Teaching Space Science to Middle School Students using Serious Games in a Post-COVID World

Brian Grey
Harrisburg University of Science and Technology
Harrisburg, PA, USA
bgrey@harrisburgu.edu

Dr. Sa Liu
Harrisburg University of Science and Technology
Harrisburg, PA, USA
saliu@harrisburgu.edu

Dr. Peter Hastings
DePaul University, College of Computing and Digital Media
Chicago, IL, USA
peterh@cdm.depaul.edu

Abstract: *SpacEscape* is a problem-solving mobile serious game designed to teach space and solar system science to middle school students through interactive problem solving. While our research was intended to be conducted in a formal, classroom setting, the worldwide shutdown due to COVID-19 in March, 2020 caused this plan to be no longer viable as schools went online. This provided us with an opportunity to study how effective serious games are in decentralized and unregulated situations. While we found that participation rates decreased significantly from our previous research, those who completed all stages of the study as specified (n = 17) showed a significant increase in pre- and post-test scores (one-tail t-test, df = 30 p = 0.021) as well as pre- and retention test scores (one-tail t-test, df = 32, p = 0.018). While these results seem promising, this represents an initial analysis and warrant further exploration.

Introduction

While the modern education system may be high evolved and specialized, it is not without its problems. One hurdle faced by schools is a question of motivation and maintaining student interest. Classroom settings are rigidly structured and artificial environments, designed around educational paradigms such as the didactic lecture that are now considered ineffective and outdated (Gee, 2007; National Research Council, 2015).

Using games in an educational or training setting is not a new concept, but the rise of the modern personal computer has sparked new interest in creating video games that could be used to enhance teaching and learning. The hope is that these specialized games, called serious games, could harness the motivational aspects of a traditional video game while also delivering the desired disciplinary content to the player (Djaouti et al., 2011).

There are many theories on why and how games form an effective medium for learning. Games provide a safe environment for experimentation, immediate and (frequently) unambiguous feedback, model intricate systems in explicit and simplified ways, and can be programmed to provide dynamically and individually appropriate challenge (Arnab et al., 2015; Clark & Mayer, 2007; Gee, 2007; Wechselberger, 2009; Zhonggen, 2019). These are just a handful of the reasons why games could be beneficial to learners and are sought out by traditional educators. But the greatest benefit in the view of many is the ability to motivate a student to want to interact with the content on a deeper level (Habgood & Ainsworth, 2011). By integrating the disciplinary content into the game in a way which necessitates learning the content to improve in-game performance, we can harness the intrinsic motivation of games to repeatedly expose students to content on a voluntary basis (Habgood & Ainsworth, 2011).

In this paper, we will discuss the design and goals of our serious game, *SpacEscape*, as well as the findings of our previous study using *SpacEscape* in a classroom setting. We will then discuss our most the design (and adjustments to the design) of our recent study, the preliminary findings of that study, any open questions that came out of the findings, and the next direction for our research.

Previous Work

SpacEscape is a problem-solving mobile serious game which was designed with the goal of teaching space and solar system science to middle school students through interactive problem solving. In the game, players play as Spark the Dog with the goal of finding their lost owner, Lucy, somewhere in the solar system. Gameplay is structured similarly to the classic game *Where in the World is Carmen San Diego?* in that both games require the player to visit various locations to learn relevant facts about those locations. In *SpacEscape*, these locations are spread throughout the solar system so that the player can learn relevant scientific facts about each celestial body. These bodies were chosen for their relevance and prominence in the solar system and include the sun, each of the eight planets, the dwarf planet Pluto, and significant moons of various planets and dwarf planets. The knowledge gained regarding each body is key to being able to follow the clues and to successful find Lucy.

