LEARNING FROM MISTAKES: EVALUATING LEARNING IN A SERIOUS GAME

Hendranus Vermeulen
James Gain
Department of Computer Science, University of Cape Town

ABSTRACT

This paper reports on evaluating learning in serious games by proposing and implementing an experimental method. A Serious Game was developed that aimed to encapsulate the execution of the A* pathfinding algorithm and utilized immediate feedback as a mechanism to promote learning. A qualitative analysis of users' experience of the Serious Game was highly positive and the feedback component rated as highly relevant for learning. However, an initial attempt at quantitatively evaluating learning in a Serious Game through formal experiment did not achieve the expected results. This paper will discuss useful insights gained and lessons learned from the implementation. In particular it will reflect on the possible contradiction between evaluation and games, where games often encourage making mistakes as a mechanism for learning and evaluation typically regards error as failure. Future development of serious games and refining the implementation of the methodology are proposed in light of the lessons learned.

Index Terms— Serious Games, Evaluation, Methodology, Education, Pathfinding, Computer Science

1. INTRODUCTION

The term "serious games" has only risen to prominence in the last decade and also recently become accepted as an academic topic. However, the origin of the term can be traced back to Clark Abt who further postulated the opportunity of making computer simulations of serious games [1].

> *We are concerned with serious games in the sense that these games have an explicit and careful thought-out educational purpose and are not intended to be played primarily for amusement [1].*

The term "serious games" is not unproblematic. Abt intended this oxymoron to unite seriousness of thought and problems with the experimental and emotional freedom of play [1]. He also acknowledged the major challenge is defining the "game" component of the term. Not surprisingly, this is also where most contemporary contention regarding the definition of serious games stems. Games come in a great variety of forms (structures and technologies). We concur with Abt's notion that a game is best considered as a way of looking at something combining analytical reason with playful creativity to solve some challenge. Making serious games and evaluating their effectiveness has also proven challenging. Games are complex systems conceptually and when in digital form also technologically. Creating games is as much art as science. There is no formula and commonly games are designed through an iterative process of prototyping and play testing. When we consider the literature of educational games and simulations over the last three decades, one of the prominent challenges that emerge is a lack of empirical evidence of the educational effectiveness of serious games and simulations [2-4, 6, 8].

This begs the question why many educators and researchers believe serious games have an important role to play in learning. This conviction stems from the perceived relevance of games to contemporary education. Games structure activities in new and challenging ways to engage participants. Games can break learning down into smaller, progressively more challenging tasks and provide instant feedback. Games can model complex processes and authentic activities, providing dynamic game objects the player can interact with and manipulate. This enables the player to transform the game system and in so doing come to understand the process behind the representation. Games are also considered to impart 21st Century skills, embody many constructivist and active learning principles; and have been shown to intensify learners' engagement and motivation [3]. We concur with these intuitions, however, it is important that they are proven objectively and thus empirically. Serious games have only recently become the subject of experimental research and the results of these studies are inconclusive [6]. There is a lack of clear methods for evaluating the effectiveness of serious games as learning tools [3]. It is not surprising that the few studies that do attempt such evaluations have conflicting results. This paper responds to this challenge by proposing an experimental method for evaluating the effectiveness of serious games as learning tools. This method was implemented to evaluate the effectiveness of a Serious Game. Mistakes were made and lessons learned, however, we regard the development of serious games and methods to evaluate them as an iterative process; and as such this should be regarded as an iteration.

