* achieved its goal of fostering *Deeper Thinking* among students; it is just that with the limited message board data we are unable to fully capture the extent of this collaboration.

Examination of the student game play data revealed a fair amount of variability in students' persistence on individual puzzles and in their engagement with the game over time. The fact that students were willing and able to increase their effort to accommodate increasing puzzle difficulty shows 1.) that students were interested enough in the game to take the time to be more purposeful in their playing and 2.) that students were able to successfully recognize when they needed to adapt or develop a new strategy for completing a more difficult level. The patterns of engagement observed suggest that student engagement may be bi-modally distributed, with some students dropping off primarily during levels that are unlocked earlier in the game and other students persisting through to levels unlocked later in the game.

With regard to student performance across subsequent puzzle attempts, results suggest that when presented with a puzzle of minimal difficulty, students were able to make consistent progress as they made repeated attempts. However, as puzzle difficulty increased, students' performance over time tended to plateau or decrease. Although these data can be considered general descriptions of the types of patterns of play students exhibited over time, they should be interpreted cautiously, for reasons described in the next section.

Finally, when attempting to identify the characteristics of the most successful teams in the Challenge, there was revealed to be a great deal of heterogeneity in how the game was implemented, the settings in which it was used, and how it was used. However, as mentioned above, the most salient, unifying characteristic that the top-performing teams seemed to have in common, was having a teacher who was enthusiastic about and interested in the game.

Limitations of the Current Study and Methodological Considerations for Future Research

The current research was conducted in a naturalistic setting, with the game being primarily implemented as it would have been by teachers in the absence of the research, with the exception of the added Challenge component. While there was great value in seeing how the *Lure of the Labyrinth* fared in the real world, with very few requirements regarding data collection and reporting, this also was the biggest limitation of the study. As described throughout the report, there was a great deal of heterogeneity in the settings in which the *Lure of the Labyrinth* was implemented and in how it was implemented. Although the results of our qualitative interviews strongly point to teacher enthusiasm being an important factor, there are likely other aspects of the implementation that were more or less effective but that were not captured by the interviews. To truly understand what components were most effective about teachers' implementation of the *Lure of the Labyrinth*, in future research there should be some basic data reporting requirements for teachers. For example, collecting minimal background information regarding teacher background and academic setting (e.g., if they teach in a typical math classroom or in a computer lab) in order to examine differences in student performance as a function of it.

Regarding the message boards, in order to truly understand how teams collaborated, it would be necessary in future research to follow a handful of teams carefully, not only observing their message board activity but also interviewing them to learn about other modes of communication, holding focus groups with teams to hear from them as a group to understand the team dynamic and also conducting classroom observations to understand the extent to which face-to-face communication (both within teams but also between teachers and teams) facilitated game play. Although the data do suggest some association between substantive use of the message boards and in-game success, the fact that in general, message board use didn't predict in-game team success supports the notion that there was a great deal of communication and collaboration happening among teams that was not captured in the message board posts.

Considering student persistence and engagement with the game, although we were able to ascertain some general tendencies of student players, there was a great deal of variability in these measures. Future research should focus on identifying student- and teacher-level characteristics that explain this variability (e.g., teacher education, student math achievement, students' attitudes toward math, etc…). Doing so would serve to inform future implementation of the *Lure of the*

Labyrinth as well as other games in which student persistence and engagement are of interest to game developers and educators.

Finally, with regard to student learning data, it should be noted that the general patterns of performance depicted are not necessarily the best indicator of student learning. As noted above, these data are presented without consideration of the minimum proportion of correct moves needed to complete the individual puzzles, so what appears to be declining performance (e.g., in the more difficult puzzles) still may be sufficient to succeed in beating the puzzle. Further, proportion of correct moves does not take into consideration the number of correct moves that would be expected simply due to chance. This is an important point because a puzzle in which students average 50% correct moves on the first attempt, but for which there are only two moves, makes a performance of 50% somewhat unimpressive because it is equal to chance. In contrast, a puzzle in which students average 50% correct moves on the first attempt, but for which there are 7 possible moves, suggests greater accuracy. Given these limitations, these data should be looked at only as general representations of patterns of performance over subsequent puzzle attempts. They should not be considered in relation to in-game performance, nor should a decrease in the proportion of correct moves over time be necessarily considered evidence of a lack of student learning.

References

Martin, T., & Schwartz, D. L. (2005). Physically Distributed Learning: Adapting and Reinterpreting Physical Environments in the Development of Fraction Concepts. *Cognitive Science*, 29(4), 587–625.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, CA: Sage Publications.

Reid, J. (1992). The effects of cooperative learning with intergroup competition on the math achievement of seventh grade students. (Research Report). (ERIC Document Reproduction Service No. ED 355 106).

Steinkuehler, C., & Duncan, S. (2008). Scientific Habits of Mind in Virtual Worlds. *Journal of Science Education and Technology*,17(6), 530-543.

Swing, S., & Peterson, P. (1982). The Relationship of Student Ability and Small-Group Interaction to Student Achievement. *American Educational Research Journal,* 19 (2), 259 – 274.

Appendix A

Teacher Interview Protocol

1) How did you decide which students would be teamed together?
Was it random? Did you purposefully choose to put stronger math students with weaker math students? Did you put friends together? Were they all in the same class?

2) How did you explain Labyrinth to the students in your class?
Did you explain that it was a math game? Did you explain that it was a contest and explain the prizes?

