Reflecting on Educational Game Design Principles via Empirical Methods

Osvaldo Jiménez, University of the Pacific, 3601 Pacific Ave, Stockton CA, ojimenez@pacific.edu

Abstract: In designing environments for children to use for learning, there are many design decisions that are made by the game's creators that can affect their effectiveness. One such example that many creators determine is the amount of story to embed in an educational game. To learn more about the importance of story in educational games, 77 fourth grade students in one elementary school were randomly assigned to play one of three versions of an educational video game with varying levels of story. The goal of the videogame is to give students practice with fractional-whole operations. In addition to logging interactions students made during gameplay, pre and post-tests that capture students' fractional knowledge in a classroom environment are reported and discussed. Results indicate that while story may not seem to be a critical factor in improving learning, its benefits and impact on learning may be more nuanced and complex.

Introduction

With one of the conference aims being about learning and becoming in practice, and foregrounding ways that learning processes are situated in practice, the research reported herein contributes to better understandings and practices involved in children's engagement with educational games. There have been many commercial and educational video games that have been studied or created by scholars. These scholars have shown that the games have diverse benefits, such as increased self-efficacy (Nelson, Ketelhut, Clarke, Bowman, & Dede, 2005), authentic scientific inquiry, (Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007), and learning to use evidence to warrant claims (Steinkuehler, 2006). However, the creation of educational video games, merely in viewing it as engineering a complex software program, is both an expensive (Squire, 2003) and difficult endeavor to undertake (Naone, 2007). While some games have shown learning differences, (Barab et al., 2007, Jiménez, Arena, & Acholonu, 2011), there are relatively few studies demonstrating the learning benefits of educational games using traditional measures (Honey & Hilton, 2011).

Because educational video games are so difficult to make, researchers and proponents of making games ponder how to create more effective educational video games. One way is to better understand the impact that certain aspects or features of educational games have in helping students learn. One area where variable amounts of resources are often placed is in a game's story and story structure. Stories have many links to learning that may be important for games, ranging from promoting role-playing (Williamson & Slivern 1991), to influencing people's judgments (Strange & Leung, 1999), to impacting individuals' comprehension and memory (Mandler & Johnson; 1977; Thorndyke, 1977).

Studies have demonstrated a potential link between stories and learning as well as links between stories and educational games. Researchers have demonstrated links between having a story in an educational video game and children's motivation to work on educational content (Malone, 1981; Parker & Lepper, 1992; Cordova & Lepper, 1996, Barab, 2007). Others have argued for a more refined link between stories and video games, arguing against some of this same research, which suggests that developers be very careful about integrating such stories/fantasy contexts with a game (Habgood, Ainsworth, & Benford, 2005). Scholars have further argued that context is largely irrelevant stating that stories chosen for their own games could easily be replaced with something else (Habgood & Ainsworth, 2011, p.176). If this is the case, story could be absent from a game altogether or could be replaced with something unappealing. Determining whether a story is important to an educational game informs game designers' future development and can help create better educational experiences for students. To further investigate this factor, one approach is to research an educational game that can be modified to have different amounts of story in the game, and use results to make decisions on how to change a game to have an efficacious amount of story.

Does Story Matter in Educational Games?

