Effects of Playing a Serious Game: A Comparison of Different Cognitive and Affective Measures

Erik D. van der Spek

Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands.
e.d.vanderspek@tue.nl

Introduction

Through a number of experiments we tried to determine general game design rules that improved the efficacy of a certain serious game. The game was called *Code Red Triage*, a total conversion modification of Half-Life 2, and was purposefully made in order to systematically empirically test the aforementioned game design rules, which in turn were based on better aligning the game with the player's cognitive system.

In Code Red Triage, a player would take on the role of a medical first responder, who has to learn how to perform a mobility (sieve) triage, or categorizing the many victims of a mass casualty event according to urgency of needed medical attention. The player started at the central hall of a train station where he (or she) received a message that a terrorist bomb strike had taken place on one of the subway platforms. The player then had to find the way to the subway platform via a number of winding corridors. Arriving at the platform, the player could triage victims by walking up to them and entering the triage menu; here, the player could choose a number of triage actions and assign one of four triage categories. From the feedback provided after performing the triage actions, as well as after categorizing the victim, the player had to deduce, iteratively through nineteen victim cases, what the correct procedure of the triage was. The feedback was presented in the form of short text snippets and a game score which detailed how well the player had performed the triage procedure on the current victim.

A total of six studies were performed with the game: a pilot study, a media comparison study where the efficacy of learning with the game was compared to learning with a conventional PowerPoint slideshow, and four value added experiments where the effect of one or more interventions were compared with a control group. The experiments were based on Mayer's cognitive theory of multimedia learning [1], which states that one goes through three cognitive processes in order to learn something: the selection of relevant over irrelevant information, the organization of the relevant information into coherent knowledge structures, and the integration of these knowledge structures into prior knowledge. The four experiments were: (1) the introduction of auditory and visual attention cues to ameliorate the selection of relevant information; (2) a 2-way comparison between just-in-time and just-in-case option complexity versus a massed or a spaced approach to victim complexity, to see which approach led to improved organization of information; (3) the introduction of an adaptation engine where, if the player scored well on a certain victim, similar victims were deleted, to see whether this led to more efficient information organization; and (4) the introduction of surprising events to stimulate better knowledge integration. A complete description of the experiments, as well as the results, can be seen in [2]. Here instead, we will focus on the used measurement instruments and how the participants' scores correlate, as an indication of their validity.

Game experience

Any experiment on learning with games may be confounded by a participant's prior game experience for a number of reasons. For instance, a novice to games and 3d games especially, may need to afford considerably more mental effort to spatially navigate an avatar through a game world than a person that has ample experience in this, which in turn could lead to less cognitive capacity that may be used for learning. Additionally, experienced gamers will be more accustomed to ways in which a game conveys important information to the learner.

One of the problems we faced was how to measure prior game experience. A conventional way would be to ask how many hours a participant spends playing games per week; however there are three problems with this measurement. Firstly, a person may have spent ample time playing games in the past, but has been too busy to play games lately, in which case he would be incorrectly categorized as an inexperienced gamer. Secondly, it is unclear what type of game experience it measures; a person may play many hours of solitaire on an office computer but still be unable to spatially navigate a 3d environment. Lastly someone may not play games, but like to read about them extensively.

We had purposefully made the way from the train station to the subway platform labyrinthine, in order to measure the time it took a participant to reach the platform, under the presupposition that experienced gamers would have less problems with spatially navigating the virtual environment, and therefore arrive quicker. However, in a pilot study we found this to be completely uncorrelated with time spent gaming [$r(19) = 0.13$, n.s.]. By surveying the paths taken and asking the participants afterwards, we discovered that experienced gamers were more inclined to wander off and explore the game world, whereas inexperienced gamers were more inclined to follow the signs towards the subway platform. In the end we therefore settled with a self-report measure that asked whether the player a) rarely played games, b) sometimes played games, or c) considered themselves a gamer. This was highly correlated to time spent gaming in the four experiments [$r(41) = 0.85$, $r(56) = 0.69$, $r(28) = 0.67$, $r(41) = 0.70$, respectively, all p's < 0.001] and, for the first (attention cueing) experiment was able to discern a significant effect of prior game experience on the ability to understand and correctly use the visual cues.