Figure 1. Screenshot of the A* Pathfinding Game

2. EVALUATION METHODOLOGY

The experimental method for evaluating the educational effectiveness of serious games proposed in this paper can be seen as an extension on the traditional controlled user experiment. We concur with Girard, et al. [6] that for a meaningful evaluation it is important to compare the serious game to another "type" of learning activity. Many studies compare groups that aim to learn from a serious game to groups that do not receive any instruction. This seems counter intuitive and although learning can be inferred this does not prove that the serious game is more effective than some other "type" of learning activity. We therefore propose a counterbalanced experimental design of two groups, where each group participates in a particular "type" of learning activity first and then in the other learning activity second. The interaction (learning effect) between the two groups and learning activity "types" can then be analyzed for statistical significant variance using a repeated measure two-way ANOVA. The method is a within-subject comparison, greatly strengthening the power of the experiment, as the measures are not generalized from a sample to a population.

We also propose that some alternative measures be incorporated into the design to provide a means to triangulate results and consider the users' subjective experience and valuation of the game and its elements. The learning effect of the different "types" of learning activities can be quantitatively operationalized as the reduction of error during the participants' execution of a task. This is typically the assumption underlying most academic assessment, i.e. assessment scores calculated from correct answers and error. The reduction of error is typically regarded as proof of learning taking place and valued as the objective of institutionalized learning. We postulate that when considering the interaction between groups participating in two "types" of learning activities over the two days, the group participating in the serious game activity first would show a significantly greater reduction in error when completing the tasks. One of the major challenges when doing experimental research is controlling for external variables that can pollute the research questions. This proved particularly challenging since the implementation of the experiment took place in situ of a scheduled "Games AI" course component. This component forms part of the 2nd year elective "Computer Games Design" course in Computer Science at our university. The experiment therefore had to take second place when conflicts arose, so not to disrupt the intended learning. However, great care was taken during the experiment, in the design of the Serious Game and alternative learning activity, to control what factors varied between the two learning activities.

3. THE SERIOUS GAME VS. TUTORIAL

The Computer Science curriculum is a fertile domain for the development of serious games. Computer Science is

primarily about solving problems through computation, which can be framed as playful experiences. We identified the A* Pathfinding algorithm [7] as the problem to embody in a Serious Game. We started the design of the game with an analysis of a pen and paper exercise that was used in tests and exams. These typically represented a tiled grid with some tiles being obstacles, a starting tile and end tile. The task required students to calculate the shortest path from start to end tile, stepping through the algorithm and doing the necessary calculations, while considering the specified heuristic. We developed two prototypes before embarking on the final game.

The first prototype focused on generating the tile grid and computationally stepping through the algorithm. The intent behind this was that we could generate different levels for the game, evaluate the players' action during each step of their execution of the algorithm, provide detailed feedback to the player and enable error tracking. Accurate immediate feedback is regarded as an important component of the judgment-behavior-feedback loop in serious games [5]. The second prototype considered the game's user interface and look and feel. The final game (see Figure 1 for a screenshot of the interface) utilized various game elements to enhance the experience of the game and the learning it intended to encourage. These elements included the game's immediate feedback, back-story, steampunk theme, user interface, graphical elements and animations. In the game story Dr. Gerasimov (a scientist that found himself as a head in a jar after a near fatal accident) requires the player to find him mechanical components in the maze like laboratories, in order to build a mechanical body. The A* Pathfinder Mechanism (see the compass like user interface element, Figure 1) in particular was developed to scaffold the learning of the A* Pathfinding algorithm. It was presented in the game story as an ingenious machine that calculates a safe (shortest) path through the laboratories and their traps. This interface element provided a view of the tile's G, H and F values; as well as a needle that pointed to the tile's parent. The needle's color also indicated the state of the tile. This information is needed for calculating the shortest path. As levels progressed the mechanism started to malfunction, requiring the students to manually set the values and so scaffolding is gradually removed. When students made mistakes, they set off traps (gas and/or electrical shocks that depleted their health serum and score) in the laboratory on the particular tile the student erred. Dr. Gerasimov also provided feedback to the player, setting and highlighting the correct values in the A* Pathfinder Mechanism.