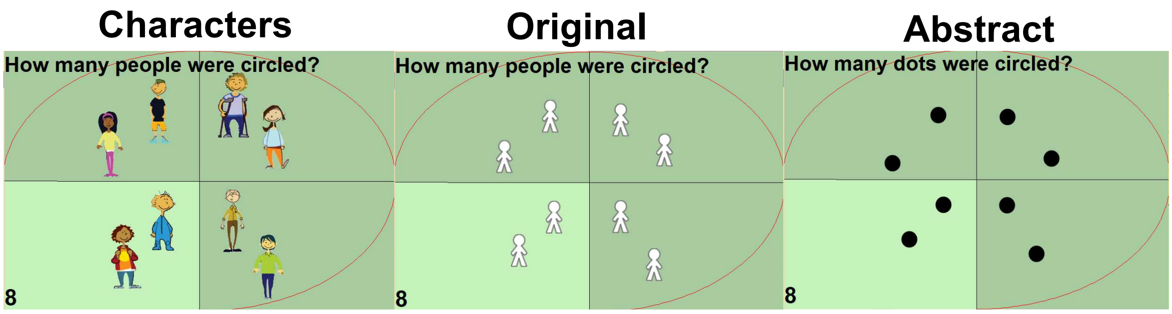
This paper reports on this type of inquiry. A computerized fractions card game that demonstrated increases between pre and post-test classroom measures (Jiménez, Arena, & Acholonu, 2013) was borrowed and used to create three versions. The three versions leveraged the game and story used in the aforementioned research. The game was appropriate for this kind of study because it dealt with a difficult learning domain, fractions. Fractions is one of the most difficult concepts for children in elementary education (Petit et al., 2010) and also one of the first concepts that causes children to stray away from a STEM pathway (Nunes, 2006).

Past research was leveraged to select and vary the critical components in each experimental condition. For example, prior studies on the use of characters and their impact on story recall and comprehension (Bower, 1978; Thorndyke, 1977) led the authors of this study to select characters as a feature that could be varied across conditions. Characters such as protagonists are a critical component of a story and a feature that could be modified to change the amount of story. Therefore one version, the *Characters version,* places more prominence on the characters in the game. Another version, the *Abstract version*, is completely devoid of any story structure, leveraging story grammars to take out all critical components, including the use of concrete characters. The final version of the game, which I refer to as the *Original* version, mirrors closest to what students played in the original fractions game cited. Nonetheless, all three versions share the same underlying code base. Only the text and images used in the software differ between them.

In the Characters version of the game, care was taken to promote and embellish the story by making the characters more salient for children. That is, rather than having generic characters as the manipulatives, the students worked with the characters present on the cards throughout the game. Figures 1 and 2 detail the differences present in the three conditions across the most important game screens.

<u>Figure 1</u>: The three conditions on the card playing screen: the Characters condition (left), the Original condition (middle) and the Abstract game condition (right).

This Characters version is displayed on the left of both Figure 1 and Figure 2. The only crucial difference is that the manipulatives in this condition were made so that they would look exactly like characters that they represented in the cards. For example, rather than seeing eight general figurines which are used to represent everyone, each figurine shown in the game would correspond to one of the members of the Johnson Family shown on the card. These figurines would also be shown during the calculation and would also be shown as *stinky* or *fresh*. Now that each character is salient, when players select characters to become "stinky", those specific characters would also be stinky on the other screen. The other conditions also give the students the opportunity to select which manipulative that they want to become stinky; however, one would think that choice would become more trivial since all of the manipulatives look alike.



<u>Figure 2</u>. The three conditions after having provided help on answering ¾ of four. The Characters condition (left), the Original condition (middle), and the Abstract game condition (right)

In the Abstract condition, which is shown on the right of Figures 1 and 2, one notices that the game has been completely stripped of the story structure. Therefore, all of the *Tug-of-War* story images were replaced with abstract images. When replacing the images and text, careful consideration was given so that the software would not be written any differently in the Abstract condition. Only the images and text displayed to the user were changed. The people changed to dots, both in images and language; cards like the stink bomb were changed to be called "Reduce" cards. All of the new animations were still kept in the game, except different images were used to explain what happened. All of the story elements were replaced by abstract representations so as to make the game playable without any insinuation of story. Taking the literature on story into consideration, one might hypothesize that the story conditions would be able to capitalize both on helping the user learn as well as engaging more with the material. However, it could be that traditional game mechanics are powerful enough for students to drive the rest of the learning and memorization in the game.

There are researchers who propose that the abstract is better, citing articles that argue against using idealized images and for using lines and shapes for student work (Goldstone & Wilensky, 2008), in learning applications such as agent-based modeling (Wilensky, 2002). There are also arguments against using a version of the game where the characters are more salient. The argument against characters may be that the characters become so salient that they distract the players from the content itself (Son & Goldstone, 2009). These complex nuances play out in games and need to be further discussed and researched. The goal of the experiment described below was to understand and detail the roles that story plays in games and how they impact and interplay with games. In order to explore this impact, a protocol was employed that assigned children to play one of the three different versions of the game.

