An Innovative Longitudinal Evaluation of a Digital Game – The First Impressions

Casper Harteveld
Delft University of Technology
c.harteveld@tudelft.nl

Abstract. A need exists to show that games work. To achieve this, proper evaluation frameworks and/or methodologies are necessary. In this paper, the author elaborates on an innovative longitudinal evaluation of a digital game. The methodological setup and the first results are presented. This way, a contribution is made to the ongoing academic discussion on how to evaluate games. It also gives an outlook on further evidence that games do work. Based on the first impressions of this empirical study, it is concluded that while its methodological setup is arduous and risky, it is very worthwhile in the end.

1. INTRODUCTION

Do games work? Many scholars have shed their light on this question and have come up with different *ideas* (see e.g., De Caluwé, Hofstede, & Peters, 2008). It is important to stress that these are foremost ideas and not *proofs*, although some of these ideas may have been based on *strong* empirical evidence. And some strong empirical evidence certainly exists (see e.g., Kato, Cole, Bradlyn, & Pollock, 2009). However, the call for more empirical evidence has been highlighted in many articles and reports (e.g., De Freitas, 2006), indicating that more proper evaluations are being asked for to show that games work.

Related to this is a search for *how* to pursue game evaluations. For proper evaluations it is first of all necessary, as Kriz and Hense (2006) argue, to use a framework and methodology. Most empirical game studies lack these (Ke, 2009), making the results unreliable and incomparable. For this reason, it is desirable to incorporate a framework and methodology. The problem is that these hardly exist. Only until recently, some scholars have made an effort to develop them (see e.g., De Freitas & Oliver, 2006; Winn, 2009).

In this paper I will highlight the methodological setup and the first results of a rather innovative longitudinal evaluation of a digital game. This way, this paper attempts to fulfill a contribution to the ongoing academic discussion on how to evaluate games. It also gives an outlook on further empirical evidence that games work. Before I get into the details of the empirical study, I will first explain what the game being studied is about.

2. THE GAME LEVEE PATROLLER

The game being studied concerns *Levee Patroller* (Harteveld, 2011). It can be described as a "single-player 3-D first person game." This means the game is solely played by one user from the perspective of the player character. It was implemented using the commercial game engine "Unreal Engine 2." In the game, the player's role is that of a "levee patroller." These are people who inspect "levees." Levees are the natural and artificial barriers that protect the land from flooding. Inspecting these levees is of importance to the Netherlands due to the high risks involved in a possible levee failure. Therefore, it was desired that patrollers are able to increase their inspection knowledge and skills. This is difficult to achieve in reality, because failures occur rarely. To make it possible for patrollers to get experience, it was decided to develop a game.

In the eventual game, the goal is to find every virtual failure in a certain region (see e.g., Fig. 1). After finding a failure, players need to fill out a report, possibly measure the failure, and contact the coordinating field office to discuss the severity of the problem. If players do well, they get a high score. If they do not do well, they get a low score. Additionally, a levee breach that floods the whole region may be the result.

Figure 1: Screen shot of *Levee Patroller*.

The game was first released at the end of 2006. Since that time, the game has been limitedly used. While the people involved are still enthusiastic about the game, they are concerned about how to employ it and whether this is worthwhile. An innovative evaluatory research study seemed necessary to find some answers to this.

3. THE SETUP OF THE EVALUATION

To evaluate *Levee Patroller* it was necessary to come up with something new. I will explain why in the next

section. I will further elaborate on the idea behind the evaluation and what eventual methods were chosen.

3.1 The problem of evaluating digital games

Aside that games are troublesome to evaluate in general, because many interrelated variables have to be taken into account, from the context in which the game is played to how it is played, the big problem with fully digital games is that they cannot be done in an hour or even not in a day. For example, for *Levee Patroller* some players already took 45 minutes to go through the training stage, in which they only learn how to play the game.[1]

This means researchers should resort to a *longitudinal* setup of evaluating *digital training games*, like *Levee Patroller*. Although this seems logical, to date, only a few studies can be qualified as being longitudinal when it comes to digital games in general (e.g., Egenfeldt-Nielsen, 2007; Kato et al., 2009; Squire, 2004).

