Experiments in serious game design
a cognitive approach

SIKS Dissertation Series No. 2011-36
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Game research for training and entertainment

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO).

ISBN: 978-90-393-5658-6
Copyright © 2011 Erik van der Spek

Experiments in serious game design
a cognitive approach

Experimenten in serious game design
een cognitieve benadering

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 26 oktober 2011 des middags te 12.45 uur

door

Eric David van der Spek

geboren op 12 september 1982
te Gouda

Promotor: Prof.dr. J.-J.Ch. Meyer

Co-promotor: Dr. H. van Oostendorp

Preface

I tilted the joystick to the right and a few blocky pixels depicting a miniature human moved in unison. This revelatory experience is my earliest memory. I was about three years old and my parents had just bought a Commodore 64 for us to play with. From then and until this day I was hooked. As I grew up, the fascination for games changed from simply enjoying the game to trying to understand why they entertained me so.

If the ability and propensity to tell each other stories is what sets Homo sapiens apart from other species, then videogames may well be one of our most interesting achievements yet. In part because games are interactive and therefore the stories can develop emergent properties, but also because, probably more so than most other entertainment media, every part of the game world is laboriously constructed and placed there for a reason, where the designer has to think in advance about the profound impact bits of the interactive world can have on the player. During my research I discovered that the game designer not only has to think about how game elements impact the player emotionally, but also cognitively.

When I was still young, other boys of my age had famous football players as their idols. For me, it was Shigeru Miyamoto, the main game designer of Nintendo, responsible for such iconic games as Donkey Kong, Mario and Zelda, whom I feel has done many things right in stimulating the player's cognition. How elements in a game's design influence the way players interact with a game fascinates me greatly and I hope that I was able to communicate a little bit of this fascination through my dissertation.

Consequently, I am very grateful that I was allowed to perform scientific research into what already was one of my favorite pastimes (as a games researcher, I sometimes felt a puzzling sensation that procrastination actually was work, and work was procrastination). For this I am indebted to my daily supervisor Herre, as well as for guiding me into the wonderful world of cognitive psychology. In addition, I would like to thank my promoter John-Jules Meyer, my colleagues at the Human Media Interaction group of Utrecht University, and Pieter and Joske in particular for their support in writing my first articles and turning me into something resembling an actual scientist. Thank you to all my family and friends, who've supported me for so long, but most of all I'd like to thank my wife, Inge, who has been my guiding light in everything. This dissertation is dedicated to you.

Table of Contents

Chapter 1

General Introduction

1.1 Learning games

There is a moment in the platform game *Super Mario Galaxy 2* that passes by unnoticed to anyone but the keenest observer. In World 1-4, Fluffy Bluff Galaxy, the eponymous protagonist (Super) Mario first encounters the cloud mushroom on top of a small island surrounded by water, turning Mario into Cloud Suit Mario. In this form Mario, among other things, can create cloud platforms on which he can stand to reach higher areas in the game. The moment Mario is transformed into Cloud Suit Mario for the first time, an information bubble shows up in the bottom right of the screen that encourages the player to create a platform by shaking the motion controller.

When the first cloud platform has been created, the camera then suddenly, but subtly, pans out to show a coin block hanging high up in the air. The player, wanting to reach that block in order to get a few coins and thereby a higher game score, jumps up and creates a new platform to stand on. Another jump and a third platform is created, the coin block now almost within reach. When the player jumps a third time, the shake of the controller inches Mario a little bit higher and makes him bump into the block which produces a coin, but no platform is created. The clouds dissipate and Mario falls back to the ground and into the water, promptly letting the cloud suit disappear. The player picks up the cloud mushroom again and continues on with the level.

In the eyes of the player, he or she encountered a new suit type, which could be used to reach a high off coin block for extra points, and then continued on with the game. The whole experience lasts no longer than three seconds, and the player will have forgotten about it an equal three seconds later. However, this is by no means a trivial moment in the game, and later on the player will find that he or she suddenly know a great deal about Cloud Suit Mario without giving it much conscious thought. In a matter of three seconds, the game has taught the player seven novel things.

Figure 1.1 – shaking the controller a fourth time makes Mario touch the coinblock, but does not create a fourth cloud (Super Mario Galaxy 2)

Firstly, that picking up a cloud mushroom turns the player into Cloud Suit Mario. Secondly, that the player can shake the controller to create platforms. Thirdly, that these platforms can be used to stand on and additionally, fourthly, that these platforms can be used to reach previously unreachable higher locations. Fifthly, that the clouds disappear over time and sixthly, that touching water makes the cloud suit disappear altogether.

But perhaps the most interesting part of this short example experience is to note that the level designer—just as it is also no coincidence that the cloud mushroom is surrounded by water, *purposefully* placed the coin block precisely at that height so the player would try to reach it and in the process deplete the three clouds. This way, the game taught the player that the cloud suit grants you a total of three clouds only.

All games, whether intended for leisure or 'serious', revolve around learning. Good games constantly throw up hurdles or obstacles a player must learn to overcome, with the aid of ever new abilities altering the existing gameplay mechanics such as in the example above, while additionally hand-eye coordination, spatial awareness and timing need to be perfected to make it across that perilous gorge or defeat the nefarious villain. In fact, much of what makes a game enjoyable in the first place comes from the mastery a player needs to develop to overcome obstacles in a game and the corresponding boost in self-efficacy and feelings of competence (cf. Klimmt & Hartmann, 2006; Ryan, Rigby & Przybylski, 2006).

Furthermore, it used to be the case that games came packaged with extensive instruction manuals that needed to be read beforehand in order to make sense of the game, but nowadays much, if not all, of this instruction has been woven into the game design (Rauterberg, Gomez-Martin, Gomez-Martin & Gonzalez-Calero, 2004). The game teaches the player a new skill, and then immediately afterwards lets the player practice this skill until near mastery, creating cycles of expertise with information that is presented to them just-in-time instead of up front (Gee, 2005).

1.2 The case for serious games

Prensky calls motivation a condition sine qua non for successful learning, and it is precisely intrinsically motivating qualities that traditional learning often lacks (2007). Given the combination of engagement and learning inherent in games, it is therefore unsurprising that games used to teach people non-trivial instructional material, so called *serious games*, are garnering more and more popular attention. But in fact motivating students to learn instructional material (Malone, 1981) is but one of the plethora of possible merits of game-based learning. Games can be used to teach declarative, procedural and conceptual knowledge, train cognitive skills such as problem solving, enhance motor skills, change a person's attitude or intrinsic motivation and teach people to communicate and work together (cf. Prensky, 2003; Ratan & Ritterfeld, 2009; Wouters, Van der Spek & Van Oostendorp, 2009; Sitzmann, 2011).

In addition, games can be used to reach learning goals that have hitherto been difficult to realize with more traditional teaching methods. For instance, in a game such as *SimCity 4*, the player has to build and govern a large city. Small changes to e.g. taxes or the sewage system will lead to changes in other interdependent constituencies making up the city's success.

After a while the player will gain a compound understanding of the functioning of the city, something Gee (2005) calls system thinking. Furthermore, military games with high-fidelity virtual environments such as *Full Spectrum Warrior* can be used to train recruits in a realistic but safe and cost-efficient setting (Shaffer, Squire, Halverson & Gee, 2005). Last but not least, when novices take on the role of an expert in a game and are presented with an authentic task, they will acquire the epistemic frame of that expert, thereby gaining expertise as compared to just knowledge, which in turn can help them successfully deal with situations outside of the intended learning domain (Shaffer, 2006).

We will leave the topic of the alleged benefits of serious games here, even though this summation barely scratches the surface. Much has already been written on this topic (e.g. Prensky, 2007), but it is not what this dissertation is about. In fact, some of the alleged benefits of using serious games for learning may previously have been overstated.

1.3 On the efficacy of serious games

In trying to ascertain the benefits, or efficacy, of serious games, we should distinguish between the effectiveness of serious games, i.e. whether serious games can reach their learning goals, and the efficiency of serious games, by which we mean whether games can reach their learning goals within a certain relatively short time period of playing the game. Note that this definition of efficiency is different from efficiency expressed in issues of cost or the time needed to set up the instruction. That there is a lack of substantiating evidence for the efficiency of game-based learning is certainly the opinion of Clark, Yates, Early and Moulton (2010), who concluded that none of the journal-published review articles could provide a case for using serious games over traditional instruction.

One can of course wonder whether efficiency is that important; after all, a person that is motivated to play a game will be inclined to do so for much longer time periods than a person that doesn't like textbooks will be studying a textbook. After such a long time-period, a serious game may be just as effective in engendering learning as a more traditional form of instruction, a notion that is corroborated by a meta-analysis by Sitzmann (2011). However, this presupposes that serious games are more motivating than traditional instruction, and even this is hardly substantiated unequivocally by empirical evidence, as was discovered by Hays (2005) in an extensive review for the military.

It should be noted however that the reviews of Clark and colleagues (2010) and Hays (2005) only looked at scientific articles that were published until 2005 at the latest. In more recent times there have been a number of successful serious games that reached their learning goals and were published in peer-reviewed articles. For instance, regarding knowledge acquisition, Squire, Barnett, Grant and Higginbotham (2004) found that players of the game *Supercharged!*, where players have to navigate a spaceship through a 3D space by controlling the electric charge of a ship, performed better on knowledge tests than participants in a guided inquiry group. Wong and colleagues (2007) discovered that players of the game *Metalloman*, as well as participants in a hypertext group, outperformed a text-instruction group on knowledge acquisition in biology. Naturally, one could argue that a hypertext is more cost efficient in this case.

There are also examples of improved cognitive skill acquisition by playing games, compared with traditional instruction. In *QuestAtlantis*, a game-like educational virtual environment, students have to work in groups and act as a fictive investigative reporter for a newspaper. In an experiment done by Barab, Warren and Ingram-Goble (2006), it was shown that the game group improved their writing skills significantly more compared with a control group given traditional instruction. Likewise, writing skills of an experimental group playing the educational virtual environment *River City*, where players have to discover the reason why citizens of a town are falling sick, was also improved compared with a control group (Dede, Clark, Ketelhut, Nelson & Bowman, 2005).

For a more detailed analysis we refer to a review we performed in Wouters, Van der Spek and Van Oostendorp (2009). After a thorough search through a great number of articles, we only found 28 articles on serious games with empirical, quantitive data on the effectiveness of serious games that had some level of validity; of these, approximately 60 percent were successful in reaching their learning goals to varying degrees. Some of these were already mentioned above; in addition it is noteworthy that motor skills (see additionally Burke et al., 2009) and persuasive games for attitudinal change (see additionally Khaled, Barr, Biddle, Fischer & Noble, 2009) especially can be considered a promising avenue for serious games (Wouters et al. 2009). As an example of the latter, the game *Re-Mission*, aimed at adolescents and young adults with cancer, has been shown to improve medicine intake and self-efficacy compared with a control group (Kato, Cole, Bradlyn & Pollock, 2008). Our review thus indicates that serious games are a viable means of learning (Wouters et al., 2009). Serious games can be effective in engendering learning; however there is definite room for improvement, both in terms of effectiveness and efficiency.

1.4 Controlled experiments in serious game design

The validity of these and all claims pertaining to the effectiveness of serious games for learning is however confounded by very different and often flawed design methodologies. In the previous section we only named the experiments with a control group, which more often than not was lacking; sometimes even a clear idea of the learning goals was lacking. This is arguably to a large extent due to the embryonic state of serious games research. With so many different modes of learning that manifest itself in a game concurrently, many of these implicit or skill-based and some possibly even unique to game-based learning, accurately measuring what is learned in a game is not only a difficult task, but in part we also lack the right means to do so. In addition, creating effective serious games is a costly, time-consuming and highly-specialized endeavor. Many researchers will likely lack the know-how or funding to create modified or control versions of an existing serious game, whereas game developers will lack a natural or financial incentive to create different, less exciting versions of their games for testing. Whatever the reason may be, not enough research is done with empirical randomized controlled trials, a sentiment that is also echoed by reviewers Vogel, Vogel, Cannon-Bowers, Muse and Wright (2006), Mishra and Foster (2007) and, according to Sitzmann (2011), also gives rise to insinuations of publication bias.

Regardless of the methodological validity, some of the serious games are not effective in reaching their learning goals, or are less time efficient than traditional instruction. The reasons for this are unclear, another problem precipitated by the lack of empirical research, especially in the area of serious game design. The question how one should design a serious game to make it more effective therefore remains unanswered. According to Gee (2005), for a possible answer to this question we should take a good look at good entertainment games. We already argued previously that all games revolve around learning. Subsequently, as games needed to incorporate instruction into their game design from the beginning, and other game developers copy the game designs that work and subsequently improve on it where possible, after three decades of evolution in entertainment games the best games also incorporate the best instructional designs. It is a compelling argument, but without any scientific proof it remains a hypothesis at best. What’s more, entertainment games have to compromise learning for fun, which means that copying these techniques could even exacerbate the poor performance of serious games to date.