The evaluation methodology we propose requires an alternative learning activity to provide a basis for comparison. We could have used the traditional pen and paper exercise. However, we decided to use a stripped down version of the Serious Game, as this would enable us to utilize the existing error tracking functionality. As mentioned earlier external variables are one of the challenges of doing experimental research and we hoped that by controlling the game elements we regarded as important for learning this would provide a clearer indication of what elements effected learning and how. The elements that were omitted from the Tutorial learning activity were immediate corrective feedback on every tile the player interacted with, back-story, theme, graphical elements and animations, i.e. all elements that contributed to the problem being framed as a game. We aimed to keep the tasks underlying the two activities as close as possible to one another to eliminate task differences from polluting the experiment. The layout of the tile grids (levels) representing the problem spaces were the same. Furthermore, both tasks had 3 levels of increasing difficulty, a time limit which decreased the score; and when completed both activities earned the participant some Experience Points (XP). XP served as external motivation as it counted a small amount (0.24% for participating in both learning activities and the evaluation questionnaire) towards the course mark and could be exchanged for rewards such as extensions on course work. It is important to note that the Tutorial learning activity also provided students with feedback regarding their performance of the task. However, there was no Dr. Gerasimov, A* Pathfinder Mechanism or trap animations to provide immediate feedback. The feedback took the form of a summative evaluation at the end of each level when all incorrect tiles were highlighted. Omitting feedback entirely would have contradicted the learning objectives of the "Games AI" course. In light of the experimental factors we aimed to control, the summative feedback of the Tutorial was regarded as significantly different from the immediate feedback from the Serious Game.

4. THE EXPERIMENT

The evaluation methodology proposed in this paper was implemented as an experiment over a period of two weeks as part of the voluntary Friday sessions of the "AI for Games" course component. The intention of the Friday sessions were that the lecturer could explore case studies and related topics not covered directly by the curriculum. Students were invited to participate in the two Friday sessions and the invitation explained the objective of the comparative evaluation of the Serious Game that would take place. Of the 55 students enrolled in the "AI for Games" component 34 participated in the first Friday session, 27 participated in the second; and 18 did not participate in the experiment. The A* algorithm was briefly introduced in two 45 minute lectures before the first Friday session. During the Friday sessions all

participants in the experiment were given access to the Serious Game or Tutorial learning activities, however, the order of presentation was permuted over the two Friday sessions. The learning activities were embedded in the university's online learning management system accessed through the internet. The system assigned students equally to two groups (Serious Game First and Tutorial First) as they logged in. The system would also assign the participants to the alternative condition in the following week or if new participants participated, it would balance the number of participants in each group. The learning activities provided instructions to the participants regarding their respective tasks, interfaces and objectives. Participants were expected to complete the task individually and both learning activities would reward the participants with a puzzle code on completion. The code could then be entered into the online system to redeem XP. The researcher's role was to support the participants in the execution of their tasks, answering questions and clarifying ambiguities. Each of the 2 Friday sessions lasted 45 minutes and both learning activities interfaces had a countdown clock which showed a time limit for each of the 3 levels. When the time limit was reached the level's score started diminishing. This was designed to encourage participants to complete the task within the time allocated. A week after the experiment all participants of the experiment were invited to complete an online evaluation questionnaire. This formative evaluation aimed to gather feedback for improving the experiment and learning activities developed; as well as provide qualitative data to triangulate the quantitative results.