3.2 The idea behind the evaluation

Besides that games are troublesome to evaluate, a large omission in most evaluatory game studies is that they are done without a clear methodology or framework. For the evaluation of *Levee Patroller* it was decided to use the idea of *Triadic Game Design* as an evaluation framework (Harteveld, 2011). In a nutshell, this design philosophy for developing games with a serious purpose stresses that:

- The design of a game poses a multi-objective problem in a design space involving three equally important worlds: the worlds of Reality, Meaning, and Play;
- Each world has its own people, practices, interests, theories, methodologies, criteria, requirements, and Weltanschauung on how to design a game;
- That various tensions can arise between the three worlds, forcing designers to make trade-offs; and
- It is fundamental to keep these three worlds in balance to create a good game.

Translating this to the evaluation of a game, using this design philosophy as a framework means looking into each of the three worlds and how they relate to each other. With the world of Play, the world that evolves around games, aspects can be considered, such as the experience and preferences participants have with games, and how much "fun" they think it is.

With the world of Reality, the world that is closely related to the physical real world and how it is represented in a game, the realism and validity of the game can be explored, as well as what background the participants have amongst other things.

Finally, with the world of Meaning, the world that is preoccupied with bringing forth some value beyond the activity itself, the previous knowledge, the expectations, and the perceived learning can be, for example, assessed when it concerns an educational or training game.[2]

With this framework in mind, I had an idea of what I needed to look into and what would possibly make it successful. For this, so it was hypothesized, a game needed to be fun, engaging, and immersive (i.e., the world of Play), realistic and valid (i.e., the world of Reality), and be able to transfer some value beyond the game itself (i.e., the world of Meaning).

3.3 The eventual setup

Eventually, the decision was made to set up a training with *Levee Patroller* of in total three weeks. This duration was chosen, because making it longer would lower the commitment of players and increase the risk that participants are not able to participate for a duration of the training. Making it shorter would either increase the workload on behalf of the participants if the same amount of exercises are maintained or lead to a lesser amount of exercises that participants need to play, meaning they will practice much less.

A training over three weeks seems nice and desirable, but how would this work? Being a levee patroller is not a full time job. In fact, most are volunteers and have a regular job during the week. To deal with this, participants are asked to play at home.

Of course this involves a risk. Participants may not feel committed to play, do not know how to play, and may not take part in any of the scientific inquiries, such as filling out a questionnaire. To prevent this, the training has a start meeting, in which a tutorial and the first exercise is played and where they get some theoretical background information, and has an end meeting, in which the final exercise is played. In between these meetings, the participants have to play 6 exercises at home.

3.4 The methods used

For the evaluation, a combination of qualitative and quantitative methods is being used. The methods are explained below.

3.4.1 Pre- and post-questionnaires

With the questionnaires the background variables (e.g., age and years of experience) attitudes, perceptions, and knowledge is measured related to the levee inspection in general and the game. To measure the knowledge of

[1] Aside from this, it is unwanted to let participants interact with a computer all day.

[2] When the game has another kind of serious purpose, such as exploration or data collection, other aspects have to be considered.

the participants, they are asked to make sense of virtual and real pictures.

3.4.2 Pre- and post-interviews

To make more sense out of the data, to get to know the people and the organization, and to control for possible adverse effects of the questionnaire, a number of participants are interviewed upfront and afterward.

3.4.3 In-game data

The great thing about digital games is that everything can be logged. For *Levee Patroller* I decided to track every little action, and use this as an input to see how participants learned throughout the training. To get an idea of how the participants experienced each exercise, they get a relatively short questionnaire after each exercise.

This "in-game data," consisting of the gameplay data and the short questionnaires, is sent to a central server over the Internet. As a backup, the information is also saved locally on the participant's computer.

4. THE FIRST RESULTS

At the moment of writing the first of four groups of participants has just finished the training. From this group, we can retrieve some preliminary findings, related to the setup of this study and to the actual results so far. But first I will give some information about this group.

4.1 About the first group

The first group that participated in this study consisted of 38 participants. Only one of them was a female and the average age concerned 49.17 years (SD = 11.29). Of this group, 5 participants did not eventually attend the last meeting, for various reasons (e.g., too busy or forgot about it). Moreover, 2 participants decided to discontinue the training. Both found it "too complicated, not for their generation, and hated to sit behind a computer."