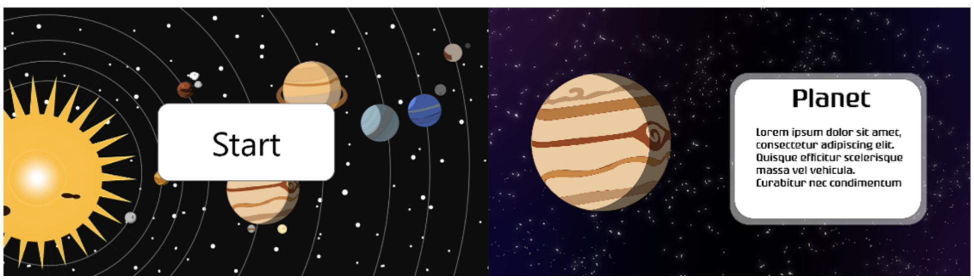

Figure 1: Planet selection screen and a detailed view of a visited planet in *SpacEscape*

Previously, we had worked with a local middle school to test our game in a classroom environment. The study, conducted in 2019, involved 228 sixth graders separated into experimental (n = 107) and control (n = 121) groups. All students in both groups had previously learned about the solar system in a content unit of their traditional science class. Students completed a 10 question pre-test on the solar system the day before the experiment. This test also contained some demographic questions about the student's gender identity and interest in learning about space. The experiment was then conducted during a 40 minute supplemental science class period where researchers were in the room to help with technical issues. The experimental group had *SpacEscape* pre-installed on their Chromebooks and were asked to play the game during the 40 minute experiment period. The control group did not have *SpacEscape* installed and were instead encouraged to search out and participate in online science-related content such as reading books or watching science-based educational videos. All participants were then given the same 10 question test as a post-test immediately following the experiment session. This test contained the same demographic questions as the pre-test as well as self-report questions about whether they found Lucy/won the game and what they think they learned from playing *SpacEscape*.

Upon analysis of the test results, it was discovered that all groupings and subgroupings of participants showed significant gains from pre- to post-test. The control group scores improved from an average of 41.57% on the pre-test to 55.54% on the post-test, representing a 13.97% improvement. The experimental group's scores improved from an average of 45.05 to 62.99, or an improvement of 17.94%. When we further decomposed the experimental group into players who found Lucy and won the game and players who did not find Lucy, the successful group increased their scores from 48.03 to 67.38% (19.35% increase) while those who failed to find Lucy increased their scores from 41.09 to 57.17% (16.08% increase). A paired-sample t-test analysis of each of the four data sets shows a significant difference ($p < 0.001$) among pre- versus post-test data.

These findings confirmed several of our hypotheses. We had hoped to see significant growth among the experimental group and hypothesized that participants in the experimental group who were able to find Lucy and complete the game would experience higher gains than those who failed. Finally, it was hypothesized that the control group would be the least improved group among the four different groupings analyzed. And, while all of these hypotheses were borne out there were also some surprising results. We were happy to see that members of the experimental group who failed to find Lucy also experienced significant improvement on the post-test. This was an open question of investigation as it has been hypothesized that mere exposure to the content within a serious game would improve the students subject matter knowledge (Mancuso et al., 2013). We were unsure of the extent to which this group would improve and were pleased to see significant rates of growth between pre- and post-test.

What we did not expect were rates of growth among our control group. While not the same as our experimental rates of growth, the improvement of the control group was similar to the rates of growth in the experimental group and statistically significant in their own right. This can be explained several ways. The most obvious is the "test effect," or the established phenomenon in educational literature which states that tests are naturally motivational and that students will improve based on repeated encounters with the same test (Hartley 1973; Marsden & Torgerson, 2012). While the pre- and post-tests were only being separated by 24 hours and this may lend credence to the "test effect" theory, it is tempered somewhat as students did not receive feedback on their pre-test performance. If the test effect fails to fully explain this improvement, then we can further explore the possibility that *SpacEscape* is an effective tool in supplementing, if not replacing, traditional in-class science instruction.

Experimental Design

We intended to conduct a follow-up study with a similar experimental design as our initial study. This was to be conducted in a classroom setting and in close collaboration with a local school district. The goal of this was to secure participation from an appropriately large ($n > 200$) population of middle school students. Researchers were initially intended to be on-site during multiple days to train both teachers and students on how to play the game, to troubleshoot technical confusion, and to ensure that data was correctly, rigorously, and completely collected.

