EXPERIMENAL METHDOLOGY FOR EVALUATING LEARNING IN SERIOUS GAMES

Hendranus Vermeulen (hendranus@gmail.com)
James Gain (jgain@cs.uct.ac.za)
Department of Computer Science
University of Cape Town
Private Bag X3
Rondebosch, 7701
Tel. +27 21 650 2663 / Fax. +27 21 650 3551

ABSTRACT

This paper reports on a project to evaluate the effectiveness of using a *Serious Game* as a learning tool. A "serious game" has an explicit and carefully thought-out educational purpose and is not intended to be played primarily for amusement. The formal experimental method was designed to compare formal learning of the A* pathfinding algorithm with a digital serious game implementation that 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 it as highly relevant for learning. However, an initial attempt to quantitatively evaluate learning in a serious game through the formal experiment did not achieve the expected results. This paper discusses useful insights gained and lessons learned from the implementation. It reflects on the possible contradiction between evaluation and games, where games often encourage making mistakes as a mechanism for learning while evaluation typically regards error as failure. Future development of serious games and refining the learning evaluation 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 research topic. However, the origin of the term can be traced back to Clark Abt who further postulated the opportunity of making computer simulated serious games (Abt, 1987).

> *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 (Abt, 1987).*

The term "serious games" is not unproblematic. Abt (1987) intended this oxymoron to unite seriousness of thought and problems with the experimental and emotional freedom of play (Abt, 1987). He also acknowledged the major challenge in defining the "game" component of the term. Not surprisingly, this is also where most contemporary contention regarding the definition of serious games stems from. Games come in a great variety of forms (structures and technologies). We concur with Abt's notion of a game as a way of looking at a challenge in such a way as to combine analytical reason with playful creativity to solve it. Making serious games and evaluating their effectiveness has also proved

challenging. Games are complex systems conceptually and when in digital form they are also technologically complex. Creating games is as much art as science and commonly games are designed through an iterative process of prototyping and play testing. A common challenge in the literature on educational games and simulations over the last three decades has been a lack of empirical evidence of their educational effectiveness (De Freitas & Oliver, 2006; Dobson & Ha, 2008; Druckman, 1995; Girard, Ecalle, & Magnan, 2013; Pierfy, 1977).

This uncertainty begs the question why many educators and researchers believe serious games have an important role to play in learning. This perhaps poorly grounded conviction may stem from the perceived relevance of games to contemporary education. Games structure activities in new and challenging ways to engage participants. They can break learning down into smaller, progressively more challenging tasks and they provide instant feedback. Games can model complex processes and authentic activities, providing dynamic game objects the player can interact with and manipulate. Such dynamic game objects enable 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 (Dobson & Ha, 2008). While we concur with these intuitions, it is important that they are empirically proven. Serious games have only recently become the subject of experimental research and the results of these studies have often been inconclusive (Girard et al., 2013) There is a lack of clear methods for evaluating the effectiveness of serious games as learning tools (Dobson & Ha, 2008; Mayer et al., 2014). Serious game evaluations have been found to utilize diverse theoretical frameworks and methods reflecting the transdisciplinary nature of the area (Connolly, Boyle, MacArthur, Hainey & Boyle, 2012). It is not surprising that the few studies that have attempted such evaluations have produced 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 particular 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 this study should be regarded as one such iteration.

2. EVALUATION METHODOLOGY

The experimental method used in this study to evaluate the educational effectiveness of serious games was designed as an adaptation of the traditional controlled experiment. We agree with Girard, et al. (2013) that for a meaningful evaluation it is important to compare the serious game to another type of learning activity. Many studies merely compare groups that aim to learn from a serious game to groups that do not receive any instruction. Although learning can be inferred, such an approach does not prove that the serious game is more effective than some other type of learning activity. We therefore designed a counterbalanced experimental design using two groups, where each group participated in one type of learning activity first and then in the other learning activity. The interaction (learning effect) between the two groups and learning activity types could then be analyzed for statistical significant variance using a repeated measure two-way ANOVA. The method was a within-subject comparison, greatly strengthening the power of the experiment, as the measures were not generalized from a sample to a population.

