Tracking the UFO's Paths: Using Eye-Tracking for the Evaluation of Serious Games

Michael D. Kickmeier-Rust, Eva Hillemann, and Dietrich Albert

Graz University of Technology, Brueckenkopfgasse 1/6,
8020 Graz, Austria

{michael.kickmeier-rust,eva.hillemann,dietrich.albert}@tugraz.at

Abstract. Computer games are undoubtedly an enormously successful genre. Over the past years, a continuously growing community of researchers and practitioners made the idea of using the potential of computer games for serious, primarily educational purposes equally popular. However, the present hype over serious games is not reflected in sound evidence for the effectiveness and efficiency of such games and also indicators for the quality of learner-game interaction is lacking. In this paper we look into those questions, investigating a geography learning game prototype. A strong focus of the investigation was on relating the assessed variables with gaze data, in particular gaze paths and interaction strategies in specific game situations. The results show that there are distinct gender differences in the interaction style with different game elements, depending on the demands on spatial abilities (navigating in the three-dimensional spaces versus controlling rather two-dimensional features of the game) as well as distinct differences between high and low performers.

Keywords: Game-based learning, serious games, learning performance, eye tracking.

1 Introduction

Over the past years, digital educational games (DEG) got in the focus of educational research and development. Several commercial online platforms are distributing (semi-) educational mini games (e.g., www.funbrain.com or www.primarygames.com) for the young children, *Nintendo DS*’ “*Dr Kawashima’s Brain Training: How Old Is Your Brain?*” is a best seller, and a growing number of companies concentrate on educational simulation and game software. Today’s spectrum of learning games is broad, ranging from so-called using off-the-shelf games (COTS) in educational setting to specifically designed curriculum-related learning games. A classification on the basis of the psycho-pedagogical and technical level of games is proposed by [1, 2, 3].

As rich as the number of games is, is also the number of initiatives and projects in this area. Historically, among the founders of the recent hype over game-based learning are doubtlessly Mark Prensky and Jim Gee. Mark Prensky published in 2001 [4] his ground breaking book “*Digital Game-based Learning*”. His idea of game-based learning focuses on the concept of digital natives. He argues that the

omnipresence of “twitch speed” media such as MTV and computer games have emphasized specific cognitive aspects and de-emphasized others which, in turn has changed the educational demands of this generation. Jim Gee focused on learning principles in video games and how these principles can be applied to K-12 education [5, 6]. His thematically origin is the idea that (well-designed) computer games are very good at challenging the players, at preserving them, and at teaching how to play. On this basis he identified several principles for successful (learning) game design (e.g., learners must be enabled to be active agents or producers and not just passive recipients or consumers; Gee, 2005, p. 6, [7]). Another pioneer was for example David Shaffer in the context of applying regular entertainment computer games (COTS) in education [8].

In the USA a strong focus of research and development has a military background (this strong background is for example mirrored by governmental institutions such the *Department of Defence Game Developers' Community*; www.dodgamecommunity.com), resulting in famous games such *America's Army*. In Europe widely a more civil approach is pursued, strongly driven by the *European Commission*. Leading-edge projects are, for example, ELEKTRA (www.elektra-project.org), 80Days (www.eightdays.eu), TARGET (www.reachyourtarget.org), mGBL (www.mg-bl.com), Engagelearning (www.engagelearning.eu), LUDUS (www.ludus-project.eu), or the special interest group SIG-GLUE (www.sig-glue.net).

A noteworthy initiative comes from the *Network of Excellence GALA* (Games and Learning Alliance; www.galanoe.eu), an alliance where Europe's most import players in the serious games sector attempt to streamline the fragmented field and to increase scientific and economic impact.

The recent hype over game-based learning is based on the natural, almost self-evident, link between pedagogical and didactic guidelines and theories and the characteristics of modern computer games [9]. Just as one example, the rich virtual environments of immersive games enable very naturally a meaningful and plausible context for learning and therefore support deeper learning processes. Evidently, computer games have the potential to make knowledge and skills a desirable and valuable asset and they have the potential to make learning a meaningful and important task. By this means, there exists the justifiable hope that also those learners can be reached who are not necessarily keen on learning and who may be reached with other educational measures. In this sense, serious games can be way more than just “chocolate covered broccoli” [10]. Recent research even argues that the immersion and gaming experience impacts the neurotransmitter systems and thus alternating cognitive functions and learning capacity [11].