In addition, the situation where there is a lack of empirical evidence for the alleged benefits that a new educational technology should enable, is similar to the introduction of animations as a learning tool. Animations (or as they sometimes incorporate narration and sounds, also described more broadly as ‘multimedia’) too were expected by their nature to facilitate better learning of complex material, particularly if the information was procedural, because this type of information presentation would alleviate problems the learner may face having to mentally animate static images (Hegarty, 1992). Like games, however, there turned out to be scant evidence that substantiated this claim. In fact, in a review, Tversky, Morrison and Betrancourt (2002) found that dynamic visualizations did not generally outperform static images; Hegarty, Kriz and Cate (2003) concluded the same after comparing both animations and static diagrams in a mechanical context with procedural information, although both did perform better than a single image.

It was asserted that a large part of the inability of animations to engender learning at the same level as static images was due to instructional designs that were ill-fitted to the cognitive system of the learner (Moreno & Mayer, 1999) and an overall higher cognitive load imposed by animations

(Ayres & Paas, 2007). As a result, the research has since then focused on how to improve learning from animation, leading to a number of cognitive guidelines for instructional design (cf. Moreno & Mayer, 1999; Van Oostendorp, Beijersbergen & Solaimani, 2008). We contend that some of these guidelines may also potentially benefit the instructional design of serious games, and therefore warrant a closer look.

Consequently, we developed a serious game from scratch that gives us full control over the game design, in order to empirically and systematically test different game design principles on learning and engagement, and come to a set of guidelines for serious game design. It's a triage trainer called *Code Red Triage*, an acronym for COgnition-based DEsign Rules Enhancing Decision-making TRaining In A Game Environment. As is implied from the acronym, our research question can be stated as follows:

C*an we discover game design techniques based on cognitive principles that improve the efficacy of serious games, without harming the engaging qualities of the game?*

Chapter 2

Research outline

Abstract—Serious games have a great potential for training and educating people in novel and engaging ways. However, little empirical research has been done on the effectiveness of serious games and, although recent findings do point in a moderately positive direction, even less is known about why some games succeed in effectively educating while others don't. We therefore propose research in order to empirically test a number of cognition-based design guidelines for serious games that ameliorate mental model construction. Our purpose is to come to a set of design guidelines through empirical experiments that enhance the instructional design of serious games and can be used in the development of effective future games. This chapter outlines the rest of this dissertation, as well as introduces some of the key concepts that the research is based on.

Parts of this chapter are based on:

Van der Spek, E.D., Wouters, P. & Van Oostendorp, H. (2008). Efficient learning in serious games: a cognition-based design guidelines approach. In Michael Simonson (Ed.), *On the Horizon: Rays of Change, Annual Proceedings of the 31st Association for Educational Communications and Technology (AECT) Conference (pp. 382-390).* Orlando, FL: Association for Educational Communications and Technology.

2.1 Towards guidelines for serious games design

2.1.1 Learning from multimedia

In order to explicate the outline of the research and our choice for the experiments later on, we need to briefly explain what we mean when we talk about learning from a serious game. We will elaborate on this further, as well as discuss the associated metrics of learning, in section 4.1 and in the experiment chapters.

On the most basic level, learning, from a cognitive viewpoint, is obtaining new knowledge, organizing this into a meaningful form and integrating that with prior knowledge into a mental model. What is learned from a game is then the total body of knowledge a person has after playing the game, minus the total body of knowledge the person had before playing the game. From an experimental design standpoint, this necessitates a pretest-posttest design, as the different participants of an experiment may have different levels of prior knowledge pertaining to the instructional material.

When we look at learning from educational technology in more detail, we subscribe to the cognitive theory of multimedia learning by Mayer (1999; 2005; 2008), who states that whenever a learner perceives and tries to learn from instructional multimedia, he or she mentally selects the information that is relevant, organizes and then either creates a new schema in the working memory or integrates this information into a preexisting schema, and finally stores the new or augmented schema in long term memory. In addition, working memory only has a limited capacity (Baddeley, 2002) and is prone to being overloaded, which will lead to faulty learning and schema construction, whereas long term memory is virtually unlimited (Mayer & Moreno, 2003). Moreover, there is evidence that our working memory can process verbal and pictorial information via separate channels and therefore relatively concurrently (Mayer & Moreno, 1998). An adapted version of the model illustrating Mayer's cognitive theory of multimedia learning can be seen in Figure 2.1.

The model has been simplified slightly to highlight the three main assumptions of the model, that information is *selected*, *organized* and finally *integrated* (cf. additionally Mayer, 1999). If serious games fail to reach their learning goals, from a cognitive standpoint it could potentially be due to problems in all three of these constituent parts of multimedia learning. We will therefore use this subdivision as a basis to conduct our experiments; with experiments targeting each of the different cognitive processes individually. In order to use this subdivision as the basis for our experiments, we additionally operated with a slightly different interpretation of the organizing and integrating processes than Mayer's theory of multimedia learning (1999). According to his theory, the verbal and pictorial models are organized and then integrated into one coherent structure, while simultaneously being integrated with prior mental models. However, this ambiguous meaning of the term integration makes it difficult to separate the organization and integration parts empirically. In this dissertation we will therefore subsequently call the organization of words and images into one coherent cognitive structure without relying on prior knowledge the *organization* of information, whereas when this structure updates a prior knowledge structure in accordance with the newly organized information, we call it *integration* of information.

In theory, although this is arguably heavily dependent on the chosen implementation, the experiments targeting each of the three cognitive processes of multimedia learning could give us an indication which of these processes would benefit the most from our design interventions, as well as avenues for future research.

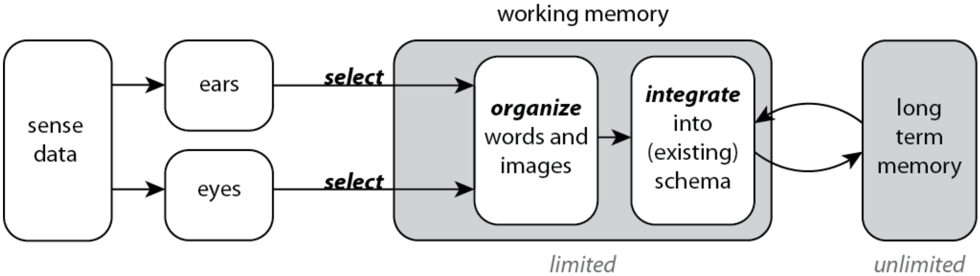


Figure 2.1 – Adapted model of Mayer's cognitive theory of multimedia learning (from Mayer & Moreno, 2003; Mayer, 2005)

Figure 2.2 outlines the framework for our research. Different game or instructional design techniques are hypothesized to target and subsequently improve the different constituent processes of mental model construction while playing a serious game. It should be stressed beforehand that while the different cognitive processes are distinct on a conceptual level, in practice when someone is e.g. 'organizing' the information, he or she will invariably select more relevant information and rely on bits of prior knowledge. Therefore, some of the game design techniques, such as signaling, can have an effect on multiple parts of this model.

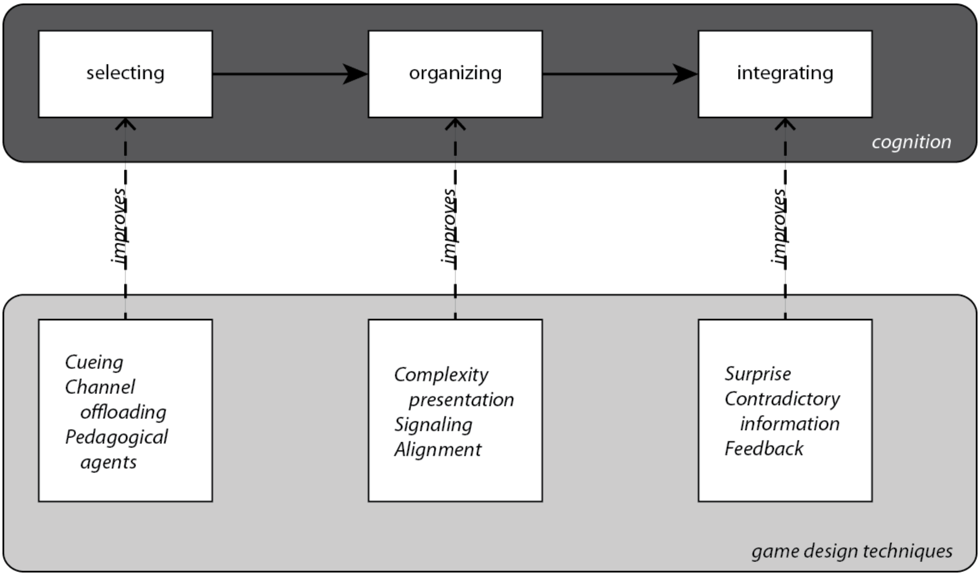


Figure 2.2 – Framework for cognition-based game design techniques hypothesized to enhance mental model construction

2.1.2 Selecting

With so much happening on-screen in a contemporary, perceptually rich virtual world, many students will not know where to look or attend to which sounds; especially as irrelevant information can be highly salient, while relevant information relatively inconspicuous. Reliably selecting relevant and discarding irrelevant information is particularly problematic for novices (Lowe, 1999; Cook, 2006) and as in a sense every person that first plays a serious game is still a novice to that game, having to learn not only the instructional material but also the manner in which the game conveys this, guidance in the selection part may be imperative to learn anything at all. In addition to problems with selection due to perceptual overload, problems can also occur when the working memory of a player is overloaded by other tasks and not enough cognitive resources can be directed towards the selection process (Brunken, Plass & Leutner, 2003).

A number of guidance techniques can be used in a game setting that could hypothetically mitigate these problems with selection due to perceptual or cognitive load. There is some evidence that *cueing*, i.e. presenting cues to guide a person's gaze or attention to the relevant material (De Koning, Tabbers, Rikers & Paas, 2010), thereby aiding in the selection of relevant from irrelevant information, may also be beneficial in the case of serious games. Steiner and Voruganti (2004) tested the effect of various forms of implicit and explicit guidance cues, such as signposts and rising smoke, on player propensity to visit a designated target in a virtual world, all of which were successful to various degrees. In research performed by Van Oostendorp, Beijersbergen and Solaimani (2008) on learning from animations, they found that interactivity and a focus cue alone did not lead to better learning, but when they were combined, did improve learning. As all serious games are interactive, adding a guidance cue may be a simple and effective way to improve the efficacy thereof.

Another technique, *channel offloading*, is based on dual coding theory (Clark & Paivio, 1991), which states that we process information from our visual and audible system via separate channels, and can do so relatively concurrently. If too much information is presented on-screen at the same time, for instance a text chat while trying to navigate a 3D world, one could provide the text audibly and thus offload information into the other channel, reducing the cognitive load on the visual channel. Mixed results of dual coding information in a multi-user virtual environment were found by Nelson and Erlandson (2008).

Last but not least, *pedagogical agents* are non-player characters that can direct a player's attention to the relevant information. They have already been shown to have some merit by serious games research, for instance by Conati and Zhao (2004), who used an agent to help students play a puzzle game to teach number factorization; by Moreno, Mayer, Spires and Lester (2001) who had Herman-the-Bug guide students through the Design-A-Plant game and thus teach them about ecology. Pedagogical agents can also be used to help players correctly organize information; as well as pose questions designed to stimulate retrieval of prior knowledge or make the player reflect on previous actions, thereby stimulating integration with prior knowledge. The latter has been shown to improve learning by Millis, Forsyth, Butler, Wallace, Graesser and Halpern (2011).

2.1.3 Organizing

In the organization phase of mental model construction, the learner orders the relevant material, chunks information and assigns causal links between different concepts. This results in basic pictorial and verbal models of the instructional material (Mayer, 1999). This process wholly takes place in working memory, and is therefore prone to cognitive overload. In addition, the ability of the learner to correctly organize the information is dependent on the way it is presented; in research from learning with animations it is for instance known that the organization of information improves when images and related explanatory narration are more closely synchronized, or when large and complex problems are segmented into bite-size pieces (Mayer & Moreno, 2003).

As serious games also impose cognitive processing apart from learning the instructional material, it is therefore important to determine beforehand how much complexity the player will have to cope with at a given time. One way to alleviate cognitive overload and improve the possibilities for information organization could be to present the information in the game just-in-time, i.e. introduce information only when the player needs to use it, (Kester, Kirschner, Van Merrienboer & Baumer 2001); this way, the player does not have to keep irrelevant information in working memory and will get a better idea of which information is related to which event.