We also designed in alternative measures to triangulate with the ANOVA results, in form of qualitative data on learner's subjective experience and valuation of the game and its elements. The learning effect of the different types of learning activities was measured from the reduction of errors committed during the participants' execution of a task. This was considered typical of measures underlying academic assessment, i.e. assessment scores calculated from correct answers and errors. Our research hypothesis was that the group participating in the serious game activity first would show a significantly greater reduction in errors committed when completing their tasks. The problem of controlling for external variables that might pollute the research questions was particularly challenging since the experiment took place in context of a scheduled "Games AI" course component. This component formed part of the 2nd year elective "Computer Games Design" course in Computer Science at our university. The experimental design therefore had to take second place when conflicts arose, so not to disrupt the intended learning. However, care was taken during the experiment, in the design of the serious game and alternative learning activity, to control variables between the two learning activities.

Figure 1. Screenshot of the A* Pathfinding Game

3. THE SERIOUS GAME VS. TUTORIAL

The Computer Science curriculum is a fertile domain for the development of serious games seeing that the subject area is primarily about solving problems through computation, which can be framed as playful experiences. We identified the A* Pathfinding algorithm (Hart, Nilsson, & Raphael, 1968) 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 representing 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. This allowed us to 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 considered an important component of the judgment-behavior-feedback loop in serious games (Garris, Ahlers & Driskell, 2002) and those without sustained elaborative feedback are not considered an effective strategy for facilitating increased student achievement (Cameron & Dwyer, 2005). The second prototype considered the game's user interface look and feel. The final game (see Figure 1 for a screenshot of the interface) utilized various game elements to enhance the gaming experience and the learning it was 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 is a scientist who has found himself as a head in a jar after a near fatal accident. Gerasimov requires the player to find him mechanical components in the maze imagined as a laboratory, 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 G value represents the cost of getting to this tile from the start, the H value represents the estimated cost of getting from this tile to the goal; and the F value represents the total estimate of the path length from start to goal, i.e. the sum of G and F. The needle's color also indicated the state of the tile. This information was needed for calculating the shortest path. As students progressed through levels the mechanism started to malfunction, requiring them to manually set the values and in this way scaffolding was 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 where the student erred. Dr. Gerasimov also provided feedback to the player, setting and highlighting the correct values in the A* Pathfinder Mechanism.

Our experimental design required an alternative Tutorial learning activity to provide a basis for comparative evaluation. 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 its existing error tracking functionality. We hoped that by controlling those game elements we regarded as important for game learning this would provide a clearer indication of what elements had affected 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 task 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 was the same for Game and Tutorial implementations. Furthermore, both tasks had 3 levels of increasing difficulty, a time limit that decreased the score; and when both activities had been completed the student earned some Experience Points (XP). These points served as external motivation as they 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. 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 Tutorial feedback merely took the form of a summative evaluation at the end of each level when all incorrect tiles were highlighted. As this feedback was not immediate, it was expected not to support the judgment-behavior-feedback loop prevalent in the Serious Game. Omitting feedback entirely would have contradicted the learning objectives of the "Games AI" course.

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 for second year students at the Department of Computer Science at University of Cape Town. These sessions were meant to allow the lecturer to explore case studies and related topics not covered directly by the curriculum. Students were invited to participate in the two weekly sessions. The invitation explained that the aim of the sessions were a comparative evaluation of the serious game. Of the 55 students enrolled in the "AI for Games" component 34 participated in the first session and 27 in the second while 18 did not participate at all. The A* algorithm was briefly introduced in two 45 minute lectures before the first Friday session. All participants in the experiment were given access to the Serious Game or Tutorial learning activities during the voluntary sessions. However, the order of presentation was permuted over the two 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. If new participants joined, the system aimed to 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. Both learning activities rewarded the participants with a puzzle code on completion. This code could then be entered into the online system to redeem XP. The researcher's role was to support the participants in their tasks, answering questions and clarifying ambiguities. Each of the 2 voluntary 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 mechanism was designed to encourage participants to complete the task within the time allocated. A week after the experiment all participants 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 additional qualitative research data.