Despite the great many of advantages of serious games and despite the large amount of related research and development activities, as outlined above, educational computer games, specifically what we termed competitive games, have not become a serious business case so far. This impression coincides with the view of many researchers in the field of game-based learning, who are arguing that those games most often are still in their infancy from a scientific and pedagogical perspective (e.g., [12], [13]). Major challenges for research, design, and development are seen, for example, in finding an appropriate balance between gaming and learning activities [3] or finding an appropriate balance between challenges through the game and abilities of the learner (e.g., [14]). One of the most important challenges for research concerns

the core strength of games, which can be summarized with their enormous intrinsic motivational potential. On the one hand, maintaining a high level of motivation requires an intelligent and continuous real-time adaptation of the game to the individual learner, for example, a continuous balancing of challenge and ability and of problems and learning progress. Essentially, this corresponds to the concept of flow – a highly immersed experience when a person is engaged in a mental and/or physical activity to a level where this person loses track of time and the outside world and when performance in this activity is optimal [15].

2 Evaluating Serious Games

An ultimate challenge in the context of serious games, however, is a scientifically sound formative and summative assessment and evaluation. On the one hand, the gaming aspect must be addressed on the other hand the learning aspects. Both, unfortunately, must be considered more than the sum of its components and, more importantly, raise in part contracting demands on the metrics of “good” and “educational” games.

When considering the criteria of “good” software in general and conventional learning/teaching software in particular, these are heavily dominated by the idea of *performance* with the software (in terms of effectivity and efficiency); the ISO standards, for example, are promoting this kind of view. Accordingly, also the metrics and heuristics focus on performance aspects (e.g., error prevention, minimization of task time, minimization of cognitive load, and so on; cf. [16]). In contrast, criteria of “good” games focus on aspects of fun, immersion, pleasure, or entertainment. Heuristics concern, for example, the visual quality, the quality of the story, fairness, curiosity, and also a certain level of challenge (in the sense of task load or cognitive load); Atari founder Nolan Bushnell states in a famous quote, “*a good game is easy to learn but hard to master*”.

Evaluating serious games must consider both, performance aspects as well as recreational aspects. On the one hand, such game is supposed to accomplish a specific goal, that is, teaching a defined set of knowledge/skills/competencies. This goes along with context conditions such as justifiability of development and usage costs, comparability with conventional learning material in terms of effectivity and efficiency, or societal concerns. On the other hand, the great advantages of the medium game are only existent if the game character is in the foreground, that is, being immersion, having fun, experiencing some sort of flow.

In the past, several approaches to measure serious games were published. De Freitas and Oliver [17] proposed an evaluation framework which focuses on (i) the application context of a serious game, (ii) learner characteristics, (iii) didactical/pedagogical aspects, and (iv) the concept of “*diegesis*”, the extent and quality of the game story’s world. Besides such general frameworks, also more specific scales were introduced. An example is the *eGameFlow* approach [12], which attempts to measure flow experience in educational games by criteria such as challenge, amount of concentration, level of feedback, and so forth. A more recent and complete approach to gauge “good” serious games comes from [18]. The EVADEG framework concentrates on the aspects (i) learning performance, (ii) gaming experience, (iii) game usability, and (iv) the evaluation of adaptive features.

The latter is a highly important yet oftentimes neglect factor. In the tradition of adaptive, intelligent tutoring systems, some modern serious games adapt autonomously to the needs, preferences, abilities, goals, and the individual progress of the players [14]. As emphasized by Weibelzahl [19], a scientific correct evaluation of adaptive systems is difficult because, in essence, personalization and adaptivity means that each user/player potentially receive different information, in a different ways, in a different sequence, and in a different format.

The introduced frameworks and approaches offer a valuable basis for evaluating serious games. A key problem, however, is how to measure all those aspects in a possibly unobtrusive, reliable, valid, and methodologically correct way. In the present paper, we took up the ideas of EVADEG and introduce eye tracking as means of studying (i) the usability of a game, (ii) the extent of learner satisfaction, and most importantly (iii) the learning efficacy. More concretely, we utilized the considerations and methods to evaluate a learning game prototype which was developed in the course of a European research project (80Days).

2.1 Eye Tracking

Observing eye movements has a long tradition in psychology in general and the field of HCI/usability in particular. El-Nasr and Yan [20] describe how perceptive (i.e., bottom up) and cognitive (i.e., top-down) processes orchestrate in the context of 3D videogames. Accordingly, while saliency of objects can grab players' attention, goal-orientation (top-down) in games is more effective for attracting attention. The big challenge for the authors was to develop a new methodology to analyze eye tracking data in a complex 3D environment, which differed considerably from the stimuli used in eye-tracking experiments conducted until then [21].