The Study

The study was conducted once the versions of the game were developed. Students were exposed to one of three versions of the game. Afterwards, they were compared on how much they had learned and how much fun they had. The method and results of that study are presented below.

Participants

77 fourth grade students participated in this study. All of these students attended one elementary school in the San Francisco Bay Area. Over 90% of the students in the study were classified as English language learners. These students were a part of classrooms in the same school, although one of the classrooms was a mixed fourth- and fifth-grade classroom. On a previous standardized benchmark test that the students took, only 27% of students tested well enough to be classified as having mastered the mathematics content taught so far at their school, with 20% struggling to learn the material. The remaining 53% were assessed as making progress towards mastery.

Materials

Both the pre- and new post-tests were an expanded version of the tests used in earlier studies of the fractions card game. This test expanded on questions that relied more heavily on students' understanding of part-whole relationships (such as "what is ½ of 8" and less on their understanding of chance, which is known to be a very difficult subject to learn (Garfield & Ahlgren, 1988). The questions that were devised in this test belonged in one of five categories. These categories were questions where students had to compare fractions to one another, questions involving decimals, complex word problems, and fractions questions. The fractions questions were further broken down into two categories: fractions they had most likely seen in the game and fractions they could not have encountered in the game. The maximum score that could be obtained on the test was 41. Most questions were given one point; however, large problems were divided into sub-problems, and each of those sub-problems was given a point.

As mentioned previously, the post-test added a fun survey as well as questions about the game to the end of the measurement. I also added additional questions about the story and playing the game. Due to student absenteeism, 73 out of 77 students completed both the pre and post-tests.

Method

Rather than assigning each classroom to a game condition, all 77 students were randomly assigned to a condition. This random assignment was stratified by previous performance on a standardized test taken earlier in the year, and controlled for their performance so that no condition had a significant difference in standardized test scores. To accommodate wishes expressed by parents about their children not being videotaped, some of the students were switched out of conditions with other students who had equivalent standardized test scores.

Once students were assigned to a condition, they were then ranked by their previous performance and randomly assigned by rank to a playing group, which I call a *pod*. This was done to ensure that all of the pods were of mixed academic ability. Once in a pod, students were then placed in mixed-ability pairs since mixed-ability pairs have been shown to help with student learning in math board games (Guberman & Saxe, 2000). The only indicator that was used for these mixed-ability groups was a standardized test score that the students had taken earlier in the year. Each condition played in one of the three classrooms, which meant that some students would play in the classroom where they had instruction. To avoid having any effects based on being in the same rooms or having the same partners, students would also switch classrooms every two weeks over the five sessions. This ensured that all students played in their classrooms at least once. When they would switch classrooms, students were also given new partners and new play groups, in another random assignment, based on their standardized test scores. Subjects switched partners and pairs to limit any particular subjects from having a bad (or good) partner throughout the study. In instances where the random assignment left a student with the same partner they had earlier, partners were then manually reassigned.

The teachers administered a pre-test assessment in their classrooms. Pre-tests were followed by one hour a week sessions in the game condition for five weeks. The students started each session in their original classrooms, and then were led to their new classroom based on condition. Once in the new classroom, students were assigned partners. Students then played the game. Playing time varied from 35-45 minutes during each session, with student pairs playing against each other.

After each session, the classroom researchers met and discussed problems that they noticed children were experiencing with the software. Simple changes to the software were made each week to help assuage any concerns or problems that had arisen. Some changes made mainly improved the language in the game, such as shortening messages given to students. Other features made the administration of the game easier, such as preventing students from starting games with other groups. Each week, the single software code base would be updated and the images and text would be replaced to create the three versions. Students would then play with the latest versions of the software.