Likewise, actions in the game should be designed to be in *alignment* with the learning goals. If a game demands a certain action of a player to reach an in-game objective, the player will believe this action to be important (Belanich, Sibley & Orvis, 2004); if this however is inconsequential to a preset learning goal, the player may be led to incorrectly assign importance to this objective, which could result in faulty learning.

Information organization can also be improved by highlighting the way in which different concepts are related, for instance underlining the text that states the way in which a specific effect is contingent on a certain cause; this is called *signaling*. Signaling is a form of cueing and therefore also helps in selecting relevant from irrelevant material (Mayer & Moreno, 2003), but in our framework we placed it under organizing to distinguish it from the cueing we mentioned previously, because structural characteristics are emphasized. In an article detailing an experiment that proves the merit of signaling, Mautone and Mayer (2001) also hypothesize that the reduced cognitive load due to signaling aiding in the selection and organization process, frees up working memory for more successful integration, however this applies to all of our design techniques.

2.1.4 Integrating

In the integration phase of learning from multimedia (in our approach, as described in 2.1.1), the newly organized information is integrated into an existing mental model via an update process, either by adding the newly organized information to the existing model, altering the existing model, or deleting from the existing model (Radvansky & Copeland, 2001). This updated mental model is then a mental representation of the newly derived knowledge, and can be stored in long-term memory. Integration therefore revolves around updating a mental model, and can be stimulated by game design techniques that necessitate the retrieval, evaluation and update of a mental model.

According to Kintsch (1998, p328), learning is contingent on whether 'we have some hooks in long-term memory to hang it on'. Meaning that for learning to occur, information is necessarily integrated with a mental model into long-term memory. One can use integration with the long-term memory to improve selection and organization in a game, for instance with the pretraining principle (Mayer & Moreno, 2003), where in a tutorial stage of the game an important concept can be presented to the player together with a retrieval cue; later on in the game, showing this cue will activate and retrieve the stored knowledge from the long-term memory, thereby increasing the capacity of the working memory in terms of the number of concepts that can be selected and organized.

However, this conflates the three different cognitive processes and makes it difficult to single out the integration process. For our research we were therefore primarily interested in an update that demands a comprehensive alteration of an existing mental model, so that an improvement in learning can be ascribed to improved integrative processing.

This update process can then for instance be forced by introducing *surprising events* in the game narrative, because surprising events, by their nature, are not a logical consequence that a player of a game expected to occur. The player has to reevaluate his or her preconceptions and determine if they have missed something (Kintsch, 1980; Hoeken & Van Vliet, 2000). This has been shown to improve comprehension of a story in research done by Hoeken and Van Vliet (2000)

This can also happen when *contradictory information* is encountered (Van Oostendorp, Otero & Campanario, 2002). Noticing the discrepancy can trigger deeper activation of prior knowledge and so enhance the updating process. Take for instance the game *Super Mario Bros.*, where enemies can be defeated by stomping on them. Later on in the game, Mario encounters an enemy with spikes on top of its shell. When Mario jumps on this enemy, he dies, which contradicts the player's previously constructed mental model that enemies can be disposed by jumping on them, forcing an update of the mental model. This can naturally also be done by other corrective *feedback* mechanisms, such as the player unexpectedly receiving penalties to the in-game score. Delayed feedback, where the feedback is not given directly but at a slightly later time, can potentially also improve building inferences in a serious game (Jarvis & de Freitas, 2009).

2.2 Proposed experiments

This was an explorative study as there was very little research on the cognitive correlates of game design decisions. At the same time, we wanted to come to a set of design guidelines that (serious) game developers can use for more effective (serious) game design. We used the adapted version of the cognitive theory of multimedia learning detailed above, i.e. that learning from multimedia incorporates selection, organization and integration of information, to ground our design guidelines in a theoretical basis. From this we set out to determine which of these cognitive processes can benefit from targeted design techniques, and tentatively, which can benefit the most. We should stress that the aim of this research is not to validate the framework itself; for a relatively recent overview of evidence corroborating Mayer's cognitive theory of multimedia learning confer Mayer (2008).

In choosing the design techniques that we were going to empirically test with our experiments, we additionally chose game design mechanisms that were preferably relatively easy to implement, in order to make them usable, and already have some basis in contemporary videogames, so that it would be clear how to implement them in other settings than our own. These were, based on the presented framework: cueing to improve selection, complexity presentation to improve organization and surprising events to improve integration. In addition, a follow-up to the complexity presentation with an adaptive game version was performed, to determine if serious games can be made more efficient.

2.3 Outline of the dissertation

In **chapter 1** we have seen that serious games offer players interactive learning environments that show great potential for letting the player reach a diverse set of learning goals in an engaging manner and in an authentic context. However, we have also seen that a large number of serious games have trouble reaching these learning goals. Until now it is largely unclear why some serious games are effective in engendering the desired mental model construction while others aren't.

We identified two prime causes for this current lack of knowledge. One, because it is still unknown what constitutes good serious game design, i.e. which game design techniques improve and which game design techniques harm mental model construction. The reason for this is that until now very little empirical research has been done on serious game design and its cognitive and affective correlates in the user. And two, the results of what little research has been done can often be questioned by weak design paradigms. This is due to a number of reasons; mainly that only few researchers have full control over the design of a serious game or that the researched game lacks clear, quantifiable learning goals.

With these two problems in mind we set out to create a serious game ourselves with a clear and quantifiable learning goal. With this game we could then systematically alter elements of the game design and measure their effects on the ability to reach the learning goals and engagement of the game, in order to come to a set of design guidelines for effective serious game design. The rationale behind the chosen experiments has been propounded here in **chapter 2**.

The game we made was a triage trainer called *Code Red Triage*. It is a 3D first-person game that trains the player in performing a categorization of the many victims of a mass casualty event according to urgency of needed medical attention. A description of the game, the instructional content, learning goal and justification for the chosen game type can be seen in **chapter 3**.

Following this, in **chapter 4** we explain what it is exactly when we say our game is effective, by relating shallow and deeper levels of learning and the associated metrics. Subsequently, we evaluate these metrics and the successfulness of our game in conveying the instructional material to ambulance personnel and laymen. In addition, in chapter 4 we describe an experiment that we performed where the game is compared with a more traditional PowerPoint presentation. The results of the experiment serve as a further confirmation of the problems serious games face in becoming as effective as traditional instruction.

Consequently, we perform the experiments as outlined in this chapter. First, we tried to improve the selection of relevant instructional material by adding subtle auditory and somewhat more explicit visual cues, and test whether these ameliorate learning and lead to more engagement due to less reliance on the help function in **chapter 5**. Reliance on this help function was much higher than previously anticipated, possibly muddling the results and those of future experiments. Chapter 5 therefore ends with a subsection that describes an alteration to Code Red Triage for the subsequent experiments.

In **chapter 6a** we report on an experiment designed to test different information presentation design paradigms on their ability to improve information organization. Here, we contrasted different ways of complexity presentation; progressive versus high variability in the complexity of consecutive victims in the game, and just-in-time presentation of the player's options to resolve these problems versus all the options from the start. The results of this experiment provided us with a confirmation as well as additional questions that justified a follow up experiment, which was performed and can subsequently be found in **chapter 6b**. In this experiment, we tested whether a modification of our game that automatically adapts to the performance of the player can improve the time efficiency for able players.

Chapter 7 details an experiment where we introduced surprising events in our game narrative at a number of decisive moments. We hypothesized that this would force an update of a mental model that was previously committed to long term memory, and therefore result in a better integration of new knowledge into already existing schemata.

Finally, **chapter 8** reflects back on all the experiments, how they relate to the theory, and provides a general conclusion to the dissertation, as well as discussing limitations and recommendations for both game developers and future researchers. Added as **appendix A** is the handout with the test that was used in the experiment of chapter 7; it is highly similar to the handout for the experiments in chapter 6a and 6b as well. The dissertation concludes with the **summaries** in English and Dutch respectively.

Chapter 3

Code Red Triage

Abstract—We created a game called Code Red Triage in order to facilitate our proposed experiments in serious game design. The game, which is a total conversion mod of Half-Life 2 is a 3D first person game where the player acts as a medical first responder and has to learn how to perform the triage procedure on the victims of a mass casualty event in a subway station. This game was purposefully made to be cognitively demanding and engendering rich affective responses, resembling high fidelity entertainment games, while simultaneously having clear learning goals that can additionally be measured via an in-game score.

Parts of this chapter are based on:

Van der Spek, E.D., Wouters, P. & Van Oostendorp, H. (2010). Code Red: Triage. Or, COgnition-Based DEsign Rules Enhancing Decisionmaking TRaining in a Game Environment. In *The British Journal of Educational Technology, 42*(3). 441-455.

3.1 Game type

Any research project that uses a single game for extensive empirical testing will invariably run into problems with generalizability. Chapter 1 already detailed a large number of different learning goals of serious games, and on top of that there are many different game types, each with their own idiosyncrasies, affordances and problems. Even within a given game type there could be many different ways of introducing gameplay mechanics, let alone narrative structures and genres with far reaching implications for learning and engagement. This has led to a Catch-22, as experiments on game design haven't been done because they can't be generalized, and generalizations on game design can't be made because there is no experimental data. We therefore strove to create a game that revolves around the training of something that incorporates a number of clearly defined, meaningful learning goals, which should have unambiguous and quantifiable right and wrong answers.

In addition, it was deemed important that the game should be comparable in fidelity of graphical quality with contemporary entertainment games, and be a 3D game where the player has to navigate a virtual environment. The reason for this choice of game type is that it introduces a new set of affordances and problems that can be affected by the type of cognitive design guidelines that were implemented in the subsequent experiments. Firstly, it induces a feeling of presence in the player, which is an important aspect for games that are intended for training real-life situations; secondly, it demands that the player is aware of and spatially navigates a virtual environment. Both place higher demands on the cognitive system of a player, potentially overloading it. This is further exacerbated by the high fidelity of contemporary graphic engines for games, where more graphical detail means more difficulties with directing attention, but is necessary to create a suspension of disbelief and thereby presence in the player (IJsselsteijn, 2003).

Three years before the start of this project, McGrath and Hill (2004) published a paper on a game they made called *UnrealTriage*, a game designed to train the triage procedure, or the categorization of the many victims of a mass casualty event according to urgency of needed medical attention, created with the *UnrealEngine 2* level editor. Training a medical first responder in the triage procedure seemed like a natural fit for a 3D serious game, as a virtual environment where you interact with victims and get confronted with dangerous and bloody situations, almost by its nature exerts rich affective responses such as stress and compassion, while simultaneously being relevant to the context of learning the triage. Because acting out the role of a medical first responder is relevant to the context, it makes the task authentic to the player, which in turn is both important for, and perhaps even necessitates suspension of disbelief or immersion in the world (Herrington, Oliver & Reeves, 2003).

The triage procedure is furthermore a relatively simple procedure that can be taught in a relatively short time period by interacting with individual victim cases, thus making it possible to constantly monitor the player's progress with every new victim.

3.2 A game for triage training

A triage, categorizing a victim according to the urgency of needed medical attention, is a relatively easy task that is set up in such a way that even laymen can quickly learn how to do it. In the Netherlands, and most other Western European countries, it roughly consists of the following consecutive steps. First, check the mobility (i.e. can the victim walk); secondly, check whether the victim's airway is free; thirdly if the respiratory rate is within a certain range; and lastly, check the capillary refill time (i.e. press on the nail bed and see when the nail regains color, as an indication of blood circulation). There are a number of factors that can alter the procedure. If the airway is obstructed, the player has to free the airway either by a chin lift or a jaw thrust. If this succeeds, the player has to place the victim in the stable recovery position. If it is too dark, a capillary refill check is unsuitable for measuring blood circulation, and the pulse should be taken instead. After obtaining the results from these checks, or performing the corresponding actions, the medic has to categorize the victims accordingly into either of four categories: T1 (red) for urgent, T2 (yellow) for delayed, T3 (green) for light or no injury and Dead (black) for death immanent or already deceased (cf. Hodgetts, 2003). The flowchart depicting the procedure can be seen in figure 3.2. Additional details pertaining to the procedure will be discussed where needed in the individual experiment chapters.

The main learning goal of the game then was to learn this procedure, which was done by triaging a number of victims of a virtual mass casualty event. Through consecutive victim cases, the player has to read hints on the procedure, form and try out hypotheses, and thus iteratively create and update a mental model of the procedure.