In total 23 participants (61%) finished all 6 exercises that they needed to play at home. 3 participants did not play at all. The remaining participants played from 1 up to 4 exercises.

4.2 Findings related to the setup

To begin with something positive, I was pleasantly surprised that so many participants completed the whole training. Few have experience with games and in fact many of them indicated that their computer literacy is somewhat limited. This means that they found the game worthwhile to invest in, as they learned from it. They confirmed this during the discussion, which took place at the end meeting.

They also mentioned that they found it pleasant to play at home, although they would have liked to get more guidance at the beginning, in playing the game and some more theoretical background that they needed to make sense of the virtual failures. From this, we see that a complete distance learning setup would most likely fail.

On the negative side, this setup is quite arduous and risky. Aside from the two meetings that each last 2.5 hours and in which only up to 15 people can participate (due to space, available laptops, and facilitation), I had to guide the participants from a distance and solve problems if needed as well during the three weeks, in which the participants had to play at home.

As for the risks involved, the data gathering methods are much dependent on ICT and the game itself, despite being tested for years, was not entirely bug free. So it happened that much of the data was lost due to errors in the game, servers that went down, and failing Internet connections. Of course, I ensured a backup, but this practically meant that I had to visit almost over 25% of the participants to retrieve the data.

4.3 Preliminary results

So far, only a partial part of the data has been analyzed. This concerns foremost the perceptions of the players about the game and about what they have learned, as indicated in the pre- and post-questionnaire.

4.3.1 Game perception results

The participants indicated on a Likert scale of 1 to 7 that in general they found the game fun to play (M = 5.19, SD = 1.36), realistic (M = 5.21, SD = 1.10), relevant (M = 5.36, SD = 0.83), and useful (M = 5.57, SD = 1.14). They also indicated that they learned from it (M = 5.36, SD = 1.19).

In the end they awarded the game a rating with a mean of 6.79 (SD = 1.32) on a scale of 1 to 10. During the discussion it became clear that many players got frustrated with some aspects of the game and they were hoping this would be improved in future versions. It was, however, not a discussion whether the game should be used. It was only stressed that not everybody may like this way of training, as many of the patrollers are not very computer minded.

Interestingly enough, the rating was strongly correlated with how much fun ($p = 0.00$) and how realistic ($p = 0.00$) they found the game, and how much they learned from it ($p = 0.00$). These aspects, fun, realism, and learning are also strongly correlated with each other. This shows that the idea of Triadic Game Design makes sense.

4.3.2 Learning perception results

Before and after the training participants had to respond to a number of statements on Likert scales of 1 to 5 or 7. These statements are related to the learning

objectives of *Levee Patroller* and to important aspects of the profession. Table 1 summarizes the statements and their results from the Wilcoxon ranks test (with $p > 0.050$ meaning that there is no difference between the pre- and post-scores).