Students played with random cards during the first four sessions, which meant they played the game as intended. In the fifth session, rather than giving everyone a random set of cards, each pod was given a fixed deck of cards arranged in the same order on that last day. While the students thought the cards were random, they were manipulated to be the same for everyone. This was done in order to gather enough data to help answer a few questions about students' gameplay. A few of the students noticed and told researchers that they had received the same set of cards as before when their games had to be restarted. However, when the researchers were asked if they felt the students knew the cards were not random, and rather the same for everyone, they answered "No". In addition, on the fifth session researchers videoed two pods during their gameplay, one from the Characters condition and one from the Abstract condition. Two cameras were used to record each pod from two different angles. Video from each pod is approximately 30-45 minutes in length.

One week after having finished the five sessions of playing the game, the students were given a post-test by their classroom teachers in their original classrooms. This post-test looked exactly like the pre-test, except it added a few items at the end which were meant to assess the amount of fun that children had with playing the game, and their experiences with the game.

Results

Table 1 below shows the average gain in scores for each condition, with the standard deviation of those scores in parentheses. Average gain is reported as the difference in student scores from pre-test to post-test (e.g. post-test score – pre-test score). When collapsed across all conditions, students had an average gain score of 5.5, which indicates about a 13% improvement. Moreover, the 5.5 gain score from pre-test to post-test was significant when compared to zero, $t(72) = 7.86$, $p < .001$, $d = 1.85$, with a large effect size. Regardless of condition, student's scores on average increased from pre-test to post-test.

Table 1: Average gain scores from pre to post by condition, (standard deviation in parenthesis)

	Abstract Group	Original Group	Characters Group
Gain Score	4.00 (5.35)	5.59 (6.74)	6.89 (5.87)

A cursory examination of the gain scores in Table 1 shows a small increase in the directions that I expected. The Characters group had a higher average score than the Original group, who in turn had a higher average score than the Abstract condition. A one-way ANOVA with planned comparisons between the Abstract and Characters conditions and Abstract and Original conditions was performed to investigate whether or not those differences were significant by condition. The results from that ANOVA show that the small increase by condition is insignificant $F(2, 70) = 1.479$, $p = .24$. The planned comparison between the Original and Abstract conditions also was not significant $t(45) = .9$, $p = .37$. However, there was a marginal difference between the Character and Abstract conditions $t(50) = 1.72$, $p < .09$. The marginal difference between the Abstract and Characters conditions led to analyzing test scores by type of test questions, which were devised *a priori*. To reiterate, the five types of questions on the assessment were: questions that involved comparisons, questions involving decimals, questions involving complex word problems, questions involving fractions students would have seen in the game, and questions involving fractions students would not have seen in the game. I refer to each of these types of questions as Comparisons, Decimals, Word Problem, Old Fractions and New Fractions respectively.

All questions were categorized into one of these five groupings, and a subscore was calculated for each of the five categories for each student. The averages of these subscores along with the standard deviations for each condition are reported in Table 2. The table's first three subscores (Comparison, Decimal and World Problem questions) show almost no gains and little difference by condition. Nevertheless, there seem to be trends in the Old Fraction and New Fraction questions in line with trends seen on the overall post-test. A one-way MANOVA then was run to determine if these trends were significant by condition. The MANOVA determined that the overall fit of group to these gain scores was not significant $F(2, 70)=.90$, $p=.54$.

Table 2: Sub-test gain scores by condition (standard deviation in parenthesis)

	Abstract Group	Original Group	Characters Group
Comparison gain	0.83 (1.66)	0.41 (1.71)	0.41 (1.80)
Decimal gain	0.38 (0.77)	0.68 (1.13)	0.56 (1.12)
Word Problem gain	0.63 (1.01)	0.68 (1.78)	0.78 (1.12)
Old Fraction gain	1.58 (3.20)	2.32 (3.56)	2.70 (3.29)
New Fraction gain	0.58 (2.10)	1.50 (2.74)	2.44 (2.39)

The next step was to look at the univariate scores for Old Fractions and New Fractions. The univariate test demonstrates that Old Fraction gain subscore by condition was not significant $F(2, 70) = .73$, $p = .49$. Nevertheless, the New Fraction gain subscore does have a significant difference by condition $F(2, 70) = 3.78$, $p < .03$. Performing a planned comparison on this gain demonstrates that while the difference between the Original and Abstract conditions was not significant, $t(45) = -1.29$, $p = .20$, there was a significant difference between the Abstract and Characters condition with a medium effect size $t(50) = 2.75$, $p < .01$, $d = .78$. The Characters condition had a higher average gain on the new fractions when compared to the Abstract condition.