The design of the game started with a design of the instructional material, for which we consulted a number of subject matter experts. First, the scenario and victim cases were devised together with Frank Rosier of Rosierenco, who has been a board member of the Major Incident Medical Management and Support (MIMMS) group and trains emergency medical personnel in the triage procedure. These victim cases were further validated with the Dutch handbook for medical management at large scale incidents, or *Geneeskundig management bij grootschalige incidenten* (Hustinx, Meeuwis & Hermans, 2004). Subsequently, the scenario was presented to Mike Bemelmans, a traumatologist at the Utrecht University hospital (UMC) and Cyrille Veltman, a medical instructor at the Ministry of Defense. These experts evaluated the victim injuries and corresponding triage categories we had devised, proposing a few extra changes to make a number of cases more challenging, as well as giving advice on the type of injuries and dispersion of bodies associated with a terrorist type bomb for added realism.

The game was then modeled with the *Source SDK*, a mod tool that comes free with Valve Corporation's *Half-Life 2* games. The player arrives in the central hall of the train station, after which he or she will need to find the way to the subway line that was struck by the explosion. We deliberately made the tunnel to the subway station lengthy and labyrinthine—a straight walk takes two minutes, the average about four; as we hypothesized that this would give an objective measure of the player's proficiency with video games. We hypothesized that experienced (First Person Shooter) gamers would breeze through the winding corridors, whereas novices may have considerably more difficulty with all the consecutive turning and maneuvering around objects. A measure for gamer experience may be necessary as some game design features, such as visual cues, could be more rapidly picked up by gamers, who have more experience with game specific guidance cues. This could be an important covariate for the effects of the cognitive games design guidelines. Screenshots of the game world, as well as overview maps, can be seen in figures 3.2 through 3.7.

Arriving at the subway platform, the player will encounter a total of 19 victim cases scattered about with varying degrees of injuries and should subsequently proceed with the triage task. A screenshot of the *triage procedure interface menu* in our game can be seen in figure 3.8, with the different checks on either side of the screen and the triage categories at the bottom. There are eight check buttons in total, the four checks mentioned above, a 'general info' button and three buttons which vary and are primarily intended to distract the player and cost valuable time. After a classification is awarded, the victim is colored accordingly.

There were some background noises such as from trains screeching on the tracks, a rumbling sound and at the subway platform one could hear people screaming and wailing in agony. In addition, we added arousing music to heighten the tension in the player, this was however turned down at the moment a player opened the triage menu, as not to interfere with learning (Mayer & Moreno, 2003).

Apart from the different versions that had to be made in order to do the experiments, there was one minor alteration in the presented information, and one major shift in the purpose of the game Code Red Triage. The first is the order in which the victims are encountered. Initially, there was some variability in the victim complexity, where easy victims were alternated with more complex victims, but the presentation of victims in terms of complexity was made more rigorous after the experiment in chapter 6.

The second was a considerably larger shift in the purpose of the game. Initially, the participants were allowed to study a flowchart of the procedure briefly before playing the game, and could also consult this on demand in the game; the focus of the game was therefore on training or practicing the triage skill. However, as the guidelines in this research were intended to target learning and mental model construction, we took out the triage procedure flowchart, thus turning the focus of the game more squarely towards learning itself rather than practicing the triage. This shift, along with the motivation, will be explained in greater detail at the end of chapter 5, however it should be noted that there is an anachronism in this dissertation. The game as a training device with some variability in the complexity presentation was used in the pilot (section 4.2) and the cueing experiment (chapter 5), but the comparison with the more traditional PowerPoint instruction (section 4.3) and the other experiments (chapters 6a, 6b and 7) used the newer version of the game.

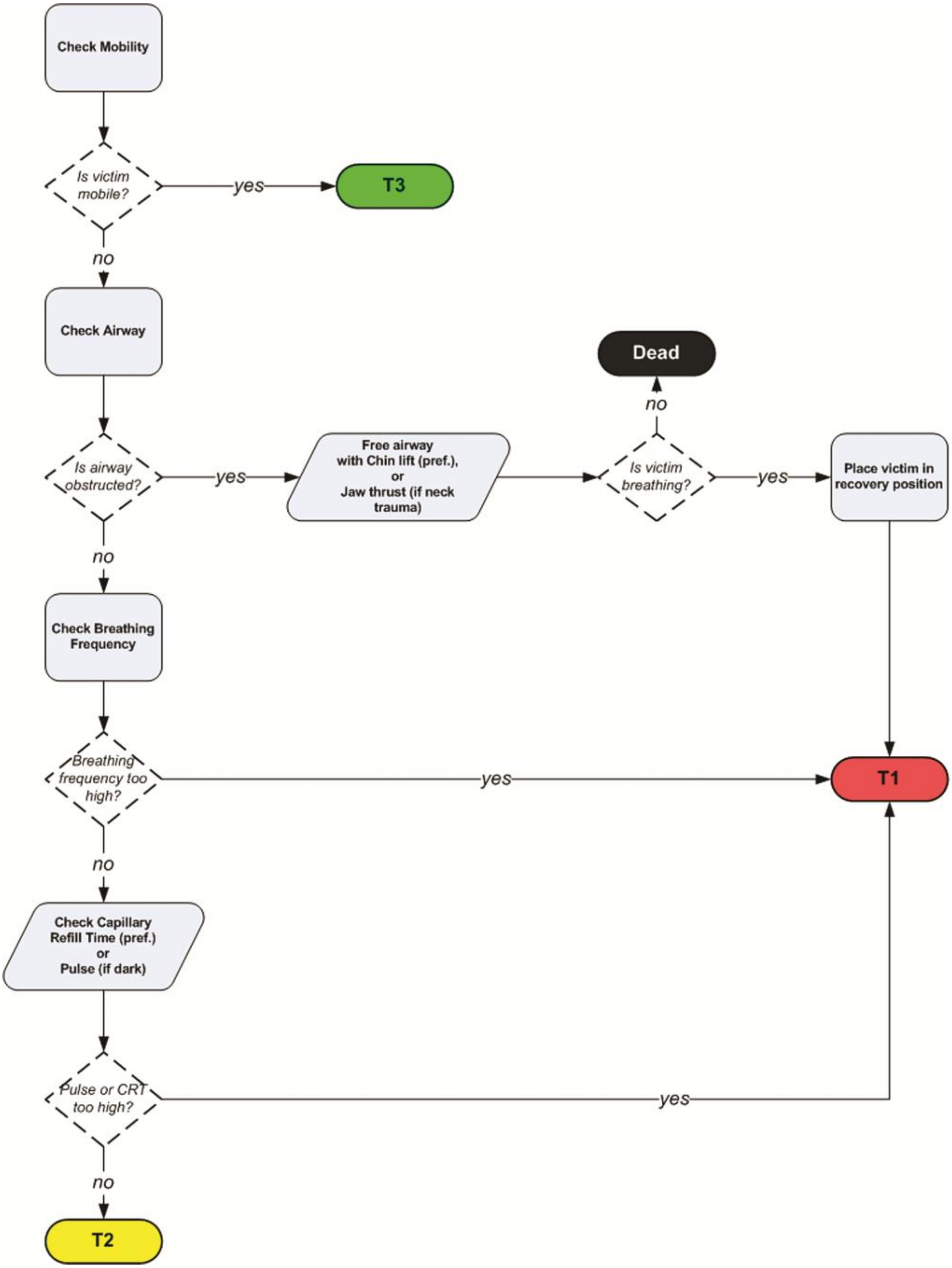


Figure 3.1 – Triage procedure

Figure 3.2 – Train station, starting point of the game

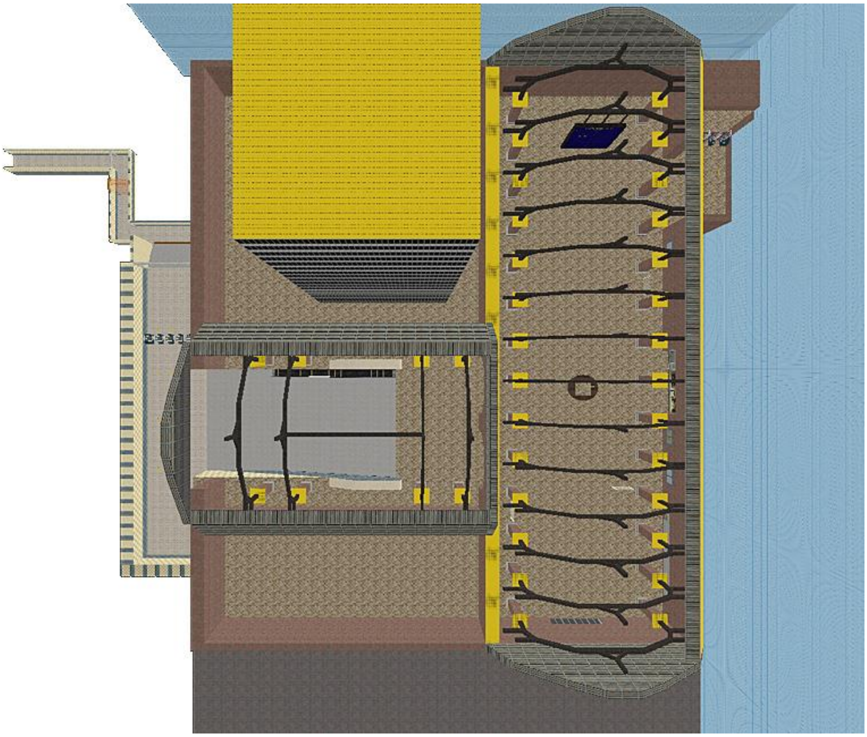


Figure 3.3 – Map of the train station

Figure 3.4 – Subway tunnels leading to the platform

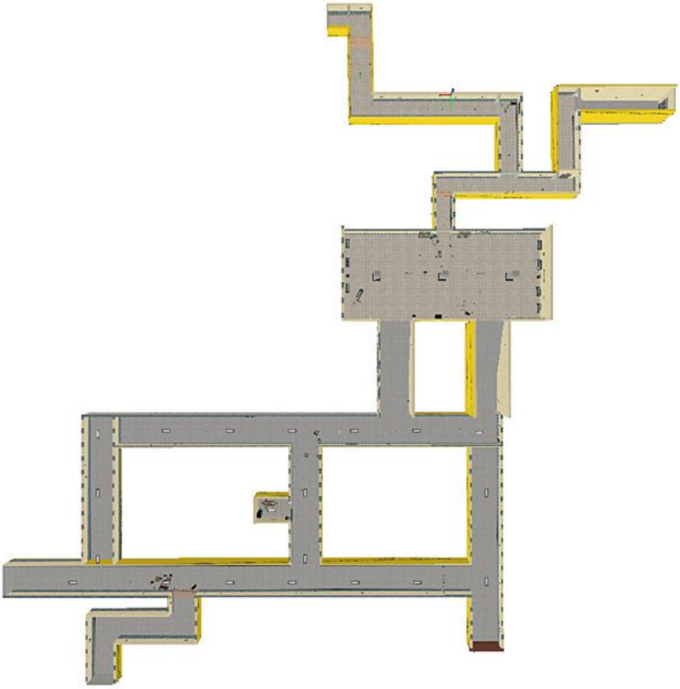


Figure 3.5 – Map of subway tunnels

Figure 3.6 – Subway platform

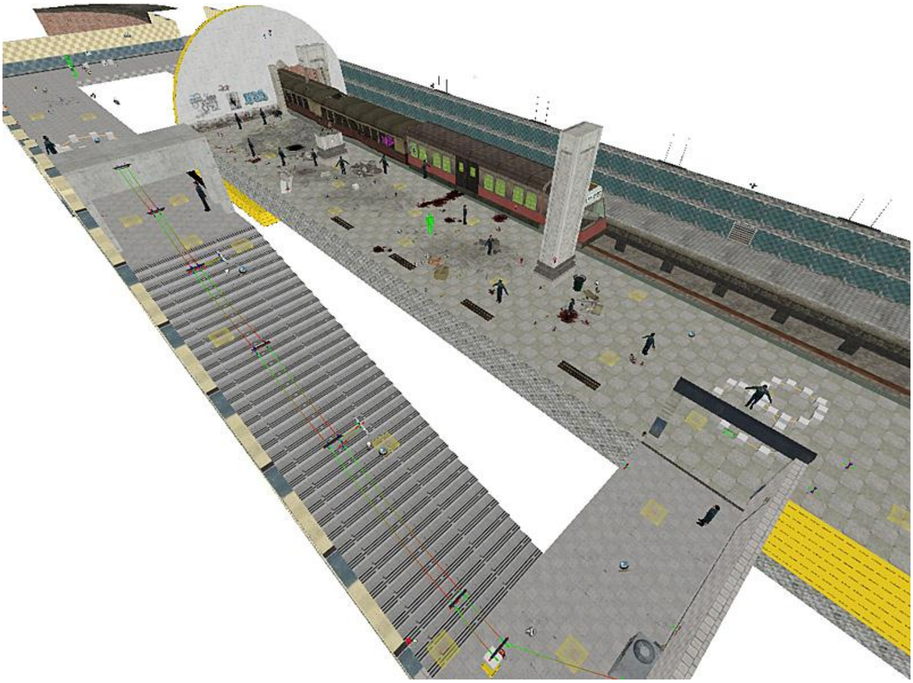


Figure 3.7 – Map of subway platform

Figure 3.8 – Triage menu in-game (first experiment version, in Dutch)

Figure 3.9 – Feedback after triaging (later experiment versions)

3.3 Scoring and logging

During the triage task the game keeps a score of how well the player is performing, and communicates this to the player. For every correct triage the player will get 100 points, which is shown as a progress bar at the top, to a maximum of 1900 points for all the victims. A penalty is subtracted from the score when the player takes longer than 30 seconds. The feedback is given immediately after a victim is triaged, as it has been shown that this is most effective for learning in a similar procedural triage setting (Jarvis & De Freitas, 2009). In later versions of the game, this score was further fine-tuned by having variable times per victim according to the number of procedure steps that should be taken, as well as awarding penalties for forgotten steps, steps that were taken out of order, or unnecessary steps taken. This was then communicated to the player via a short text message (see figure 3.9).