For this second experiment, we did hope to replicate the results of the original and show that the experimental group showed statistically significant learning between the pre- and post-test. However, we also added new goals within our study design. Within the study, we added a retention test which was to be taken two weeks after the post-test was completed. This would measure how persistent the knowledge gain was. This retention test was identical to the post-test with the addition of an opportunity to provide feedback about the game.

We also added two control questions into our pre-/post-/retention tests. These questions were facts on space science that were not taught by *SpacEscape*. We chose questions within the topic of space science in order to not call attention to them as having some "special" purpose within the study. The goal of these control questions was to control for the test effect more acutely within our experiment, as we hypothesized that gains on these questions should only be attributable to the test effect. We also added event logging within the game so that we could analyze a player's actions in the order and timing that they originally occurred. Finally, we changed some things within the *SpacEscape* game itself. These were largely quality of life improvements that we didn't have time to implement before our first study or were recommended to us through participant feedback.

Unfortunately, the worldwide shutdown due to COVID-19 in March, 2020 caused this plan to be no longer viable as all school districts went online by the end of that month. This caused logistical complications for our study and even threatened to cancel it until school reopened. Without an in-class presence of some sort, it would be virtually impossible to ensure that participants completed the experiment and used the reporting tools properly. Even if we were able to attract a large population, we feared that the data would be incomplete at best.

However, turning the lens around, we realized that this would provide an opportunity to study how effective serious games are in decentralized and unregulated situations. Most learning is intended to be done in environments outside of a regimented classroom setting, so being able to better understand what effects a game has under in such an environment and what level of participation can reasonably be expected under those conditions would be useful information to teachers and researchers. We modified our experimental design to allow us to still conduct the original experiment but in this decentralized environment. We provided written instructions on the game and experiment to the teachers whom we original intended on working with. They were encouraged to provide their students with the instructions for installing, logging into, and playing the game as well as how and when to take the three tests (pre-, post- and retention) we were requiring. The most significant change to our experimental design is the lack of a control group to compare our experimental group with. However, we hoped that the inclusion of the control questions would mitigate this issue in some respects.

Given the lack of a teacher or researcher present with the students during play to help with any issues, we anticipated lower levels of participation than we would have expect had we been with the population in the classroom. We also anticipated lower completion rates across all three experimental tests (pre-test, post-test, and retention test) as students lost interest or focus and dropped off. Given our expectations of complications and motivational issues, we expected lower levels of knowledge gain as measured by the post-test as compared to the initial study. We expected this lower level of knowledge gain to be reflected further in the retention test as it is expected that there would be a drop off in recall of the content over time (Semb & Ellis, 1994). Finally, while we anticipated performance on our control questions to remain relatively level while acknowledging the potential for a slight bump in the post-test due to the test effect before settling to a level near the pre-test in the retention test.

Results

Upon a preliminary analysis of the data collected, our participation hypothesis was confirmed. We originally had anticipated over including over 200 participants in this study. This is due both the size of the 6th grade class we were using as a participant pool but also accounting for attrition due to some students missing school on relevant days of the study. However, as we anticipated with our fully online format, the number of participants plummeted to only 137 unique participants who completed any version of the test (pre, post, or retention). Drilling down further, we had only 17 participants out of the 137 (or 12.4%) who completed all three stages of the experiment. These rates of participation and completion were in line with our hypothesis that student engagement rates would be reduced in our online format.

However, when we drilled into our test results and tracked participant performance, we were excited to see that our expectations on performance were unfounded. We examined the test data of 17 participants who completed the experiment as designed and completed all three tests. A simple statistical examination show an increase in test performance of 13.5 points between pre-test (64.1 points) versus post-test (77.6 points) scores when considering only experimental questions. These scores were functionally maintained between the post-test and the delayed retention test, where students scored 78.8 points on the retention tests.