Krathwohl (1998) drew a useful distinction that in qualitative studies the explanation (hypotheses) guides the development of the study, while in qualitative research the explanation follows from the data collected. By using both quantitative and qualitative methods this research methodology aimed to triangulate the findings by using both deductive and inductive approaches respectively.

5. THE RESULTS

The Quantitative data collected as a measure of learning was the number of participant errors. The back-end system recorded these 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. Not all students participated in both sessions. To meet the requirements for the repeated measures ANOVA, we therefore included the data of those who had participated in both. Some students had not completed all the levels of the activities, and therefore some levels were analyzed for fewer participants. These restrictions resulted in the selection of 10 participants for the "Serious Game First" group and 11 participants for the "Tutorial First" group. *The research hypothesis proposed that students who learned through the Serious Game first would make significantly less errors during their tutorial activity than the students learning from the Tutorial first*. 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). However, the reason for this reduction in errors is unfortunately unclear. Several factors could have contributed to this, including the intervening lectures on A*, a test and the practice students had received during 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). 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 all errors are considered. ANOVAS for all 3 levels were also calculated to provide a more fine-grained analysis of the data. Statistical significance of 0.045 was calculated at alpha = .05 (0.531 observed power) for Level 3, however, in favor for the Tutorial learning activity. *This meant that for Level 3: students who learned through the Tutorial first made significantly less errors during their Serious Game activity than the students learning from the Serious Game first.* This was contrary to our expectations.

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) 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 (ranking 1 - low to 5 - high)	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

This data showed that the Serious Game was well received by students. There were a few negative comments on the Game, including some reports of frustrations at later levels. In particular when students were penalized for making interaction errors. However, during interviews and in the comment section of the questionnaire students reported that they found it easier to visualize the execution of the A* algorithm as it was stepped through using the dynamic and interactive serious game. They found the visual aspects and animations of the game appealing. Students rated the immediate feedback and the Pathfinder Mechanism highly when considering what contributed to their learning. When asked "What, if anything, did you learn about the A* pathfinding algorithm from the A* game?" Typical responses were "*It helped me learn the basics of the algorithm"*, "*the feedback allowed me to become more confident in my understanding"* and *"It helped me follow the execution of the algorithm more strictly"*. With the pen and paper alternative, they would sometimes forget to do a step and then have to go back and rework sections of the algorithm. Students also reported that the serious game played an important role in the reduction of errors made. They found the scaffolding of increased difficulty with levels facilitated their mastery of the algorithm. The more they played the easier it became. When it was time for the students to use the algorithm on their own game, they reported that the serious game had benefited their implementation.

6. DISCUSSION

The quantitative and qualitative results of this study contradicted each other. Although the participants believed that the serious game was a better learning tool than the tutorial exercise, a measurable superiority was not found. Possible explanations include that the game was indeed less effective despite it being preferred, or that there was a flaw in the evaluation of its effectiveness in promoting learning. The "Hawthorn Effect" (Adair, Sharpe & Huynh, 1989) could also be considered, however, we believe that this is doubtful, as students in both learning activities received the same attention. Furthermore, the randomized control experimental design would have controlled for this effect. We believe instead that there were flaws in our implementation. One oversight was 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. This might have been a mistake on our part 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*.