At the same time, we implemented an extensive logging system in the Half Life 2 source code, automatically writing all the steps that were taken by the player to a text file, how much time they spent triaging a victim and the corresponding score obtained afterwards. Trying to assess how well people learned from a game post-hoc could therefore be considered redundant, as games *are* a continuous assessment of how much they are learning. However, the triage score may not be an exact or conclusive measure for learning. We will elaborate on this in chapter 4.

We contend that games should be goal-directed, a competitive activity and conducted within a framework of agreed rules, while constantly providing feedback to enable players to monitor their progress towards the goal (Wouters et al., 2009). The goal of our game is to teach the player to perform triages, the performance of which is communicated by giving feedback in the form of a score and short verbal explanations, which in turn adds a competitive factor; we therefore believe our training is fully a game, instead of a simulation.

Chapter 4

Assessment of Code Red Triage and serious games in general

Abstract—In order to assess learning from a serious game, we propose a number of different techniques; paper knowledge tests, a structural knowledge assessment and the in-game score, in order to measure surface and deeper levels of knowledge. To measure affective responses, we used the engagement subscale of a presence questionnaire. These measures were subsequently used and validated in a pilot experiment, and an experiment where we contrast our game to a PowerPoint presentation of the same instructional material. In the latter we discovered that Code Red Triage is less efficacious than the PowerPoint presentation, and that this is likely due to a higher cognitive load. The difference in efficacy disappeared after a week

Parts of this chapter are based on:

Van Dijk, V., Van der Spek, E.D., Van Oostendorp, H. & Erkens, G. Serious games versus tradition instruction; learning, enjoyment and cognitive load. *Manuscript submitted for publication*.

4.1 Determinants of a successful serious game

Underlying the assumption that serious games are struggling to reach the same levels of efficacy as traditional instruction in teaching the intended learning goals, as was argued in chapter 1, is the notion that the learning goals are similar and clearly measurable. We already mentioned in section 1.2 'the case for serious games' that in fact you can teach a lot more with serious games than with traditional instruction; notably, information is not only learned in a game, but can also immediately be applied, thus creating the possibility for skills development (Dondlinger, 2007). While application of the triage procedure is arguably a skill, it is one that is multi-faceted, including for instance procedural knowledge but also coping with stress, and is thus not easy to measure accurately. Perhaps more importantly, the skill is quite contextualized, and results may therefore be less well generalizable to other serious games than is the case with our framework of selecting, organizing and integrating information (section 2.1), which are cognitive processes that occur in any game.

As a result, we chose to focus solely on the cognitive aspects (i.e. testing knowledge acquisition and construction) of the triage skill, and the result of our game design interventions thereon. In addition, these effects of game design interventions can be measured quantitatively, so that we can get a clear picture not only if, but also how well the guidelines improve learning. First this would demand an elaboration on what we mean by learning from serious games.

4.1.1 On learning

Mayer (2005) distinguishes between two types of instructional paradigms: those that promote information acquisition and those that promote knowledge construction. In the first, a learner's mind is an empty vessel waiting for the instructor to pour in new information; in the second the learner actively tries to make sense of new information and has to integrate new information into preexisting knowledge (Mayer, 2005; Moreno & Mayer, 2007), as was explained previously. A clear cut distinction between the two may be impossible, as one could argue that in order for a person to make sense of new information, this information at least has to be acquired and what subsequently happens in the learner's mind is contingent on a person's epistemology, where in fact in a radical constructivist view (cf. Jonassen, 1991) all knowledge is constructed. We'll leave the philosophical debate here, but suffice to say that, while we do not adhere to a constructivist epistemology per se, in serious games at least the learner has an active role and has to experientially make sense of bits of information by inference and mental model construction (cf. Graesser, Singer & Trabasso, 1994).

While the terms information acquisition and knowledge construction may have an unwanted epistemological connotation, they do allude to an important difference in modes of learning, and related learning goals. From here on, we will denote this as the difference between *surface learning* and *deep learning* (Graesser, Chipman, Leeming & Biedenbach, 2009). It should be stressed beforehand that this distinction is not decisive, as both are correlated and could be interdependent (cf. Kintsch, 1994). In the case of triage training, the distinction may be especially difficult to delineate, which we will argue next.

4.1.1.1 Surface learning

Surface learning roughly corresponds to accumulating facts, or declarative knowledge; it is useful when the information has to be reproduced afterwards and is often associated with rote memorization (Entwistle & Tait, 1990; Kintsch, 1994). As could be seen in section 2.2, a triage is a relatively straightforward, simple procedure where every step is of the form *if* x *then* y (*else* z), and every branch in the corresponding decision tree terminates in a certain triage category. It furthermore contains such declarative knowledge as knowing which color belongs to which triage category, and which steps are a part of the triage and which aren't. This means that for the triage procedure, being able to reproduce declarative knowledge from memory may already be some indication of whether a person who has played Code Red Triage has learned how to perform a triage.

Surface learning, learning of declarative knowledge especially, can be assessed well via pen-and-paper multiple choice knowledge tests (Shavelson, Ruiz-Primo & Wiley, 2005). Therefore we used paper knowledge tests with multiple choice questions in all of our experiments as one of the main pillars for measuring learning gains. In addition to using strictly verbal questions, we included a number of questions with a screenshot from a victim in the game, or another virtual scene and character created with the same game engine, where the participant was given a snapshot from the triage procedure and had to choose one of four possible next steps in the triage procedure. This was because in research by Belanich, Sibley and Orvis (2004) it was discovered that the players of a game recalled visual information better than written information. If videogames favor the acquisition of visually encoded knowledge, screenshots of the game may help the test participant to retrieve the correct answer and therefore provide a better assessment of what is learned in the game.

Learning of procedural knowledge is more difficult to categorize. The triage is a procedure, and therefore knowledge of the triage is procedural knowledge; however procedural knowledge is sometimes seen as distinct from declarative knowledge, in that it is compiled or tacit knowledge (Driscoll, 2005). Conversely, every step in the triage procedure is explicit; it is a conditional statement with a clear resolution. Even categorizing the victim in a triage category has the form of an if-then statement, albeit preceded by a number of similar decisions before it. We therefore contend that in the case of the triage procedure, procedural knowledge, like declarative knowledge, can to some extent be 'surface knowledge'. One can still arrive at the declarative knowledge by deep learning, but assessing procedural knowledge in our case does not necessarily imply deep learning has occurred. Therefore, we need another measurement device.

4.1.1.2 Deep learning

If surface learning corresponds to learning *that*, then deep learning is learning *why*. Deep learning, as opposed to learning in order to replicate information, signifies that the learner is actively trying to make sense of, i.e. comprehend the instructional material (Kintsch, 1994). Comprehension involves the learner constructing a mental model inside his or her mind that is a representation of reality and can be used for rapid inferential reasoning (Johnson-Laird, 1983).

One of the ways to assess the quality of a constructed mental model is by performing a structural knowledge assessment (Ferstl & Kintsch, 1999). Numerous forms of structural knowledge assessment have been proposed over the years, such as hierarchical cluster analysis (Adelson, 1981) and multi-dimensional scaling (Gonzalvo, Canas & Bajo, 1994). For a more detailed breakdown of different techniques that were used, see Ferstl and Kintsch (1999) and Wouters, Van der Spek and Van Oostendorp (2011). We chose Pathfinder, which is a computer program that creates Pathfinder networks (PFNets) or (un)directed graphs from relationship ratings of concept pairs (e.g. 'water' is highly related to 'ice', but less related to 'tree'), where concepts are nodes in the graph, and nodes that are closely related appear closer together in the graph than those that are less related (Schvaneveldt, 1990). These ratings are relatively easy to perform for participants (cf. Trumpower, Sharara & Goldsmith, 2010) and Pathfinder is better at extracting a latent structure as opposed to other techniques such as multidimensional scaling (cf. Jonassen, Beissner & Yacci, 1993). The graphs that were constructed for our research were minimum-cost networks, i.e. graphs with the least amount of links, with the Minkowski distance metric set to infinity and the scope of the minimum cost paths set to n-1 (cf. Chen, 1998).

A PFNet can be compared with another PFNet via a similarity measure. One way to compare two PFNets for similarity is by dividing the number of shared links by the total number of unique links in both networks, or the intersection divided by the union of the two sets of links, which results in a similarity score between 0 and 1, where ratings closer to 1 indicate higher similarity (Trumpower et al., 2010). In addition, the probability that two random networks share a number of links can be determined with the hypergeometric probability distribution; by subsequently subtracting the similarity that is expected based on chance from the similarity that is observed via the previous metric, one derives at a similarity score corrected for chance (Gomez, Hadfield & Housner, 1996). This score is negative when the observed similarity is less than what would be expected from chance and positive when the networks are more similar than would be expected from chance (Thompson, Gomez & Schvaneveldt, 2000). The chance-corrected similarity score can be used for variance analysis (Gomez et al., 1996) and was the one we used in our experiments, because it is more sensitive than the raw similarity measure (Gillan & Cooke, 2001).

Pathfinder has been successfully used in other experiments to measure knowledge on a structural level, by calculating the similarity of a participant's PFNet to that of a referent expert PFNet, for instance by Goldsmith, Johnson and Acton (1991) and Kraiger, Salas and Cannon-Bowers (1995). Day, Arthur Jr. and Gettman (1994) used the change in Pathfinder similarity score to measure learning from a complex videogame, and found that it was even predictive of skill retention and skill transfer.

We already mentioned the difference between declarative and procedural knowledge; a third type of knowledge is strategic knowledge, which is the combination of declarative and procedural knowledge in a mental model (Kahler, 2002). By combining declarative and procedural knowledge in a mental model the student not only knows what and how to do something, but also why and when to do this, resulting in a deeper level of comprehension. Additionally, whereas declarative and procedural knowledge are very domain-specific, strategic knowledge is more generally applicable (De Jong & Ferguson-Hessler, 1996); there is evidence that strategic knowledge may be important for the transfer of the knowledge to other settings (Wenden, 1998).

After obtaining the strategic knowledge structure a player has created while playing Code Red Triage by the structural knowledge assessment outlined above, this structure can be compared to that of an expert via a similarity score. This score is then an objective measure of the quality of deep learning that has occurred during gameplay and possibly an indicator for skill transfer.

4.1.2 On engagement and presence

One of the main reasons to use serious games as instructional material is that they're more fun than traditional instruction, even though evidence for the engaging qualities of serious games is mostly anecdotal (Wouters et al., 2009). While we already argued in the introduction that learning, to an extent, is what makes games fun to play, there is a real possibility that design guidelines used to improve learning could harm the enjoyment of the game. Discovering new things by oneself may be an important reason why games are fun (e.g. Garzotto, 2007), as is the challenge inherent in overcoming difficult obstacles (cf. Garris, Ahlers & Driskell, 2002). Techniques that are designed to guide the player in learning the relevant material may come off as hand-holding, both in terms of discovering new things and in problem-solving, and thereby lead to less engagement. In fact, the feeling of autonomy during gameplay is one of the strongest motivational pulls for games in a study done by Przybylski, Rigby and Ryan (2010). Conversely, in the flow model by Csikszentmihalyi and Csikszentmihalyi (1988), when challenge exceeds the abilities of the user, the user (in our case: player) will feel anxious and leave the optimal, or 'flow'-experience. The feeling of being in a flow is both pleasurable and can improve learning (Kiili, 2005), which would justify hand-holding in difficult areas of the game.