Investigating the Relationship between Learning and Fun across Conditions

In addition to investigating the learning, it was also important to investigate the association between fun and learning across all three conditions. Figure 3 displays the gains that individual students had on the overall fractions test by condition and the gains that students had on the New Fractions subscore. The height of each digit on both graphs represents the gain that a particular student had from the pre and post assessments. The particular digit represented on the graph is that particular student's fun score. Upon an initial examination of both graphs, the Characters condition exhibited a pattern of having higher fun scores associated with higher learning gains. Nonetheless, this pattern does not seem apparent in the Abstract or Original conditions.

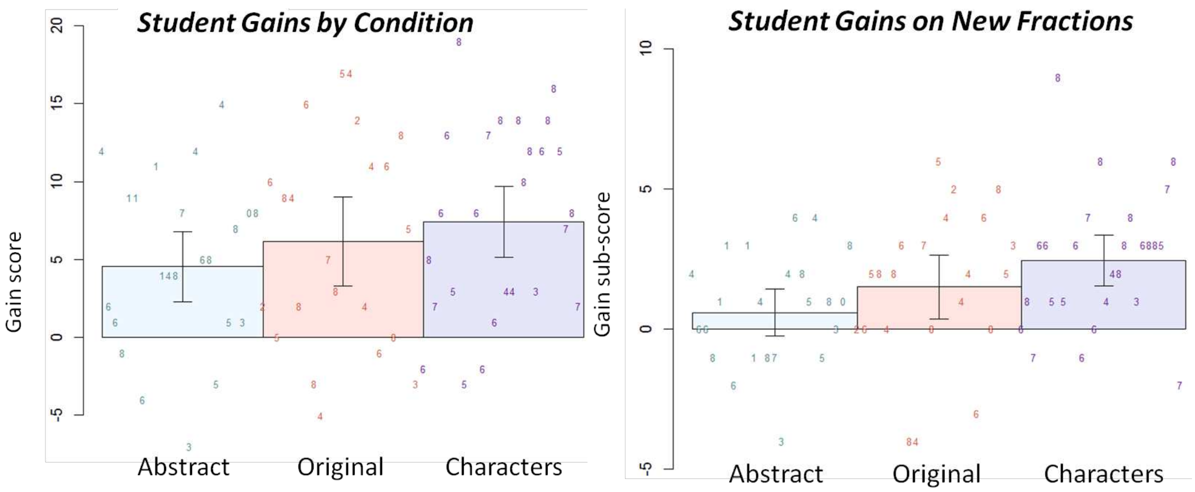


Figure 3. Student Gain Scores by Condition (Overall: left, and New Fractions: right). The height of each digit represents the gain score for a particular student, and the digit used represents the student's fun score rating

Table 3 below gives the correlations between the gain score (or subscore) and students' reported rating of fun. Positive correlations would provide evidence that there is a positive association between fun and learning. In a positive correlation, students who reported having more fun would be linked to stronger gains from pre-test to post-test. The correlation table does not display a positive correlation for either the Abstract or the Original conditions, but rather a negative one. Furthermore, the Abstract condition has a stronger correlation than the Original condition, but in the *negative* direction. Nonetheless, the Abstract condition exhibited no significant correlations in either the fraction gain scores $z(22) = -.89$, $p = .37$, or the sub-gain scores $z(22) = -.68$, $p = .50$, which means that there is not a likely association between the gain score and fun rating. Because the Original condition was closer to zero than the Abstract condition, no significance tests were performed.