5. THE RESULTS

The Quantitative data collected as a measure of learning was the number of errors. The back-end system recorded participant errors, reporting the error type, tile and level at which it occurred. The only difference between the Serious Game and Tutorial error recording was that the Serious Game recorded errors at each step of the A* algorithm's execution, while the Tutorial reported a summary of all the errors at the end of each level. It was necessary to include only the data of students who participated in both Friday sessions (required for the repeated measures ANOVA) as not all students participated in both sessions. Also, we found that some students did not complete all the levels of the activities, so some levels were excluded to match the data for both days. This resulted in the selection of 10 participants for the "Serious Game First" group and 11 participants for the "Tutorial First" group. The null hypothesis for the experiment was: *The interaction between the conditions (Serious Game First vs. Tutorial First) over the two days will have no significant effect on the variance of errors made.* Although the ANOVA calculated that the reduction of errors over the two days were of 0.06 significance calculated at alpha = .05 (0.839 observed power) the reason for this reduction in errors is unfortunately unclear. There are many factors that could have contributed to this, including the intervening lectures on A*, a test and the practice students had between the two sessions. Importantly the interaction between the two conditions over the two days did not report statistical significance (0.21 significance calculated at alpha = .05 and 0.234 observed power) and when considering the conditions more closely the Tutorial learning activity measured a greater reduction in error than the Serious Game. *We therefore accept the null hypothesis when considering all errors*. ANOVAS for all 3 levels were also calculated to provide a more fine grained analysis of the data. We found statistical significance of 0.045 calculated at alpha = .05 (0.531 observed power) for Level 3, however, in favor for the Tutorial learning activity. *We can therefore reject the null hypothesis for Level 3 errors for the Tutorial learning activity.* The data for the qualitative evaluation of the Serious Game was collected through the online evaluation questionnaire as well as interviews with participants. The Serious Game was well received and participants reported enjoying and regarding it as beneficial for learning the A* pathfinding algorithm. Students continued playing after the experiments and some asked for access to practice for tests. The questionnaire included a Likert Scale (1- strongly disagree to 5- strongly agree), rating questions (5 scale where a higher number rates as more effective) and open questions. In total 22 students completed the questionnaire. See tables 1 and 2 below, for the means and standard deviations of noteworthy responses:

Table 1. Noteworthy Likert Scale responses to statements.

Statement	**Mean**	**SD**
The A* game is good for introducing the A* algorithm.	4.1	0.77
I enjoyed playing the A* game.	3.95	0.8
I prefer the "Tutorial" exercise to the A* game.	2.1	1
The A* game should NOT be part of next year's course.	1.48	0.6
I believe the A* game improved my performance in the test.	3.76	1.09
I did NOT learn anything from playing the A* game.	1.62	0.8

Table 2. Elements ranked highly according to perceived contribution towards learning.

Serious Game element	Mean	SD
Immediate feedback when correct action was taken or mistakes made.	4.48	1.12
Pathfinder Mechanism as Interface to tile editing.	4	1.3

6. DISCUSSION

There exists a clear dissonance between the quantitative and qualitative results of this study. This begs the question: What was the cause for this contradiction within the results? Was it because the Serious Game was less effective despite it being preferred? Or, was it that there was a flaw in the evaluation of the Serious Game's effectiveness in promoting learning? We believe that although the evaluation methodology is sound there were serious flaws in our implementation. One major oversight was the fact that we conflated the evaluation phase with the learning/play phase. Due to time and scheduling constraints the experiment was limited to the two Friday sessions. This meant that errors were recorded from the very first attempt participants made. We believe this was a mistake as some participants later reported utilizing the Serious Game's immediate feedback as a strategy for learning. The immediate feedback of the Serious Game therefore promoted experimentation through mistakes by participants. The Tutorial on the other hand only provided feedback at the end of each level and it was therefore more difficult to utilize this mechanism as a strategy for learning. *We recommend that when doing an evaluation of a serious game using the method proposed in this paper, there should be distinct phases dedicated to learning and evaluation; and ample time for participants to undertake the learning phase*.