Basic variables in eye tracking studies are fixations (processing of attended information with stationary eyes) and saccades (quick eye movements occurring between fixations without information processing) [22]. The sequence of fixations establish scan paths through a visual field. Although such eye tracking measures are commonly used, their interpretations remain malleable. To give an example, an important indicator for the depth of processing is fixation duration; the longer a stimulus is attended, the higher is cognitive load and the deeper is the processing]. In immersive games such relationship may not be quite as stable and valid as the might be for other visual processing tasks. In the work of Jennett [23], for example, who investigated the immersion in a game using eye tracking a decrease of fixations per second in the immersive condition was found as compared to an increase in a non-immersive control condition. Jennett argued that in an immersive game the attention of the players becomes more focused on game-related visual components and therefore less "vulnerable" to distracting stimuli.

3 An Eye Tracking Study on Learning and Gaming

3.1 The 80Days Game Prototype

The investigated game prototype was developed in the context of the European 80Days project (www.eighytdays.eu). The game is teaching geography for a target

audience of 12 to 14 year olds and follows European curricula in geography. In concrete terms, an adventure game was realized within which the learner takes the role of an Earth kid at the age of 14. The game starts when a UFO is landing in the backyard and an alien named Feon is contacting the player. Feon is an alien scout who has to collect information about Earth. The player wants to have fun by flying a UFO and in the story pretends to be an expert in the planet earth. He or she assists the alien to explore the planet and to create a report about the Earth and its geographical features. This is accomplished by the player by means of flying to different destinations on Earth, exploring them, and collecting and acquiring geographical knowledge. The goal is to send the Earth report as a sort of travelogue about Earth to Feon's mother ship. In the course of the game, the player discloses the aliens' real intentions – preparing the conquest of the earth – and reveals the “real” goal of the game: The player has to save the planet and the only way to do it is to draw the right conclusion from the traitorous Earth report. Therefore the game play has got two main goals: (1) to help the alien to complete the geographical Earth report, and (2) to save the planet, which is revealed in the course of the story, when the player realizes the true intention of the alien. Figure 1 gives some impressions of the game. Details are given by [14].

Fig. 1. Screenshots of the 80Days demonstrator game; an action adventure – on the basis of an Alien story – to learn geography according to European curricula

3.2 Study Design

The study presented in this paper is only one of a long sequence of experiments in several European countries, to evaluate to demonstrator game and to conduct in-depth research on the relationships and mechanisms in the context of using computer games for learning. Due to the vast complexity of this research battery, we must focus on a rather concise snapshot of this work only. The present results are based on data of 9

Austrian children, 4 girls and 5 boys. The participant's age ranged between 11 and 16 years with the average of 13 years ($SD = 1.61$).

3.3 Material and Apparatus

To record gaze information, the *Tobii 1750* eye tracker was used, a device that works with infrared cameras and therefore enables a fully unobtrusive recording of gaze information (Figure 2a). For the pre and post assessments of knowledge, we utilized a paper-pencil knowledge test, and in addition motivational, usability-related, and attention-related scales were issued. For the analyses, we selected three individual scenes: Flying to Budapest, instructive cockpit scene in Budapest, terraforming simulation (cf. Figure 2b).

Fig. 2. Panel a (left) shows an image of the eye tracking set up. Panel b (right) a screen shot of the game's terraforming simulation. The colored rectangles indicate predefined areas of interest (AOI) for gaze data analyses.

4 Results

4.1 Learning Performance

The average score of the knowledge test prior to the gaming session was 32.33 ($SD=9.45$) and that of the posttest 39.00 ($SD =10.22$). The difference is statistically significant ($T=-3.814$, $df=8$, $p=0.005$). For girls the average score was 26.25 ($SD=11.09$) and 31.25 ($SD =10.63$) respectively. There is no significant difference between pre and posttests. For boys the average score in the pretest was 37.20 ($SD=4.44$) and that for the posttest 45.20 ($SD=4.02$), which is a significant difference between these two test scores ($T=-6.136$, $df=4$, $p=0.004$). The results are illustrated in Figure 3.

4.2 Eye Movements

For this investigation it was important to get information on how much time in the three different situations participants spend while playing the game and on which parts their eyes fixed. By this way the relative fixation numbers, the duration of the situation as well as the total duration, and the saccade lengths were analyzed.

Fig. 3. Results of the knowledge tests before and after playing the learning game

These results imply that females spend more time in playing the game. Females' total duration is $M=1018.86$ with $SD=102.58$ in contrast to males' total duration with $M=868.80$ with $SD=280.72$. Especially on the simulation situation females spend more time with $M=913.20$ ($SD=142.28$) than males with $M=726.41$ ($SD=278.41$).