Table 3: Correlations of fun ratings with gain or new fractions sub-gain scores by condition

	Abstract	Original	Characters
Correlation (r) of Overall Gain with Fun	-.14	-.08	.45
Correlation (r) of New Fractions with Fun	-.11	-.09	.34

While the Abstract and Original conditions did not exhibit any correlations, there does seem to be a positive correlation between students' self-reported ratings of fun and the gains that they made from pre-test to post-test in the Characters condition. The Characters condition exhibits significant positive correlations between the fun ratings and gain scores $z(25) = 2.92$, $p < .01$ as well as the New Fractions sub-gain scores $z(25) = 2.17$, $p < .03$.

The fact that there exists a positive correlation in the Characters condition, but no correlation in the other two conditions warrants further review. To make sense of what was happening, the following section details analysis of in-game measures during gameplay.

Investigating In-Game Measures

In addition to measures taken outside of the game, there were also internal measures that were created and stored in the software. These measurements were logs detailing activity from each user, such as when a question was answered without first getting it wrong and whether or not a student asked for help. To determine if the play by students differed by condition, a MANOVA was performed to investigate any differences by condition on several measures that were determined to be of possible use in earlier studies, such as the number of times they asked for help, the number of questions solved, and the average amount of time students took to answer a question. Performing the MANOVA demonstrates an overall fit of the data to the general model $F(2, 70) = 1.99$, $p < .01$, which would indicate that the type of play students exhibited seemed to differ by condition. Looking at the univariate scores, however, one notices that there is only one variable that carries a significant difference, the average number of questions a student would answer correctly on their first attempt in the last session $F(2, 70) = 4.72$, $p < .02$. By the last session, the effects that students have in playing the game should be different, which may explain the differences reported in Table 4 below.

Table 4: Select in-game results by condition

	Abstract	Original	Characters
Number of 1st Attempts Correct in Last Session	6.00 (2.80)	8.32 (4.34)	9.15 (3.95)
Number of Incorrect Moves	8.13 (3.46)	7.86 (4.89)	7.59 (3.21)
Percentage of Incorrect Moves to Correct Moves	24.7 (9.1)	24.4 (9.1)	20.4 (6.6)

Performing a planned comparison between the Abstract and Original conditions yields a significant difference between them, $t(45) = 2.10$, $p < .05$, as well as when comparing the Abstract condition with the Characters condition $t(50) = 3.00$, $p < .01$.

One line of reasoning about the Characters condition or story conditions was that if the students understood the story and its rules, they would be less likely to make errors. If they were less likely to make errors, then the software would have logged less of those errors on average. Table 4 above reports the average amount of times that a student would incorrectly make an incorrect move across all sessions. The MANOVA's results displayed no significant differences across the three conditions $F(2,70) = .12$, $p =.88$. However, thinking about the question critically, solely investigating the number of incorrect moves was not as appropriate a measure as the *percentage* of incorrect moves to total moves, as students who played more would naturally make more incorrect moves. Analyzing how this measure differed by condition yielded a larger F-statistic but demonstrated no significant difference $F(2, 70) = 2.13$, $p = .13$. Performing planned comparisons between the Original and Abstract conditions yielded no significant difference $t(45) = .12$, $p = .90$; however, there was a marginal difference between the Abstract and Characters conditions $t(50) = 1.84$, $p < .07$. There is a marginal trend towards students having less of a tendency for making illegal moves. Based on the earlier analysis that correlated student's fun with learning, it may be that interest is driving students to make fewer errors. When we look specifically at whether such a measure correlates with the amount of reported interest, we do get correlations in the correct direction for both the Original ($r = -.26$, $t(20) = -1.18$, $p = .25$) and Characters conditions ($r = -.17$, $t(24) = -.82$, $p = .42$), though neither of these are significant, while the Abstract condition shows no correlation between the percentage of incorrect moves and the amount of interest that students have ($r = 0$). While the in-game measures have generated insight into the student's behaviors in the game, the data logging techniques did not catch any of the social behaviors that the children exhibited *while playing the game*. In the future I hope to analyze and report qualitative findings generated from the video captured in the Abstract and Characters conditions.

Discussion

The main question this experiment sought to answer is whether or not story is a necessary feature in developing effective educational games. Exploration of the experimental study's results indicates a complex scenario for the role that story may play in educational video games. The initial results seem to indicate support for arguments indicating that story may not be necessary. This conclusion could be inferred based on the experimental results that demonstrated that students increased from pre-test to post-test regardless of condition.