We also speculate that there is a conflict at the heart of evaluating serious games, analogous to the *Heisenberg Principle* (1930) in physics. When we use serious games in academia and evaluate them using assessment criteria such as the measurement of errors, this very act might disrupt the evaluation of a serious game. Comparing the two cultures of "games" and "academia", the former 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" for the achievement of learning objectives.*

We also consider the Tutorial and Serious Game activities to have been too similar to each other. The Tutorial activity was essentially constructed by omitting what were regarded as important game elements. Yet, the core of these 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 involved was more akin to an academic assessment than a game. With these constraints the stripped down Tutorial activity could possibly be more efficient and accessible for students as it was closer to the activities that students are familiar with within the university context. Although participants rated the immediate feedback of the Serious Game as beneficial for learning as noted earlier they also reported it to be 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 aspects. *Furthermore, we propose that feedback can be scaled down as the player progress or configured by the player, so not to cause frustration*.

Finally, when considering the serious game as technologically mediated instruction, the lack of statistical significance within the quantitative results might be attributed to the "No Significant Difference Phenomenon" so called by Russell (1998). He compiled a bibliography of 355 media

comparison studies for the period of 1928 to 1998, all of which found no significant difference between the technological and live delivery methods. This was argued in support of Clark's (1983) theory that the instruction medium has no effect on the learning outcome. However, important criticisms have been raised against that position. A later study by the Institute for Higher Education Policy (1999) found that many of the studies reported in Russell's bibliography suffered from: poor design, inadequate control of extraneous variables, lack of random selection of subjects, poor validity and reliability of instruments used for measuring outcomes and lack of adequate controls of "reactive effects". A common criticism raised against media comparison studies is that they fail to control for extraneous variables (Joy & Garcia, 2000; Lockee, Moore & Burton, 2001). Indeed the research reported study suffered from this defect. Controlling external variables while evaluating the learning effectiveness of serious games in context of a learning institution can be complex and difficult. Furthermore, 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 experimental and developmental contexts can be very rewarding and afford much learning to take place, for teacher, experimenter and students. *When in a situation of conflict, what is important is that the learning objectives take precedence.* A deeper understanding of the human context of the experimental implementation could facilitate a better control over extraneous variables.

7. CONCLUSION

In this paper we proposed an experimental method for evaluating the effectiveness of serious games as learning tools. We designed one such Serious Game and an alternative Tutorial learning activity and used them as alternative media in a symmetrically balanced learning experiment. It was found that participating students reported favorably on the Serious Game as a learning tool, but that subjective impression could not be objectively detected in the relative learning outcomes. It might be necessary to separate the processes of learning from evaluation of its success more strictly than we did in this first experiment. There remains a shortage of experimental evidence for the effectiveness of serious games in achieving learning objectives. 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

Abt, C. C. (1987). *Serious games*. University Press of America.

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research*, 59(2), 215-228.

Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research*, 16(3), 243–258.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, 53(4), 445–459.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2), 661-686.

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

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

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

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33(4), 441-467.

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

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *Systems Science and Cybernetics, IEEE Transactions on*, 4(2), 100-107.

Heisenberg, W. (1930). *The physical principles of quantum mechanics*. U. Chicago Press, Chicago.

Institute for Higher Education Policy (1999). *What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education*. Retrieved from http:// http://www.ihep.org/research/publications/whats-difference-review-contemporary-research-effectiveness-distance-learning

Joy, E. H. & Garcia, F. E. (2000). Measuring Learning Effectiveness: A New Look at No- Significant-Difference Findings. *Journal of Asynchronous Learning Networks*, 4(1): 33- 39.

Krathwohl, D. R. (1998). *Methods of educational and social science research: An integrated approach*. Longman/Addison Wesley Longman.

Lockee, B., Moore, M., & Burton, J. (2001). Old Concerns with New Distance Education Research. *EDUCAUSE Quarterly*, 24(2): 60-62.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., Lo, J., Kortmann, R. & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, 45(3), 502-527.

Pierfy, D. A. (1977). Comparative simulation game research: Stumbling blocks and steppingstones. *Simulation & Games*, 8(2), 255-268.

Russell, T. L. (1998). *No Significant Difference Phenomenon*. Raleigh.