Table 1: Learning results

Statement	Result
Expertise: how much knowledge do I have about levee inspection (on 1 to 7)?	$M_{pre} = 3.31$, $SD_{pre} = 1.32$ $M_{post} = 4.00$, $SD_{post} = 1.39$ $p = 0.015$
Impact: do I know what the consequences are of levee failures (on 1 to 7)?	$M_{pre} = 5.23$, $SD_{pre} = 1.32$ $M_{post} = 5.57$, $SD_{post} = 0.836$ $p = 0.116$
Occurrence: do I know what kind of levee failures could occur (on 1 to 5)?	$M_{pre} = 2.74$, $SD_{pre} = 0.657$ $M_{post} = 3.29$, $SD_{post} = 0.600$ $p = 0.003$
Generalizability: do I know where levee failures could occur (on 1 to 5)?	$M_{pre} = 2.74$, $SD_{pre} = 0.611$ $M_{post} = 3.25$, $SD_{post} = 0.518$ $p = 0.004$
Recognition: can I recognize failures (on 1 to 5)?	$M_{pre} = 2.68$, $SD_{pre} = 0.768$ $M_{post} = 3.25$, $SD_{post} = 0.585$ $p = 0.004$
Information: do I know to what I need to pay attention to (on 1 to 5)?	$M_{pre} = 2.54$, $SD_{pre} = 0.611$ $M_{post} = 3.04$, $SD_{post} = 0.649$ $p = 0.004$
Judgment: can I assess the severity of situation (on 1 to 5)?	$M_{pre} = 2.61$, $SD_{pre} = 0.556$ $M_{post} = 2.87$, $SD_{post} = 0.626$ $p = 0.285$
Predictability: can I predict how a failure will develop (on 1 to 5)?	$M_{pre} = 2.44$, $SD_{pre} = 0.660$ $M_{post} = 3.08$, $SD_{post} = 0.628$ $p = 0.005$
Interpretation: can I determine the cause of a failure (on 1 to 5)?	$M_{pre} = 1.91$, $SD_{pre} = 0.853$ $M_{post} = 2.61$, $SD_{post} = 0.685$ $p = 0.007$
Action: do I know what measures to take when necessary (on 1 to 5)?	$M_{pre} = 2.47$, $SD_{pre} = 0.662$ $M_{post} = 3.11$, $SD_{post} = 0.685$ $p = 0.002$

From Table 1, we can retrieve that after the training the participants significantly have more knowledge, know better what failures could occur and where, are able to recognize them better and know what to pay attention to, have a better idea of how failures develop over time, what causes them, and, finally, what action they need to take. Only what impact failures could have and how to assess the situation are not significantly different.

These results are very remarkable, but it needs to be stressed that only a few people can be characterized as "experts" in the area of levee inspection. Most are laymen and it is, therefore, not surprising that the game taught them a lot. On the other hand, this was exactly the purpose of the game and so in this regard it can be seen as successful.

5. CONCLUSION

In this paper the preliminary findings are highlighted of an innovative longitudinal evaluation of a digital game. It is innovative, because researchers and practitioners are struggling with how to use and evaluate digital games, like *Levee Patroller*, and the way this study has been setup has not been done before.

From this study we can see that the game as well as the setup is quite successful, although the setup can be characterized as arduous and risky. It has also been shown that the framework of Triadic Game Design is useful in evaluating games, as the aspects it considers are all important in determining the success of a game.

Further study needs to give more depth and meaning to these preliminary findings. With this, we may get more strong empirical evidence that games work and some fundamental insights in how to use and evaluate games.

REFERENCES

De Caluwé, L., Hofstede, G.J., & Peters, V. (Eds.). (2008). *Why Do Games Work? In Search of the Active Substance*. Deventer, the Netherlands: Kluwer.

De Freitas, S. (2006). *Learning in immersive worlds: a review of game-based learning*. Bristol, UK: Joint Information Systems Committee.

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

Egenfeldt-Nielsen, S. (2007). *Beyond Edutainment: The Educational Potential of Computer Games*. London, UK: Continuum Press.

Harteveld, C. (2011). *Triadic Game Design: Balancing Reality, Meaning and Play*. London, UK: Springer.

Kato, P. M., Cole, S. W., Bradlyn, A. S., & Pollock, B. H. (2009). A Video Game Improves Behavioral Outcomes in Adolescents and Young Adults with Cancer: A Randomized Trial. *Pediatrics, 122*(2), 305–317.

Ke, F. A (2009). A Qualitative Meta-analysis of Computer Games as Learning Tools. In R.E. Ferdig (ed.), *Handbook of Research on Effective Electronic Gaming in Education* (Vol. I, pp. 1-32). Hershey, PA: Information Science Reference.

Kriz, W.C., & Hense, J.U (2006). Theory-oriented for the design of and research in gaming and simulation. *Simulation & Gaming, 37*(2), 268-283.

Squire, K.D. (2004). *Replaying History: Learning World History through Playing Civilization III*. Unpublished Dissertation, Indiana University, Bloomington IN.

Winn, B.M. (2009). The Design, Play, and Experience framework. In R.E. Ferdig (ed.), *Handbook of Research on Effective Electronic Gaming in Education* (Vol. III, pp. 1010-1024). Hershey, PA: Information Science Reference,.