Another important aspect of the experience of games that could be affected by the serious game design guidelines we're investigating, is the feeling of presence in the virtual environment, or in other words the computer-mediated sense of being in the environment (cf. Witmer & Singer, 1998). Inducing presence can be an important goal of the game designer, as it is positively correlated with the height of experienced emotions (Riva et al., 2007) and may therefore be a prerequisite for learning goals that require the player to experience being a part of the virtual world, for instance when it is needed to conjure up feelings of fear or train situational awareness (cf. Herrington et al., 2003). It is furthermore induced the strongest by games with a first-person view (Kallinen, Salminen, Ravaja, Kedzior & Sääksjärvi, 2007) and high-quality graphics (Nunez & Blake, 2003). As Code Red Triage has both, and training crisis situations requires feelings of stress and anxiety, our game would be a relevant testing environment. If the guidance that we incorporate in the game turns out to be too ostentatious, for instance a large blinking sign cueing the relevant material, this could harm the feelings of presence, which in turn could interfere with the goals of a game designer.

Although it is not exactly clear whether factors such as interest and engagement are determinants or correlates of presence, they do seem closely related (Lessiter, Freeman, Keogh & Davidoff, 2001). As we already have two quite extensive knowledge metrics in our experiments, we chose the Engagement subquestionnaire of the Independent Television Committee's Sense of Presence Inventory (Lessiter et al., 2001) to get an amalgamated measure of both. In this research by Lessiter and colleagues (2001), the subquestionnaire was shown to be reliable with a Cronbach alpha coefficient of 0.89. The questionnaire will be elaborated on further in the independent experiment chapters.

4.1.3 On games as assessment

Finding ways to assess learning and enjoyment in games is frankly a peculiar undertaking. Games are assessment: they continuously challenge the player with problems that need to be overcome by applying the knowledge or skills that were taught in the game (Gee & Shaffer, 2010). Serious games like Code Red Triage can therefore be considered an assessment of the skills and knowledge of a player, at least when the learning goals are aligned with the goals of the game. If an obstacle in the game is too difficult to overcome with the prior knowledge of a player then progressing past the obstacle, e.g. by training the needed motor skills or solving the puzzle, is testament to the newfound knowledge or skills of the player. Finishing the game may be proof enough that the game is successful in teaching the educational material.

In practice however, for many purposes it may prove impossible to adequately align the instructional content with the game rules or obstacles within the game, at least without losing the engaging flow of the game. Nonetheless, as games are digital media it is relatively easy to assign a direct numerical score to actions pertaining to learning goals or the correct carrying out of instructional material. If this score is then fed back to the player, as the main score of the game, via bonus points or Xbox-like Achievements, this will also increase the competitiveness of the game and thereby the engaging characteristics thereof. A high score would thereby imply that the game was successful in teaching the relevant material.

Some caveats to this paradigm should be noted. Firstly, chances of trial-and-error gameplay reaching to high scores should be minimized as much as possible. Secondly, overcoming obstacles in the game only implies newfound knowledge if the player did not have this knowledge in the first place, consequently the experimenter still has to test for prior knowledge of the participant. Thirdly, while game scores, and the progression thereof, can be used to compare the results between the subjects of an experiment, scoring is very game specific, and therefore difficult to use as a comparison with other games or instructional media.

Most importantly however, (partially) failing a task and receiving a lower score does not eo ipso mean a failed understanding of the task afterwards. In fact, as we explained in section 2.1.4, doing something wrong and being told it was wrong can lead to reflection on the task, inferences on what the correct procedure should have been and thereby a deeper understanding of the task than when the player did it right on the first go.

4.2 Pilot experiment and target audience

In order to examine the usability of our game and the measurement scales particularly, we've performed a small-scale pilot with an early version of the game, which did not yet include any of the guidelines or sounds and music. The purpose of the pilot was to determine whether the triage procedure was easy enough to teach to laymen, to ascertain the potential validity and usefulness of the measurement scales, and to correct possible mistakes in the training and the game. We were furthermore interested in how professionals would fare, as these could also provide useful insights on the training itself, and therefore first tested the game with EMS personnel.

A total of 10 emergency physicians participated in our pilot experiment. They were all male subject matter experts (mean years on an ambulance 10.4, SD = 9.0) and ranged in age from 21 until 55 (M = 41.5, SD = 9.13). The pilot had a pre-posttest design with a number of different instruments to measure knowledge acquisition and construction in the videogame. First, the participant had to rate a total of 78 word pairs for the structural knowledge assessment. These words were a mix of procedural steps, such as 'classify victim' and 'airway', medical concepts, such as 'blood circulation' and situational concepts, such as 'darkness', in order to ascertain the strategic knowledge of the participant (see 4.1.1.2). After this they received a knowledge test with 10 multiple choice questions on triage concepts. The participants then played the game for approximately 18 minutes, right after which they filled in the engagement questionnaire of the ITC-Sense of Presence Inventory (cf. Lessiter et al., 2001), followed by the same tests as before the game: the mental model elicitation and knowledge test respectively. In addition to this, we gave the participants two short questionnaires with procedural questions, one strictly verbal, the other with the aid of screenshots.

Ten physicians may be too little a number to come to any valid conclusions on learning gains, but some tentative trends can be discerned. Qualitatively, nearly all of the participants responded positively to the game (ITC-SOPI engagement questionnaire mean of 3.6/5), saying they liked how it looked, that it engaged them and that they found it particularly useful, as some admitted to not knowing the exact procedure anymore. Evidence for this can be found in the results of the pre and post knowledge tests, where the participants scored significantly better after playing the game than before (N = 10, pretest M = 7.70, SD = 1.16, posttest M = 8.30, SD = 1.16, $t(9) = -2.71$, $p < 0.05$).

The same however cannot be said for the knowledge structures elicited by Pathfinder. While the participants reported no real problems in rating the concepts, most of them finishing the 78 pairings in under ten minutes, the corrected similarity measures for the pre and post mental models do not differ significantly, and also do not correlate significantly to the scores obtained by the paper test. It is possible that the use of Pathfinder in this context is inadequate, but a number of alternative explanations for this unexpected result can also be given. Firstly, while a significant increase in test scores was found, this increase was only very small; ambulance personnel already know how to perform a triage quite well, so it is improbable that a real change in mental model structure was to be expected, especially with so few participants. Secondly, the mental models of ambulance personnel, being medical experts, may also have too much encapsulated knowledge (cf. Boshuizen & Schmidt, 1992), and are therefore more akin to those of the experts. Because of this it may be unlikely that playing the game will lead to a convergence with the theoretical mental model. However, comparing the mental models of the ambulance personnel to the expert reference model also did not yield significant results.

We then performed the same pilot with nine medical novices, mean age = 31.11, SD = 14.21. These also showed a marked improvement in knowledge test scores after the game (M = 6.89, SD = 1.76) in comparison with before (M = 4.00, SD = 1.23), $t(8) = -3.51$, $p < 0.01$. A paired-samples t-test here also did not show a significant increase in the structural knowledge assessment scores [$t(7) = -1.78$, $p = 0.12$], but with one participant's data unusable, the valid N = 8 may be too small to reach an actual conclusion anyway. The increase in scores from before (M = 0.10, SD = 0.07) to after the game (M = 0.17, SD = 0.08) does, on the face of it, indicate that with a higher amount of participants the structural knowledge assessment may become useful. In addition, for novices the structural knowledge assessment was significantly correlated with the verbal procedural questions ($r(8) = 0.73$, $p < 0.05$).

The engagement scores were on average slightly lower for the novice participants (M = 3.36, SD = 0.37) compared with the ambulance personnel, but this was not significant $t(14) = -1.02$, $p = 0.32$. This can probably be explained by the higher relevancy the learning material has to the ambulance personnel, as is signified by the fact that one of the appraisal statements on the engagement questionnaire was 'the content of the game appealed to me'.

Another interesting thing that came out of the pilot is that most of the participants did not have any experience in playing 3D shooter games, but still had little problems navigating the world with the mouse and keyboard, something we feared would be too difficult for non-gamers. The navigation task through the corridors of the subway station before the actual triage task that had been devised to assess the player's proficiency, actually turned out to be a good way for the participants to get to grips with the controls.

One person was not engaged by the game (SOPI score < 3), experienced problems with situational awareness as evidenced by that person trying to categorize victims that had already been triaged, and reported symptoms of cybersickness. This could be due to Half-Life 2's usage of a smaller Field of View than normal, which is known to be nauseating for some (Bos, De Vries, Van Emmerik & Groen, 2010), but could also indicate a greater problem with (3D) serious games, that needs special attention in the future.

Summarizing, we contend that novices are best suited for our experiments, as they do not yet have encapsulated knowledge and show greater learning gains that could potentially be altered by the manipulations of the experiments.

4.3 Code Red Triage versus a PowerPoint instruction of the triage procedure: learning, enjoyment and cognitive load

4.3.1 The efficacy of Code Red Triage

Before ascertaining how to improve learning with Code Red Triage via game design techniques, it may be worthwhile to determine how Code Red Triage fares in comparison with more traditional instruction. Is our game more or less effective, and are there reasons to assume learning and mental model construction is hindered in a game, so that it is justifiable to target the selection, organization and integration processes of mental model construction via design techniques.

4.3.2 On cognitive load in games

In the introduction of chapter 1 we already stated that serious games are rarely more efficient than traditional instruction, and, although occasionally marred by lax research designs, also are not as effective as they could be. One reason for why some of the serious games that were validly researched were not as effective as they could have been, may lie in something we have until now only touched on very briefly, namely the cognitive load serious games impose on their player. As we mentioned in chapter 2, most of the selection, organization and integration of information into coherent knowledge structures takes place inside the working memory, which only has a limited capacity and is prone to being cognitively overloaded. If playing a game leads to cognitive overload, this could hinder the selection, organization and integration process, thereby leading to erroneous learning. It may therefore be worthwhile to see if this is the case, in order to mandate improving these cognitive processes via targeted game design techniques.

And there is ample reason that cognitive overload may occur, as playing a serious game often is a complex task, where the player has to visually attend different locations on the screen and select relevant from irrelevant information; navigate with a mouse or joystick; interpret different visual, auditory, verbal and/or tactile cues; make sense of information by organizing and integrating it into a mental model, and subsequently use this to solve problems, often within a time limit (see chapter 3 and Wouters, Van der Spek & Van Oostendorp, 2009).

From a cognitive load perspective (cf. Sweller, Van Merriënboer & Paas, 1998) one could question if this complex task (playing a serious game) is fully compatible with the human cognitive architecture, which is characterized by a limited working memory. This possible incompatibility with the human cognitive architecture is especially the case for novice learners, who have no or very limited schemas associated with the domain of knowledge in long-term memory. This possible cognitive overloading of the player during mental model construction could be a reason why a relatively large percentage of serious games fail to reach their learning goals (Van der Spek, Wouters & Van Oostendorp, 2010).

There’s a growing body of researchers that stress the importance of balancing the cognitive load a learner encounters while playing a (serious) game, e.g. Greitzer, Kuchar and Huston (2007) and Clark and colleagues (2010), and in fact they were preceded by research in multimedia instruction such as learning from animations, that encountered the same problems with then-new educational technology and led to a number of different cognition-based design guidelines (e.g. Mayer, 2003).

However, here too, while the argument seems compelling, there is only very limited actual empirical evidence for cognitive load in serious games, whether this is higher than in more traditional instruction, and its corollary effects on learning and enjoyment. In a recent experiment by Smets, Abbing, Neerincx, Lindenberg and Van Oostendorp (2010), a significantly higher mental effort was found to be exerted by a game-like environment compared to a static learning environment; however they measured the effects on the speed with which a task was performed, the results of which were mixed, instead of learning itself. In order to determine whether there really are problems with selection, organization and integration in working memory and the related cognitive load this both exerts and simultaneously impacts learning, we first have to ascertain whether cognitive load is higher in our serious game, and whether this results in poorer learning. Consequently, we performed an experiment where we compared our serious game for triage training, Code Red Triage, with a static PowerPoint presentation of the same learning material as in the game, and tested it for cognitive learning goals in the short and long term, enjoyment and cognitive load (see section 4.1.2).

Incidentally, and unbeknownst to us at the time of the experiment, Knight and colleagues (2010) also compared their triage training game with a more traditional means of teaching the triage procedure, in their case a card-sorting game. Here, they found that their serious game, called Triage Trainer, resulted in a better triage performance (as measured by the number of correct procedural steps taken) in a transfer setting than the card-sorting game. However, their research is different from ours on a number of key points. Primarily, they used the game(s) to practice the triage procedure instead of learning it per se. The procedure itself was taught in a 15-minute tutorial *before* playing the game, whereas our experiment uses the game for teaching the procedure to novices (at least in this version of the game, which is the newer version, see sections 3.2 and 5.9). Other than being told what a triage is for and which triage categories there are, everything had to be learned inferentially by the participant while playing the game. In addition, they did not test for enjoyment, mental model construction, cognitive load or delayed retention.