Learning

Arguably the most important part of measuring the efficacy of serious games for learning is measuring how much is learned. Three measures for learning were used: an in-game score, a pen-and-paper knowledge test and a structural knowledge assessment.

In-game score. The in-game score was a measure for how well the player performed the triage procedure on a victim, ranging from 1 to 100 for any victim, to a total of 1900 for all 19 victims. Aligning the thing to be learned with the core game mechanics leads to better learning [3]; but at the same time it also makes for an excellent stealth assessment [4], where the player's learning progress can be continuously monitored without him (or her getting) disengaged from the game. A possible problem with using the in-game score as a measure for learning could however be that choosing the wrong option will lead to a lower score, but provides the player with the opportunity to learn from his mistake, possibly even more so than a person that chose the right answer and doesn't reflect on his actions. The performance in the game may therefore not be an accurate depiction of actual learning after the game is finished.

Pen-and-paper knowledge test. We consequently also had the participants perform a knowledge test before and after the game. The questions were in the form of 'given a victim where you just measured x, then the next step in the primary triage procedure would be?'. Eight of the questions were verbal only, eight additional questions also contained a screenshot of a victim as it would appear in the game. The pictorial versions of the questions were added because, in research done by [5], it was found that people who played a game remembered visual information better than text. In the four value-added experiments the participant had to choose one of four possible answers per question, in the comparison experiment with a PowerPoint presentation, each question had one of six possible answers and an additional retention test was performed a week after the intervention.

Structural knowledge assessment. While pen-and-paper knowledge tests are a good way to measure the learning of declarative knowledge and how well a participant can replicate this knowledge afterwards, it is ill-suited to measure deeper learning, or how well the learner stores knowledge structurally. We therefore used another instrument, the structural knowledge assessment, specifically to measure deeper learning. For this, the participant had to rate pairs of concepts of the triage procedure on their degree of relatedness. From these relatedness ratings, a computer program (PCKNOT) uses the Pathfinder algorithm to create graphs, where concepts that are closely related to each other only have one or a few links separating them, whereas concepts that are unrelated appear further away in the graph [6]. These graphs can subsequently be compared with those

of an expert leading to a similarity score [7]. This score is then a measure of how well a participant's knowledge structure resembles that of an expert and an indication of transfer of training [6].

Correlations and conclusions. Due to space constraints, we will only report the interesting correlations between different measurement types. For the four value-added experiments they are printed in Table 1. In addition, in the study comparing the game to a PowerPoint, both the verbal knowledge test and the pictorial knowledge test were correlated with the in-game score after playing the game [$r(48) = 0.90$, $p < 0.01$; $r(48) = 0.60$, $p < 0.01$ resp.], as well as for the delayed knowledge tests [$r(47) = 0.78$, $p < 0.01$; $r(47) = 0.75$, $p < 0.01$ resp.]. The structural knowledge assessment after playing the game was correlated with both the verbal knowledge test and the pictorial knowledge test after the game [$r(47) = 0.30$, $p < 0.05$; $r(47) = 0.30$, $p < 0.05$], but not with the in-game score.

Table 1. Correlations between the different learning metrics over the four different experiments. SKA = Structural Knowledge Assessment, * = $p < 0.05$, ** = $p < 0.01$

	Score – Verbal knowledge test	Score – Pictorial knowledge test	SKA – Verbal knowledge test	SKA – Pictorial knowledge test	SKA - Score
Experiment 1	0.573**	0.744**	0.364*	n.s.	n.s.
Experiment 2	0.291*	0.505**	n.s.	0.363**	0.293*
Experiment 3	0.546**	n.s.	n.s.	n.s.	n.s.
Experiment 4	0.409**	n.s.	0.378**	n.s.	n.s.