These differences represent statistically significant increases in performance. Using a one-way ANOVA of the pre-, post-, and delayed retention test data, we found statistical differences in these groups ($F(2, 48) = 3.304$, $p = 0.045$). Further analysis found that these statistical differences were present both with between the pre- and post-test and the pre- and retention test performances. Both differences between pre- and post-test scores (one-tail t-test, $df = 30$ $p = 0.021$) and pre- and retention test scores (one-tail t-test, $df = 32$, $p = 0.018$) represent statistically significant knowledge gains. For completeness, we will also note that the difference between the post- and retention test scores is not statistically significant and that scores actually increased by 1.2 points which is counterintuitive.

Finally, our hypothesis with respect to the control questions were also incorrect. These questions, which were questions within the disciplinary content of planetary science but were not content that could be learned by playing *SpacEscape*, showed a similar improvement trend as the experimental questions. Participants answered the control questions correctly 50% of the time on the pre-test. This rate rose to 67.65% on the post-test and continued to rise to a rate of 70.59% on the delayed retention test. In fact, if we include these questions into our statistical analyses of our test data, the differences between tests become even more statistically significant. Each test returns a stronger result showing that the test results are of different populations (one-way ANOVA: $F(2, 48) = 4.015$, $p = 0.024$; pre-/post one-tail t-test: $df = 29$, $p = 0.013$; pre-/retention one-tail t-test: $df = 32$, $p = 0.012$).

Discussion

Despite these promising results, there are some confounding results and open questions. The most notable issue with our results is the increase in statistical significance of test results when including the control questions. One potential explanation could be that the participants failed to follow the experimental instructions properly and simply completed the retention test far too soon. Experimental instructions asked students to wait two weeks after the post-test to complete the retention test. If they had instead completed the retention test almost immediately after the post test, then this could explain the similarity between the scores. However, we found this not to be the case. Examining the time between the post-test and the retention test, we found that the 17 participants that we have been analyzing universally waited a significant amount of time after the post-test before they completed the retention test. This time ranged from a minimum of 10 days and to a maximum of 25 days with an average of 18.88 days. So, while some students completed their retention test "too soon" with respect to the explicit instructions laid out within the experiment, these students still waited at least a week and a half to complete the delayed retention test and, under the circumstances, this seems like more than enough time to count as a delayed test.

Another possible explanation, and one that could also explain the marginal increase between the post- and retention tests both with respect to only the experimental questions or with respect to the experimental and control questions in combination, is one of a small sample size While our goal was to have over 200 participants, we were only able to achieve a relatively small participant pool (n = 17). Obviously, statistical measures are designed to attempt to account for the sample size and this sample size is more than adequate for a pilot study. Even given this, our results are statistically significant. However, this cannot be completely discounted as a factor and will be able to be explored more as we move forward with follow-up experiments.

The participants' innate interest in space science may also account for either of the above phenomena. In each test, we included a survey question about if the student is currently interested in the disciplinary content. ("Are you interested in learning about our solar system now?") Potential participants were overwhelmingly interested in the content, regardless of playing *SpacEscape*. Out of 102 participants who completed the pre-test, 82 individuals (80.3%) replied in some positive way that they were interested in learning about the solar system. When examining the 17 participants who completed all three phases of the experiment, consistently 15 individuals (88%) replied in some positive way that they were interested in learning about the solar system.

Given this interest of students in space science, it is reasonable to ask if playing *SpacEscape* encouraged students to do additional research on the solar system. Students may have run into some facts or unanswerable questions (such as the control questions) during the experiment and decided to do outside research on the topic. This is a desired benefit and documented phenomenon with serious games. A classic example of this is that of *WolfQuest*, which ties information about wolves and their habitat into a first-person perspective game similar to a first-person shooter. Researchers found that 75% of surveyed participants increased their interest enough that they did or intended to do additional research about wolves while over half of the respondents tied their experience with *WolfQuest* to renewed interest in zoos and parks (Psotka, 2013). This shows the motivational potential of an effective serious game and would hopefully be the primary reason for participants' potential increases on control questions and the retention tests in general.