A key question of this investigation is whether good and low performers in terms of learning have distinct gaze patterns/scan paths. Participants who learned more spend more time in playing the game with $M=940.10$ with $SD=23.52$ in contrast to participants who learned less with $M=781.79$ with $SD=344.46$. These results were present in the different situations as well. For participants who learned more the duration of the flying situation is higher with $M=100.09$ ($SD=79.33$) than for the other group with $M=59.71$ ($SD=31.75$). The duration for the instruction for a high learning effectiveness is $M=70.97$ ($SD=29.33$) and $M=45.53$ ($SD=20.28$) for a low learning effectiveness. In the simulation situation the duration is $M=769.04$ ($SD=77.21$) for better learners and $M=676.55$ ($SD=362.57$) for persons who learned less. Regarding the fixation number/sec for the three different situations participants who learned more have smaller values than participants who learnt less. The total fixation number/sec for better learners is $M=0.72$ ($SD=0.04$), in the flying situation the fixation number/sec is $M=2.08$ ($SD=0.41$), in the Instruction situation is $M=1.94$ ($SD=0.85$), and in the Simulation situation is $M=2.22$ ($SD=0.34$). For the other group the total fixation number/sec is $M=0.77$ ($SD=0.12$), the flying fixation number/sec is $M=2.40$ ($SD=0.33$), the instruction fixation number/sec is $M=2.22$ ($SD=0.50$), and the simulation fixation number/sec is $M=4.48$ ($SD=4.14$). Regarding the first situation participants who learnt more have a longer average fixation length with $M=0.45$ ($SD=0.10$) than persons who learnt less with $M=0.40$ ($SD=0.04$). Regarding the second situation better learner also have a longer average fixation length with $M=0.57$ ($SD=0.37$) than the others who have an average fixation length with $M=0.42$ ($SD=0.06$). Regarding the third situation a longer average fixation length with $M=0.45$ ($SD=0.06$) is found for good learners. Participants who learnt less have an average fixation length with $M=0.35$ ($SD=0.21$) in the simulation situation. In the flying situation the saccade lengths are higher for participants who showed a higher learning

effectiveness with $M=73.44$ ($SD=29.64$) in contrast to the other group where $M=53.05$ ($SD=27.13$). In the second and third situation both groups have nearly the same average saccade lengths.

Although the descriptive data shows some differences, MANOVA results showed no significant differences for gender and attention for duration of playing, fixation rate, and saccade lengths. Only learning effectiveness has a significant effect on the duration of the simulation situation with $F(1)=186.652$, $p=0.047$. Regarding the fixation lengths in the different situations especially in the simulation part significant differences could be found. On the one hand gender has a significant effect on the fixation length with $F(1)=195.77$, $p=0.045$. On the other hand the learning effectiveness has a significant effect with $F(1)=372.982$, $p=0.033$.

4.3 Areas of Interest

Areas of Interest (AOI) are particular display elements which are predefined by the researcher. AOI analysis is used to quantify gazed data within a defined region of the visual stimulus. The number of fixations on such a particular display element should reflect the importance of that element. More important display elements will be fixated more frequently.

When aiming at evaluating serious games, such information is crucially important since it provides very clear indications of which regions on the screen are attended (sufficiently) and, therefore, whether all instructional aspects are attended.

The most distinct results of these analyses are that players with high learning performance have in total a larger number of fixations (Figure 4). By the prototypical example of AOI 6, which was the most often attended one, we can say that children who learnt more spend 38.67 % ($SD=0.65$) on it and children who are not so good learners spend 32.23% ($SD=4.32$) on it (Figure 4). Persons with a higher attention value spend 36.24% ($SD=2.77$) on AOI6 in contrast to persons with a lower attention who spend more time with 38.53% ($SD=0.86$) on it.

Similarly, we found that high performers showed significantly longer saccades in the flying scene of the game. This is an indication that children who learnt well in general exhibited a more “calm” and smooth distribution of fixations on the screen.

Fig. 4. Average fixation duration and saccade length

5 Summary

The empirical results indicate that the children can benefit from playing computer games for learning purposes. The most distinct finding we presented here is the fact that extreme groups such as high and low performers exhibit different visual patterns. While the good learners scan the visual field evenly with longer saccades and attend relevant areas on the screen more frequently and in a more stable fashion.