Given the results mentioned above and in Table 1, there doesn't seem to be a measure that is entirely redundant. The verbal knowledge test is always correlated with the in-game score, which corroborates the notion that the in-game score is a valuable measure for learning in the game. Interestingly, the pictorial knowledge test seems to more closely measure what was learned in the game in the first two experiments, but then is uncorrelated to the in-game score in the latter two experiments. The reason for this ostensible switch is unclear. In addition, Table 2 shows the measures that revealed a significant effect of the experimental condition.

In Table 2 one can see that the structural knowledge assessment has been able to pick up effects of an experimental condition that a pen-and-paper knowledge test was unable to discover, meaning that game design decisions sometimes activate the deeper processing capabilities of a player without affecting more superficial learning, or even the in-game score. Given the previous considerations, we therefore advise the usage of all three different measures for learning.

Table 2. Measures that showed a significant effect of condition, * = significant effect but unimportant due to differences in the way the condition was set up

	Verbal knowledge test	Pictorial knowledge test	Struc. Knowledge Assessment	Score
Experiment 1	n.s.	n.s.	Sig.	Sig.
Experiment 2	n.s.	n.s.	n.s.	Sig.*
Experiment 3	Sig.	Sig.	Sig.	Sig.*
Experiment 4	n.s.	n.s.	Sig.	n.s.

Affect

We lastly want to briefly discuss the affective measure(s) that we used. Primarily we wanted to test whether the game design interventions did not negatively influence the engagement of the game, as some cognitive guidance techniques could take the player out of the experience. We therefore needed to measure both the enjoyment/engagement of the player as well as their feelings of presence. For this we used the engagement subscale of the ITC Sense Of Presence Inventory (SOPI) [8], as this seemed to give a good mix of the two while

not overburdening the participant with questionnaires. However the Cronbach alpha's were poor (< 0.60) for two of the four experiments and in between eight experimental interventions, only one saw a small significant improvement in engagement. We offer three main reasons for this: One, small cognition-based game design interventions do not impact engagement. Two, the scale is rated on a five point Likert scale and participants rarely used the extremes, making any possible effect very subdued. Two, in letting the participant play only one condition of the serious game, it is unclear what the user's reference point was to indicate how engaged he was. Was it another form of instruction, in which case the game could have been appraised favorably, or was it a (entertainment) game, in which case our comparably dull serious game may have suffered. We therefore propose that researchers use 7 point Likert-scales as well as give the player a clear referent to measure the serious game (condition) against.

References

1. Mayer, R.E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning*, 31-49. New York, NY: Cambridge University Press.

2. Van der Spek, E.D. (2011). *Experiments in serious game design: a cognitive approach*. Doctoral dissertation.

3. Habgood, M.P.J., Ainsworth, S.E. (2011). Motivating children to learn effectively: exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences* **20**(2), 169-206.

4. Shute, V.J., Ventura, M., Bauer, M.I., Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects,* 295-321. New York, NY: Routledge.

5. Belanich, J., Sibley, D.E., Orvis, K.L. (2004). *Instructional characteristics and motivational features of a PC-based game* (Research Report No. 1822). Alexandria, VA: U.S. Army Research Institute.

6. Wouters, P., Van der Spek, E.D., Van Oostendorp, H. (2011). Measuring learning in serious games: a case study with structural assessment. *Educational Technology Research and Development* **59**(6), 741-763.

7. Thompson, L.A., Gomez, R.L., Schvaneveldt, R.W. (2000). The salience of temporal cues in the developing structure of event knowledge. *The American Journal of Psychology* **113**(4), 591-619.

8. Lessiter, J., Freeman, J., Keogh, E., Davidoff, J. (2001). A cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence: Teleoperators and Virtual Environments* **10**(3), 282-297.