4.3.3 Method

4.3.3.1 Participants and design

A total of valid N = 48 undergraduate students participated in this experiment, 12 males and 36 females, who were divided equally into the game and the PowerPoint condition, i.e. 6 males and 18 females for a total of 24 participants per condition. Mean age was 21.15, SD = 1.87.

The participants performed a structural knowledge assessment, and knowledge retention and application tests both before the intervention (pretest), directly after the intervention (posttest) and a week later (delayed test). In addition we measured their cognitive load after nine victim cases and the enjoyment and cognitive load (once more) directly after the intervention. The procedure is shown in figure 4.1.

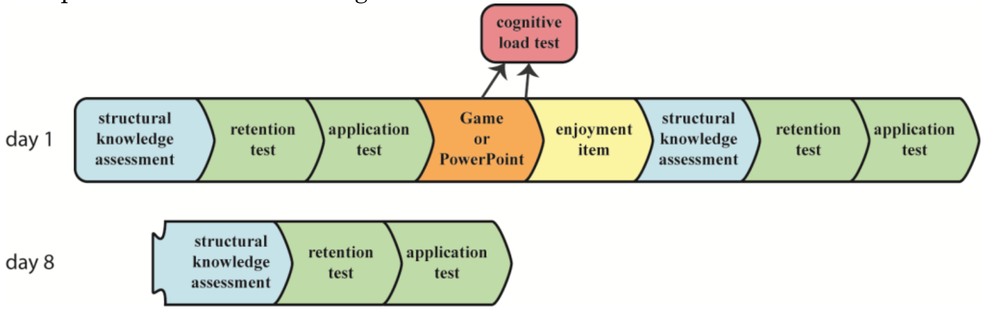


Figure 4.1 – Procedure of the experiment that compares Code Red Triage with a PowerPoint.

One participant from the PowerPoint condition did not respond to the delayed test. In addition, one participant from the Code Red Triage condition returned the tests too late and was consequently also not included in the delayed test analysis.

4.3.3.2 Materials

Participants either played the game Code Red Triage, or browsed through a PowerPoint presentation. Both of these started with a description of the context, which was a terrorist strike on the subway, and that a triage needs to be performed by a medical first responder to categorize the victims according to urgency of needed medical attention. Both the game and the PowerPoint incorporated the same 19 victim cases designed to teach the user the triage procedure, with essentially the same instruction, albeit in different forms (see figures 4.2 and 4.3 for examples).

Code Red Triage

In the case of the game Code Red Triage, the moment a player arrives at a victim, he or she can bring up the triage menu with a press of the e-button in order to start the triage. This menu is a Heads-Up Display (HUD), with eight buttons on the sides of the screen depicting the different procedure actions a person can take (e.g. ‘Airway’ to check the airway and ‘Chin lift’ to perform a chin lift); see figure 4.2. Pressing a button gives the player some general information in text on what the button depicts and roughly when this check or action is needed; this information was the same for every button for every victim. This is followed by a blank line, after which the information specific to this victim is printed; clicking the mobility button may for example return the information shown in quote 4.1.

At the bottom of the triage HUD are four triage category buttons, T1, T2, T3 and Dead, which the player can use to categorize the victim. Upon pressing this button, the victim will turn into the color corresponding to the category, and the player sees the game score change to reflect how well he or she has done, together with a short text that says whether the player has forgotten steps, took unnecessary steps and whether these were taken in the correct order. From this, the player has to deduce the correct order of button presses.

As further help to the player, the complexity of the user interface gradually increased. At the first victim the player only has two buttons to choose from. As the complexity of the victim gradually increases, that is, the number of buttons needed to resolve the correct triage category, so do the number of buttons. This also corresponds closer to the PowerPoint presentation, where only the necessary steps are given.

Figure 4.2 – the triage menu in game

> *"Determine whether victims can leave the disaster site of their own volition. If a victim is able to walk, this means that he or she is lightly injured and you can categorize him or her in a triage category.*
>
> *I can walk. I want to get out of here. Where should I go?"*

Quote 4.1 – text shown after pressing the mobility procedure button

PowerPoint
In the PowerPoint presentation of the instructional material, victims were presented per slide, with at the left side information about the victim, the necessary checks of the triage procedure for this victim and the classification of the victim into one of four triage categories; at the right side was a static picture of the victim, which was a screenshot from Code Red Triage. This information was the same as in the game, however, only the relevant information was printed and in the correct order, whereas the players in the game have to select the relevant from the irrelevant information and organize this themselves.

The general information that didn't change per check was printed in bold so that, similar to the game condition, it could be skipped if the learner wanted to. A picture of a PowerPoint slide can be seen in figure 4.3. As the participant did not have to navigate a virtual environment, the maximum allowed time with the material was shorter than in the game, 14 compared to 17 minutes. The PowerPoint was user-paced.

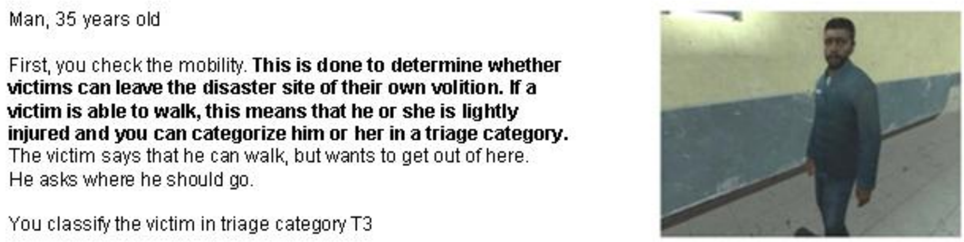


Figure 4.3 – Slide of the PowerPoint presentation

4.3.3.3 Measurements

Learning

We subdivided the knowledge test into a retention test and an application test. Both consisted of 5 verbal multiple choice questions pertaining to the triage procedure, and 5 questions with a screenshot of the game, albeit decontextualized (i.e. without seeing it's a subway station with signs of carnage), in order to eliminate bias for one of the two instructional forms. In the retention test, the participants had to answer conditional statements, e.g. 'if the victim is mobile, the next step in the triage procedure is..'. In the application test, the participants had to determine the next step in a case, e.g. 'A man of about 40 years, his respiratory frequency is normal and he has a free airway; what do you have to do next in the triage procedure?'. In both tests, the participant had to choose one of six possible answers.

The subdivision into a retention test and an application test was done to make a distinction between more declarative and more procedural knowledge. However, as we had already argued in section 4.1.1.1, the triage procedure is essentially one large set of explicit conditional if-then-else statements. Therefore, this distinction may be nonexistent, and the only difference is that the application test adds more information, making it possibly a little bit more complex. We therefore took it out of future experiments.

For the structural knowledge assessment we once again used the relationship ratings of pairs of concepts, and the subsequently created PFNets with Pathfinder software. However this time we only looked at a subset of the triage procedure, so that a more manageable 8 concepts, and consequently 28 pairings, sufficed. The referent structure was derived by averaging the elicited knowledge structures of the three supervisors of the experiment, Pieter Wouters, Erik van der Spek and Vivianne van Dijk, and can be seen in figure 4.4.

Cognitive load
We used a subjective measure of cognitive load, based on the assumption that participants are able to introspect on their cognitive processes and the amount of mental effort afforded for learning the material (Sweller et al., 1998). The measure was a 7-point rating scale, ranging from very, very difficult to very, very easy (Kalyuga, Chandler & Sweller, 2001). Recently concerns have been raised that a one-item measure cannot distinguish between different forms of cognitive load and may not be very well generalizable (De Jong, 2010), however pausing the game for a longer time period in order to administer a more elaborate questionnaire could harm the enjoyment of the game.

Enjoyment
As the participant cannot really be immersed in a PowerPoint presentation of text and the occasional image, we did not use the same engagement scale as in the pilot experiment, but instead had the participants rate how much they enjoyed the respective instructional form on a 7-point scale, ranging from very, very stupid to very, very enjoyable.

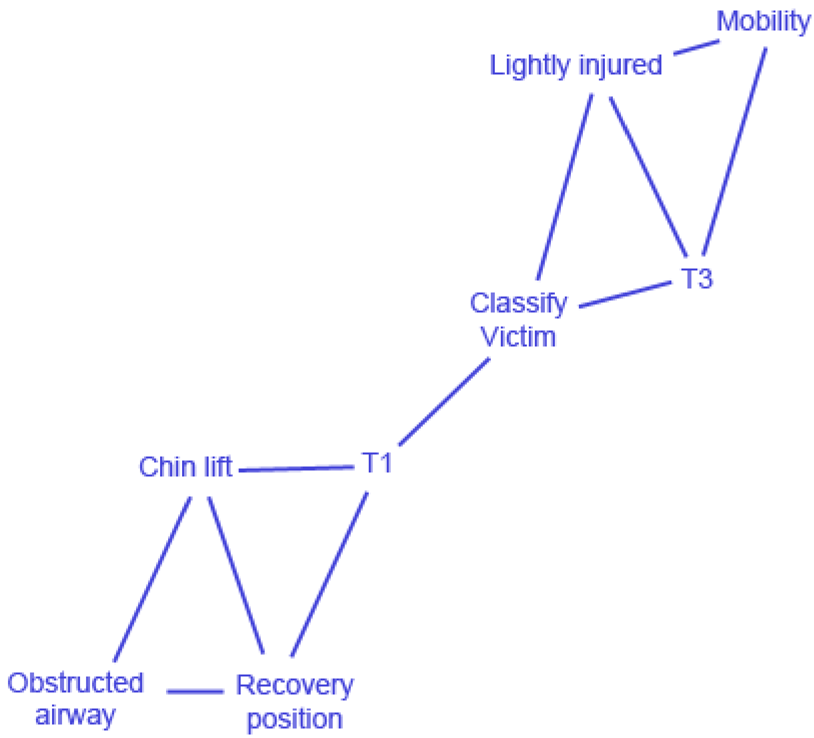


Figure 4.4 – Referent PFNet or mental model representation elicited by structural knowledge assessment

4.3.4 Results

Learning

In accordance with Vickers (2005) who demonstrated ANCOVA is superior to a Mann-Whitney U test in nearly all cases for this type of research, we performed a Repeated Measures ANCOVA, with pretest as covariate, posttest and delayed test as repeated measures and condition as between-subjects fixed factor. This returned no significant main effect of condition for the application test [$F(1,44) = 3.08$, $p = 0.09$] and the structural knowledge assessment [$F(1,44) = 0.77$, $p = 0.39$]. The main effect result of the retention test is borderline significant: $F(1,44) = 3.93$, $p = 0.05$. This implies that overall, the null hypothesis is maintained and the Serious Game and the PowerPoint version of the same information perform comparably for teaching the triage procedure, however the PowerPoint is probably superior to the Serious Game in terms of retention of facts.

More importantly, there is a significant interaction effect in the way the retention of triage facts progresses over time, for the two different instructional conditions [$F(1,44) = 7.86$, $p < 0.01$, partial eta squared = 0.15]. A graph of this can be seen in figure 4.5, Similar, albeit non-significant, trends can be seen in the other tests, as shown in table 4.1.

	Pretest	Posttest*	Delayed test*
PowerPoint retention	1.54 (0.98)	6.42 (0.50)	5.22 (0.47)
Serious Game retention	1.50 (1.25)	4.50 (0.50)	4.50 (0.46)
PowerPoint application	1.54 (1.06)	5.59 (0.40)	5.20 (0.48)
Serious Game application	1.62 (1.25)	4.45 (0.40)	4.44 (0.47)
PowerPoint structural knowledge assessment	0.04 (0.13)	0.20 (0.02)	0.15 (0.03)
Serious Game structural knowledge assessment	0.04 (0.10)	0.20 (0.02)	0.20 (0.03)

* column reports estimated marginal means (standard error), instead of mean (standard deviation)

Table 4.1 – Knowledge test and structural knowledge assessment scores

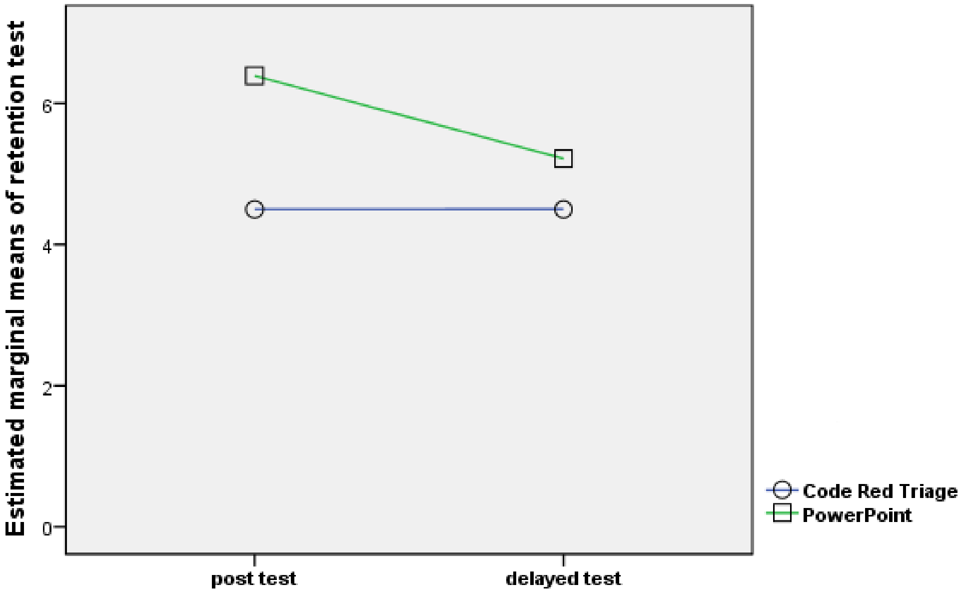


Figure 4.5 – Results of retention post and delayed test.