The results of our investigation also show that there are distinct gender differences in the interaction style with different game elements, depending on the demands on spatial abilities (navigating in the three-dimensional spaces versus controlling rather two-dimensional features of the game) as well as distinct differences between high and low performers in terms of learning. In addition to the comparisons on the level of participants, on the basis of gaze density maps aggregated from the data in combination with qualitative interviews with the subjects, we identified design recommendations for further improvements of the game prototype in particular as well as games in general.

Finally, our study showed that using eye tracking can be successfully applied to measure critical aspects with regard to the quality of serious games.

Acknowledgements. The research and development introduced in this work is funded by the European Commission under the seventh framework programme in the ICT research priority, contract number 215918 (80Days, www.eightydays.eu).

References

1. Kickmeier-Rust, M.D.: Talking digital educational games. In: Kickmeier-Rust, M.D. (ed.) *Proceedings of the 1st International Open Workshop on Intelligent Personalization and Adaptation in Digital Educational Games*, Graz, Austria, October 14, pp. 55–66 (2009a)
2. de Freitas, S.: Learning in immersive worlds. A review of game-based learning (2006), http://www.jisc.ac.uk/media/documents/programmes/elearning_innovation/gaming%20report_v3.3.pdf (retrieved August 28, 2007)
3. Van Eck, R.: Digital game-based learning. It's not just the digital natives who are restless. *Educause Review*, 17–30 (March/April 2006)
4. Prensky, M.: *Digital game-based learning*. McGraw-Hill, New York (2001)
5. Gee, J.P.: *What video games have to teach us about learning and literacy*. Palgrave Macmillan, New York (2003)
6. Gee, J.P.: *What video games have to teach us about learning and literacy* (2nd revised and updated edn.). Palgrave Macmillan, New York (2008)
7. Gee, J.P.: Learning by design: Good video games as learning machines. *E-Learning and Digital Media* 2(1), 5–16 (2005); Gee, J. P.: *What video games have to teach us about learning and literacy*. Palgrave Macmillan, New York (2003)
8. Shaffer, D.W.: *How computer games help children learn*. Palgrave Macmillan, New York (2006)
9. Kickmeier-Rust, M.D., Mattheiss, E., Steiner, C.M., Albert, D.: A psycho-pedagogical framework for multi-adaptive educational games. To appear in *International Journal on Game-Based Learning* 1(1), 45–58 (in press)

10. Habgood, J.: Wii don't do edutainment. In: *Proceedings of Game-based Learning 2009*, London, UK, March 19-20 (2009)
11. Demetriou, S.: Motivation in computer games: The impact of reward uncertainty on learning. In: Gómez Chova, L., Martí Belenguer, D., Candel Torres, I. (eds.) *Proceedings of Edulearn 2010*, Barcelona, Spain, July 5-7 (2010)
12. Fu, F.-L., Su, R.-C., Yu, S.-C.: EgameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education* 52(1), 101–112 (2009)
13. Oblinger, D.: Games and learning. *Educause Quarterly Magazine* 29(3), 5–7 (2006)
14. Kickmeier-Rust, M.D., Albert, D.: Micro adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning* 26, 95–105 (2010)
15. Csikszentmihalyi, M.: *Flow: The psychology of optimal experience*. Harper and Row, New York (1990)
16. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L. (eds.) *Usability Inspection Methods*. John Wiley & Sons, New York (1994)
17. de Freitas, S., Oliver, M.: How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers and Education Special Issue on Gaming* 46, 249–264 (2006)
18. Law, E.L.-C., Kickmeier-Rust, M.D., Albert, D., Holzinger, A.: Challenges in the development and evaluation of immersive digital educational games. In: Holzinger, A. (ed.) *USAB 2008. LNCS*, vol. 5298, pp. 19–30. Springer, Heidelberg (2008)
19. Weibelzahl, S., Lippitsch, S., Weber, G.: Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz* 3(2), 17–20 (2002)
20. El-Nasr, M.S., Yan, S.: Visual Attention in 3D Video Games. In: *Proceedings of ACE 2006*. Hollywood, California (2006)
21. Law, E.L.-C., Kickmeier-Rust, M., Albert, D., Holzinger, A.: Challenges in the development and evaluation of immersive digital educational games. In: Holzinger, A. (ed.) *USAB 2008. LNCS*, vol. 5298, pp. 19–30. Springer, Heidelberg (2008)
22. Land, M.F.: Eye Movements and the Control of Actions in Everyday Life. *Progress in Retinal and Eye Research* 25, 296–324 (2006)
23. Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A.: Measuring and Defining the Experience of Immersion in Games. *International Journal of Human Computer Studies* 66(9), 641–661 (2008)