Conclusions & Future Work

In this study, we conducted an experiment where we attempted to teach 6th grade students about the solar system via our serious game, *SpacEscape*. While we had intended to be in-person for the experiment, the COVID-19 pandemic caused us to move our experiment into an online and distributed format. As anticipated, our participation levels in this setting were far lower than we had hoped for in our original experimental design. However, the participants (n = 17) who completed the experiment and three associated tests (pre-, post-, delayed retention test) showed significant improvement in test scores both between pre- and post-tests ($p = 0.021$) and pre- and retention tests ($p = 0.018$). While the relative sample size may have contributed slightly, the effect should be both minimal and largely accounted for. Another possible factor could be participants' innate interest in the disciplinary content.

These open questions will continue to be examined in future research. Researchers have available to them detailed play logs of all students who played *SpacEscape*. These can be potentially studied and mined for relevant content in an attempt to understand player behavior during game play sessions, including differences between students who manage to find Lucy and win the game and those who fail to find Lucy. Additionally, while the current research attempts to account for overall participant interest in the disciplinary content, it fails to account for relative changes by a participant. Questions will be added to later post- and retention tests asking participants not just if they are interested in solar system science but if they are more, less, or relatively similarly interested as they were before playing *SpacEscape*. We will also be adding questions to understand what sort of additional activities students have or are intending to carry out in order to learn more about the solar system on their own. Finally, in an attempt to make the *SpacEscape* experience more user-friendly as well as to make it available to more participants, researchers are currently porting the game from mobile platforms into a web application which could be access through any internet-facing web browser. This will allow participants to play on desktops, laptops, and Apple devices, not just Android devices as currently constituted.

References

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., Freitas, S. de, Louchart, S., Suttie, N., Berta, R., & Gloria, A. D. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391–411. https://doi.org/10.1111/bjet.12113

Clark, R. C., & Mayer, R. E. (2008). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning, 2nd ed* (pp. xvi, 476). Pfeiffer/John Wiley & Sons.

Djaouti, D., Alvarez, J., & Jessel, J.-P. (2011). *Classifying Serious Games: The G/P/S Model* [Chapter]. Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches; IGI Global. https://doi.org/10.4018/978-1-60960-495-0.ch006

Gee, J. P. (2007). *What Video Games Have to Teach Us About Learning and Literacy. Second Edition: Revised and Updated Edition* (2nd edition). St. Martin's Griffin.

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

Hartley, J. (1973). The effect of pre-testing on post-test performance. *Instructional Science*, *2*(2), 193–214. https://doi.org/10.1007/BF00139871

Mancuso, V. F., Hamilton, K., Tesler, R., Mohammed, S., & McNeese, M. (2013). An Experimental Evaluation of the Effectiveness of Endogenous and Exogenous Fantasy in Computer-Based Simulation Training. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, *5*(1), 50–65. https://doi.org/10.4018/jgcms.2013010104

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583–616.

Michael, D. R., & Chen, S. L. (2005). *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade.

National Research Council. (2015). *Reaching Students: What Research Says About Effective Instruction in Undergraduate Science and Engineering*. The National Academies Press. https://doi.org/10.17226/18687

Psotka, J. (2013). Educational Games and Virtual Reality as Disruptive Technologies. *Journal of Educational Technology & Society*, *16*(2), 69–80.

Semb, G. B., & Ellis, J. A. (1994). Knowledge Taught in School: What Is Remembered? *Review of Educational Research*, *64*(2), 253–286. https://doi.org/10.3102/00346543064002253

Wechselberger, U. (2009). Teaching Me Softly: Experiences and Reflections on Informal Educational Game Design. *T. Edutainment*, *2*, 90–104. https://doi.org/10.1007/978-3-642-03270-7_7

Zhonggen, Y. (2019, February 3). *A Meta-Analysis of Use of Serious Games in Education over a Decade* [Review Article]. International Journal of Computer Games Technology. https://doi.org/10.1155/2019/4797032