Considering this we also speculate that there is an important conflict at the heart of evaluating serious games. It would seem that there is a proverbial *Heisenberg Principle* at work when evaluating serious games. When we use serious games in academia and evaluate them using assessment criteria such as the measurement of error, there is a chance that this very act can disrupt the evaluation of a serious game. When we consider the two cultures of "games" and "academia", we note that games are often regarded as less strict, playful, opportunistic, incorporating activities of cheating, collusion and trying to "game" the system. While in academia activities are more formal, strict, regulated and uncompromising in particular the rules of assessment practice. Imposing academic assessment criteria and constraints on a serious game could be in conflict with the playing of the game. When these two "cultures" meet in the form of a serious game we often see the playful and opportunistic practices of games sneak in to subvert academic assessment practice. *We do not believe that this should be a reason not to utilize serious games in academic activities, but rather we should utilize these "game cheats" to afford the achievement of learning objectives.*

We also consider the Tutorial vs. Serious Game activities to have been too similar. The Tutorial activity was essentially derived by omitting what were regarded as important game elements. Yet, the core of the activities were the same. In light of the time constraints of the experiment and the fact that the evaluation took place at the same time as the learning activity, the activity produced was more akin to an academic assessment. With these constraints the stripped down Tutorial activity could possibly be more efficient. Although participants rated the immediate feedback of the Serious Game as beneficial for learning they also reported it frustrating at later levels. In particular when they were penalized for making interaction errors. The evaluation metric of errors is also quite limited in its ability to evaluate the conceptual learning and understanding of the subject material. *In light of this we would recommend that the evaluation not only be conducted in a distinct phase in time but also in form. The evaluation should ideally include practical as well as conceptual assessments of the subject matter.* In the case of our research, we could have utilized a pen and paper exercise as evaluation, including questions that assess both the mechanistic execution of the A* algorithm and theoretical. *Furthermore, we propose that feedback can be scaled down as the player progress or configured by the player, so not to cause frustration*.

Lastly we would like to acknowledge that doing experimental research and evaluating the learning effectiveness of serious games in context of a learning institution can be chaotic and challenging. The role of teacher/experimenter is particularly precarious as the demands made from these two roles are often in conflict. Notwithstanding, it is our contention that such situations can be very rewarding and afford much learning to take place, for teacher, experimenter and students. *When finding yourself in a situation of conflict, what is important is that the learning objectives take precedence.*

7. CONCLUSION

In this paper we proposed an experimental method for evaluating the effectiveness of serious games as learning tools. We designed a Serious Game and alternative Tutorial learning activity with the aim of implementing the method proposed. During the implementation we experienced the challenge of experimentally evaluating the effectiveness of serious games as tools for learning. Although we regard our

first attempt as a failure we draw on the game paradigm to frame our mistakes as opportunities to learn and "level up". We believe that it is important to pursue experimental research into serious games. Currently there is a shortage of experimental evidence for the effectiveness of serious games in achieving learning objectives. It is important to have different measures and data to triangulate our results. We aim in the next year to improve on our implementation of the method proposed in this paper in multiple experiments.

8. ACKNOWLEDGMENTS

We would like to thank the National Research Foundation of South Africa for funding this research and http://illustratorg.deviantart.com for the graphics used to produce the Serious Game.

9. REFERENCES

[1] Abt, C. C. *Serious games.* Viking Press, 1987.

[2] De Freitas, S. and Oliver, M. How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education*, 46, 3 (2006), 249-264.

[3] Dobson, M. W. and Ha, D. Exploring interactive stories in an HIV/AIDS learning game: HEALTHSIMNET. *Simulation & Gaming*, 39, 1 (2008), 39-63.

[4] Druckman, D. The educational effectiveness of interactive games. *Simulation and gaming across disciplines and cultures: ISAGA at a Watershed*, (1995), 178-187.

[5] Garris, R., Ahlers, R. and Driskell, J. E. Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming*, 33, 4 (2002), 441 - 467.

[6] Girard, C., Ecalle, J. and Magnan, A. Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29, 3, (2013), 207-219.

[7] Hart, P. E., Nilsson, N. J. and Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *Systems Science and Cybernetics, IEEE Transactions On*, 4, 2 (1968), 100-107.

[8] Pierfy, D. A. Comparative simulation game research: Stumbling blocks and steppingstones. *Simulation & Games*, 8, 2, (1977).