While the initial results from the final study could be used to argue that students can learn from a game without story, it is sensible to recognize the marginal differences that were present overall between the Abstract and the Characters conditions. In particular, the Characters condition displayed marginal trends in having larger gains from pre-test to post-test when compared to the Abstract condition. These results, combined with the high gains across all conditions, lay the foundation for a general viewpoint about the role of stories in educational games: a story may not be necessary for creating a successful educational video game, but its presence may help in increasing student's gains from pre to post.

The role of a story becomes clearer with respect to learning when looking specifically at the types of questions asked on the tests. Once the pre and post-tests were broken down by the type of fraction question, the results indicated that most of the types of questions did not differ by condition. Nevertheless, the results did indicate that one question category did differ by condition – new fractions questions. Students in the Characters condition had significantly higher gains when answering new fractions questions when compared to students in the Abstract condition. This means that students in the Characters condition had higher gains from pre-test to post-test when answering fraction questions they had not seen in the game, like "What is 2/5 of 15?". Not only does the story with characters marginally allow students to learn more, those students were better able to answer fractions questions they had not seen before. Because students in the Characters condition were able to answer questions they had not seen before, one may conclude that the Characters condition was better preparing students to think about fractions in the future. This finding coincides with earlier research with commercial video games and preparation for future learning (Arena, 2012). While the research presented in Arena's work establishes that video games can be used as preparation for future learning in a domains such as history, I would argue that the results from this experimental study both support Arena's research and further reveal the role that story has as being a potential critical aspect for that preparation. While game play may be strong enough to drive learning in a specific domain, having story may help students be better prepared to learn new material related to that domain in the future.

Another finding from the study was the high correlation present between the amount of fun that the students reported, and the amount that a student's score increased from pre-test to post-test. While such a correlation displayed no correlation in the Abstract and Original conditions, the Characters condition exhibited a strong correlation between fun and learning. Therefore, the results reveal a strong positive association between the amount of fun a Characters condition student had and the amount that they learned. One possible speculation for this comes from previous research done on transportation (Green & Brock, 2000). That research highlighted the role that the amount that a student was absorbed in a narrative would affect that amount that they were persuaded by such a narrative. Green & Brock's research may be one way to explain the high correlation – students who enjoyed and were more absorbed by the story were more persuaded and thus motivated to learn the material.

The results suggest that children who played in the story with characters condition got marginal learning benefits overall. Having a slightly better understanding of the role that story has in games can help the educational community think about how many resources might be devoted to the creation of a story and its incorporation into a game. If a designer does a good job in creating a powerful story that most people enjoy, that designer may have an easier time engaging students to work with the material. It may better prepare students for future learning, as well as providing them more entertainment.

References

Arena, D. (2012, June 5). Commercial Video Games as Preparation for Future Learning..

Barab, S. A., Sadler, T. D., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating Narrative, Inquiry, and Inscriptions: Supporting Consequential Play. *Journal of Science Education and Technology*, 16(1), 59–82.

Bower, G. H. (1978). Experiments on story comprehension and recall*. *Discourse Processes*, *1*(3), 211–231. doi:10.1080/01638537809544437

California Department of Education (Safe and Healthy Kids Program Office), & WestEd (Health and Human Development Department). (2012). *California Healthy Kids Survey*. Retrieved from http://chks.wested.org/resources/chks-elem-bilingual-1213.pdf

Cordova, D., & Lepper, M. (1996). Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. *Journal of Educational Psychology*, 88(4), 715.

Garfield, J., & Ahlgren, A. (1988). Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*, 19(1), 44–63. doi:10.2307/749110

Goldstone, R. L., & Wilensky, U. (2008). Promoting Transfer by Grounding Complex Systems Principles. *Journal of the Learning Sciences*, 17(4), 465–516. doi:10.1080/10508400802394898

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. Journal of Personality and Social Psychology; *Journal of Personality and Social Psychology*, 79(5), 701–721. doi:10.1037/0022-3514.79.5.701

Guberman, S. R., & Saxe, G. B. (2000). Mathematical Problems and Goals in Children's Play of an Educational Game. Mind, Culture, and Activity, 7(3), 201–216.