3) How did you use Labyrinth in classroom? Did you discuss the game in the classroom? Did you use it in math lessons? If so, how? If not, why not? Did you ask how the students were progressing in the game?

4) Did you play the game? How many times? Regularly?

5) Did you moderate the message boards? Did you delete a lot of inappropriate material?

6) Did you post on the message boards? If so, did you make it clear that you were the teacher? Did you pose as a kid? What types of posts did you post on the message boards? Encouragement? Tasks? Reminding them to behave? Asking for help?

7) When students were playing the game, what did that typically look like? Were they all playing at the same time in a computer lab at separate computers? If so, was there a lot of talking between the students when they were playing? Were kids sharing computers?

8) Did you use the administrator tools to monitor students' progress during the game? Why or why not? If so, did it guide your behavior in any way e.g. did you spend more time with students on certain concepts?

Technical Appendix

For each of the nine puzzles examined, a three-level hierarchical linear model was used. The first level of the model, the student level, there were multiple (5) puzzle attempts over time nested within students. In the second level of the model, there were students nested within teams. Finally, in the third level of the model, there were multiple teams nested within teachers. In each of these models, the outcome measure of interest was the proportion of correct moves on each puzzle attempt.

Preliminary models included random intercepts at level- 1, 2 and 3 and random slopes at level-1. However, across all models, none of the random slope effects was statistically significant, and therefore these were trimmed from final models. In some cases, particularly on the more difficult levels, the level-2 and level-3 intercepts were not significant. However, these non-significant intercepts were retained in the models to allow a hierarchical model to be fit to the data and to allow for a consistent set of analyses across the levels. Regarding change over time, in each preliminary model a linear, quadratic and cubic growth term was included in the level-1 model. All non-significant growth terms were trimmed in subsequent models. In Tables 1 through 9, the fixed and random effects coefficients associated with each of the 9 final models are reported.

Table 1. Employee Cafeteria

Fixed Effect		*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}		0.50	0.004	99.74
Average learning rate, γ_{100}		0.04	0.002	17.45
Average learning rate squared, γ_{200}		-0.004	0.001	7.97
Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.02			
Level 2, r_{0ij}	0.005	1842	6248.13	<0.001
Level 3, u_{00j}	0.01	403	1679.04	<0.001

Table 2. Manager's Cafeteria

Fixed Effect		*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}		0.60	0.01	57.44
Average learning rate, γ_{100}		-0.06	0.01	4.76
Average learning rate squared, γ_{200}		0.03	0.01	4.24
Average learning rate cubed, γ_{300}		-0.004	0.001	3.38
Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.04			
Level 2, r_{0ij}	0.02	890	3886.17	<0.001
Level 3, u_{00j}	0.01	191	473.05	<0.001

Table 3. Executive Cafeteria

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.75	0.01	91.21
Average learning rate, γ_{100}	-0.004	0.002	1.98

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.02			
Level 2, r_{0ij}	0.02	329	1860.92	<0.001
Level 3, u_{00j}	0.0001	108	114.28	>0.05

Table 4. Testing Lab #1

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.51	0.001	92.05
Average learning rate, γ_{100}	0.02	0.003	6.68
Average learning rate squared, γ_{200}	-0.001	0.001	2.37

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.03			
Level 2, r_{0ij}	0.01	1763	6892.96	<0.001
Level 3, u_{00j}	0.01	344	1356.38	<0.001

Table 5. Testing Lab #2

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.48	0.01	65.70
Average learning rate, γ_{100}	-0.08	0.01	7.30
Average learning rate squared, γ_{200}	0.03	0.01	3.60
Average learning rate cubed, γ_{300}	-0.003	0.001	2.61

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.05			
Level 2, r_{0ij}	0.02	708	2211.82	<0.001
Level 3, u_{00j}	0.002	172	239.43	<0.001

Table 6. Testing Lab #3

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.38	0.01	30.00
Average learning rate, γ_{100}	-0.09	0.02	4.85
Average learning rate squared, γ_{200}	0.04	0.01	3.59
Average learning rate cubed, γ_{300}	-0.01	0.002	3.197

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.03			

Level 2, r_{0ij}	0.02	195	982.85	<0.001
Level 3, u_{00j}	0.0002	83	100.23	>0.05

Table 7. Employee Lounge

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.68	0.005	143.38
Average learning rate, γ_{100}	0.06	0.005	12.43
Average learning rate squared, γ_{200}	-0.01	0.003	5.55
Average learning rate cubed, γ_{300}	0.001	0.0004	3.34

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.021			
Level 2, r_{0ij}	0.005	1574	4779.97	<0.001
Level 3, u_{00j}	0.002	279	668.74	<0.001

Table 8. Manager's Lounge

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.77	0.001	123.29
Average learning rate, γ_{100}	-0.01	0.002	5.11

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.03			
Level 2, r_{0ij}	0.02	627	2638.94	<0.001
Level 3, u_{00j}	0.0002	159	170.28	>0.05

Table 9. Executive Lounge

Fixed Effect	*Coefficient*	*se*	*t Ratio*
Average initial status, γ_{000}	0.27	0.01	27.45
Average learning rate, γ_{100}	-0.15	0.02	7.31
Average learning rate squared, γ_{200}	0.04	0.01	3.721
Average learning rate cubed, γ_{300}	-0.05	0.002	2.42

Random Effect	*Variance Component*	*df*	χ^2	*p Value*
Level 1, e_{tij}	0.03			
Level 2, r_{0ij}	0.01	342	815.73	<0.001
Level 3, u_{00j}	0.0003	121	151.55	<0.05