Habgood, M. P. J., Ainsworth, S. E., & Benford, S. (2005). Endogenous fantasy and learning in digital games. *Simulation & Gaming*, 36(4), 483–498.

Habgood, M. P. Jacob, & Ainsworth, S. E. (2011). Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences*, 20(2), 169–206. doi:10.1080/10508406.2010.508029

Honey, M., & Hilton, M. (2011). *Learning Science: Computer Games, Simulations, and Education*. Washington, D.C.: National Academies Press.

Jimenez, O., Arena, D. A., & Acholonu, U. (2011). Tug-of-War: a Card Game for Pulling Students to Fractions Fluency. In C. Steinkuehler, C. Martin, & A. Ochsner (Eds.), *Proceedings of the Games, Learning, & Society Conference 7.0* (pp. 119–127). Madison, WI: ETC Press.

Jimenez, O., Arena, D., & Acholonu, U. (2013). Educational Video Games: The Devil is in the Details. *Paper presented at the annual meeting of the AERA*, San Francisco, CA.

Malone, T. W. (1981). Towards a Theory of Intrinsic Motivation. *Cognitive Science*, *4*, 333–369.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9(1), 111–151. doi:10.1016/0010-0285(77)90006-8

Naone, E. (2005, July 12). Virtual Labor Lost: The failure of a highly anticipated, multiplayer game shows the limist of academic virtual worlds. *Technology Review*. Published by MIT. Retrieved from http://www.technologyreview.com/web/19817/page1/

Nelson, B., Ketelhut, D. J., Clarke, J., Bowman, C., & Dede, C. (2005). Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. *Educational Technology*, 45(1), 21–27.

Nunes, T. (2006). Fractions: difficult but crucial in mathematics learning. *Teaching and Learning Research Programme (TLRP) Research Briefing*. Retrieved from www.tlrp.org/pub/documents/no13_nunes.pdf

Parker, L. E., & Lepper, M. (1992). Effects of fantasy contexts on children's learning and motivation: making learning more fun. *Journal of Personality and Social Psychology*, 62(4), 625.

Petit, M. M., Laird, R. E., & Marsden, E. L. (2010). *A focus on fractions: Bringing research to the classroom*. New York: Routledge.

Son, J. Y., & Goldstone, R. L. (2009). *Contextualization in Perspective. Cognition and Instruction*, 27(1), 51–89. doi:10.1080/07370000802584539

Squire, K. D. (2003). Video games in education. *International Journal of Intelligent Games & Simulation*, 2(1), 49–62.

Steinkuehler, C. (2006). Massively Multiplayer Online Video Gaming as Participation in a Discourse. *Mind, Culture, and Activity*, 13(1), 38–52.

Strange, J. J., & Leung, C. C. (1999). How Anecdotal Accounts in News and in Fiction Can Influence Judgments of a Social Problem's Urgency, Causes, and Cures. *Personality and Social Psychology Bulletin*, *25*(4), 436 –449. doi:10.1177/0146167299025004004

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, 9(1), 77–110. doi:10.1016/0010-0285(77)90005-6

Wilensky, U. (2002). Modeling nature's emergent patterns with multi-agent languages. In *Proceedings of EuroLogo* (p. 1–6). Retrieved from http://ccl.northwestern.edu/papers/MEE/

Williamson, P., & Silvern, S. (1991). Thematic-fantasy play and story comprehension. *Play and early literacy development*, 69–90.

Acknowledgments

I would like to thank Shelley Goldman, Dan Schwartz, and Kristen Pilner Blair, whose suggestions and support have undergirded this study from its inception. I would also like to thank Dylan Arena and Ugochi Acholonu for helping create the game's initial ideas. Finally, I would like to thank the research facilitators from the AAA lab, as well as the teachers and students who participated. This work was supported by a grant from the National Science Foundation (NSF#0354453). Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.