Performing separate ANCOVAs, with pretest as covariate and posttest and delayed test as dependent variable respectively, shows that the significant interaction effect is due to the fact that immediately after the intervention, the PowerPoint condition significantly outperforms the Serious Game condition: $F(1,45) = 7.36$, $p < 0.01$ (robust for Bonferroni adjustment), but this effect diminishes over time, and is no longer apparent a week later [$F(1,44) = 1.19$, $p = 0.28$].

Cognitive Load

Halfway into the intervention, the self-reported cognitive load was already significantly higher in the Serious Game condition than in the PowerPoint condition ($M = 4.25$, $SD = 1.15$, vs. $M = 2.75$, $SD = 0.53$ resp., $t(32.38) = 5.79$, $p < 0.001$). Directly after the intervention, this effect remained significant, with the Serious Game condition $M = 4.29$, $SD = 1.37$ and the PowerPoint condition $M = 3.29$, $SD = 0.69$ [$t(34.01) = 3.20$, $p < 0.005$].

The cognitive load during the intervention was also negatively correlated with the reported enjoyment, for both the PowerPoint condition [$r(22) = -0.54$, $p < 0.01$] as well as the Serious Game condition [$r(22) = -0.42$, $p < 0.05$].

Enjoyment

Contrary to our hypothesis, there was no significant difference between the PowerPoint and the Serious Game condition in the reported enjoyment of the teaching material. PowerPoint $M = 4.62$, $SD = 0.65$; Serious Game $M = 4.88$, $SD = 1.30$; $t(33.79) = 0.85$, $p = 0.4$.

Note that the standard deviation in the Serious Game condition is twice the size of the PowerPoint condition, the variance in the two conditions is significantly different (Levene's test $p < 0.01$). Some of this may be explained by a number of the participants confessing after participating in the experiment that they in fact did not like playing games at all. However there was no 'preference for playing games' question in the questionnaire, so this evidence is anecdotal at best.

Factors correlated with learning
From the data we can discern a number of plausible reasons as to why the PowerPoint condition outperformed the Serious Game, at least directly after the experience. We will present the results here, and discuss them in the conclusion.

The experienced enjoyment of the game is significantly correlated with the scores on the retention post-test [$r(22) = 0.62$, $p < 0.005$], retention delayed test [$r(22) = 0.49$, $p < 0.05$], application post-test [$r(22) = 0.69$, $p < 0.001$] and application delayed test [$r(22) = 0.68$, $p < 0.001$]. The in-game score and post-test enjoyment rating of the game are furthermore highly and significantly correlated [$r(22) = 0.64$, $p < 0.005$]. Conversely, the participant's enjoyment rating was not significantly correlated to any of the test scores for the PowerPoint condition.

A second factor that could contribute to the results of the knowledge tests is the cognitive load experienced during the time the participants spent with the two forms of instruction. The averaged experienced cognitive load in the Serious Game condition was significantly negatively correlated with the retention post-test directly after the [$r(22) = -0.52$, $p < 0.01$], as well as for the retention delayed test [$r(22) = -0.56$, $p < 0.01$]. Cognitive load was also negatively correlated with the application posttest [$r(22) = -0.65$, $p < 0.005$] and application delayed test [$r(22) = -0.55$, $p < 0.01$] for the Serious Game condition.

The same does not hold true for the PowerPoint condition, where none of the post and delayed tests were significantly correlated with the experienced cognitive load during the experiment.

The structural knowledge assessments of the constructed mental model were not significantly affected by cognitive load in this experiment.

4.3.5 Conclusion on the efficacy of Code Red Triage

The aim of this experiment was to determine if there was a difference between our game Code Red Triage and a more traditional PowerPoint instruction in terms of learning gains, and whether factors can be discerned that influence learning, so that these can be targeted by game design techniques.

We found that in the short-term, the PowerPoint presentation is superior to the game, at least for the retention test. Over longer time periods this difference starts to disappear however, giving some credibility to the possibility that what is learned by playing a serious game appears to stay with the learner longer. It is unclear whether this trend continues over time, maybe even leading to the, arguably unlikely, moment that the serious game outperforms the PowerPoint presentation in the long run, but it's interesting to note just how robust the serious games scores stay over time in table 4.1. There was no apparent difference in the structural knowledge assessment scores.

We also found that the reported enjoyment in the game condition was positively correlated with the post knowledge test scores. Therefore, motivating the players of a serious game seems paramount for learning to occur. Conversely, the participant's enjoyment rating had no significant influence on any of the test scores for the PowerPoint condition. Consequently, one could argue that the slightly worse performance by the Serious Game could be due to some of the participants not liking games in general, or this game in particular, meaning they were not invested in trying to reach a high score.

However, the correlation could also run the other way. After triaging a victim in the game, the player would get feedback on how well he did. Much of what makes a game fun is the feeling of self-efficacy a player builds up while playing a game, so the positive feedback provided to those who are good at the game may go some way in explaining the correlation between enjoyment and post knowledge test scores. This is corroborated by the fact that the in-game score and the posttest enjoyment score were also highly correlated. There was no correlation between reported enjoyment and the structural knowledge assessment, which would indicate that deeper learning occurs regardless of whether the game is enjoyed.

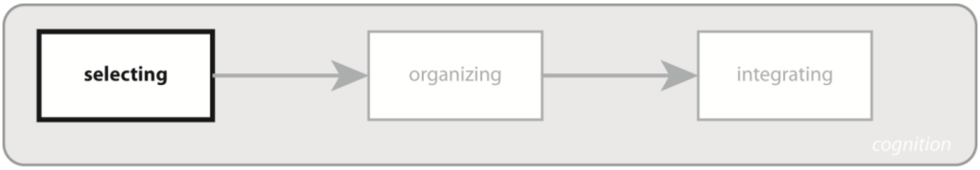
Perhaps more importantly, we discovered that the serious game exerted a significantly higher cognitive load on the user, and that this was negatively correlated with the learning gains. This indicates problems with cognitive load in the working memory at the time during which the triage procedure has to be learned. The PowerPoint and the serious game both incorporated the same instructional information, and both were supplied to the player via text, so the difference has to lie in other aspects pertaining to presentation form.

The serious game has a much higher perceptual richness, placing a potentially higher load on the selection process of the player. At the same time, the serious game is also more interactive, where the information is not presented in a linear fashion and preorganized by the system. Instead, the player first has to select the relevant bits of information, and then also has to organize these into the correct, meaningful order him or herself. It is likely that the higher taxation of these cognitive processes led to a higher cognitive load, as well as to poorer learning. Therefore we contend that it is justified to decrease cognitive load by aiding the player in the selection and organization process via game design techniques, which in turn should lead to better mental model construction.

Curiously, we found no difference in the rated enjoyment of the game, meaning that the participants enjoyed Code Red Triage and the PowerPoint presentation to an equal degree. This does correspond to the findings of Hays (2005) that there is little proof for the extra enjoyment that games would provide; however there's a problem with our research design that casts a doubt on the validity of this result. The participants only witnessed one condition and consequently lacked a clear reference with which to judge it by. It could very well be that the participants in the PowerPoint condition likened the instruction to other types of educational material such as an algebra book and, due to the interestingness of the triage procedure, responded to this favorably. Participants in the Code Red Triage condition may have expected to play an action-packed entertainment game, only to find out that it is an educational game with no shooting aliens whatsoever; as a result they could've rated it lower. Even though we now lack the data to back this up, we still think that it is likely that, when compared side to side, people will prefer the serious game.

Chapter 5

A comparison of auditory and visual guidance cues in a serious game



Abstract—Games, and serious games especially, revolve around learning new material and integrating this into a mental model. However, playing games can be cognitively demanding and novices may fail to distinguish between relevant and irrelevant information. To overcome this, one can try to subtly cue the attention of the player to the relevant material. We empirically tested the use of auditory and visual cues to guide the player in learning the correct procedure in a 3D serious triage training game. Learning and mental model construction was significantly worse in the auditory cueing condition, as compared to a control group with no cues. The visual cueing condition was on par with the control condition in terms of overall learning gains, but we found that within the visual cueing condition, experienced game players significantly outperformed participants that did not play games, likely due to the experienced gamers being familiar with visual guidance cues. Implications for game design and future research are subsequently discussed.

Parts of this chapter are based on:

Van der Spek, E. D., Van Oostendorp, H., Wouters, P., & Aarnoudse, L. (2010). Attentional Cueing in Serious Games. In K. Debattista, M. Dickey, A. Proenca, & L. P. Santos (Eds.), *Second International Conference on Games and Virtual Worlds for Serious Applications VS-GAMES 2010* (pp. 119-125). Braga, Portugal: IEEE. doi:10.1109/VS-GAMES.2010.8

5.1 Introduction

All games incorporate learning. Good games constantly throw up hurdles and obstacles a player must learn to overcome. In fact, much of what makes a game enjoyable in the first place comes from the mastery a player needs to develop to overcome obstacles in a game and the corresponding boost in self-efficacy (cf. Klimmt & Hartmann, 2006). Given this combination of engagement and learning, it is therefore unsurprising that games used to teach people non-trivial instructional material, so called serious games, are garnering more and more popular attention; however the results of serious games in terms of learning goals so far have been mixed (Vogel et al., 2006; Wouters et al., 2009).

One important reason why serious games sometimes fail to reach their learning goals, is because games offer a rich multimodal and thereby cognitively too demanding experience. As a person's working memory only has a limited capacity (cf. Paas, Renkl and Sweller, 2003), and selecting, organizing and integrating sensory data into a mental model happens within the boundaries of this working memory (Mayer, 2001), such a rich experience may be too cognitively demanding, resulting in incorrect mental models of the instructional material, as was evidenced by Nelson and Erlandson (2008). In addition, we previously witnessed a similar outcome with our game in chapter 4, where the higher cognitive load in Code Red Triage compared with a PowerPoint version of the same game likely led to lower learning gains. Finding ways to optimize cognitive load during the design of the game is therefore an important step in creating effective and instructionally sound games (cf. Kiili, 2005).

5.2 Improving selection by attentional cueing

A number of ways in which the cognitive load in games may be mitigated have been proposed by, among others, Moreno and Mayer (2007), and Van der Spek, Wouters and Van Oostendorp (2008; also see chapter 2). Here, we will focus on using cues in order to guide a player's attention to the relevant information items at specific moments occurring in a game.

The rationale behind using cues is twofold. Firstly, a player may sometimes not know where to look for the information that is relevant to understanding the underlying instructional principles. This is paramount when the relevant information is less salient than irrelevant information, especially for novices who do not yet know how to distinguish relevant from irrelevant material (Lowe, 2003; Van Oostendorp, Beijersbergen & Solaimani, 2008), but also when the perceptual system is overloaded with too much concurrent information to attend to and select the relevant material in the first place (Lavie, 2005). As this visual search for relevant instructional material is for a part extraneous, and thus unnecessary, cognitive load, one generally wants to limit this as much as possible.

Secondly, a person's working memory can be overloaded with too much concurrent information, which hinders the correct processing of relevant material. In the case of our serious game Code Red Triage, we noticed that the selection, organization and integration of the relevant material under time pressure led to a higher cognitive load and lower learning gains (see chapter 4).

Attentional cues then, are a way of guiding the attention of the player to the relevant material, or away from the irrelevant material, by means of deliberately placed sensory stimuli. It is therefore a way to improve the selection process in multimedia learning. This can be achieved, for instance, by artificially making the relevant information more salient, by inserting sounds or smells that the learner associates with the relevant information, or by introducing visual arrows that point to the objects to be learned.

Attentional cues are a natural fit for gaming and are starting to become widely applied, for instance in *Mirror's Edge* where the process of wayfinding is sped up by coloring the objects one has to run towards bright red, or in *Fable 2* where non playable characters (NPCs) emit a colored halo as a way of communicating to the player which NPCs are friendly and which hostile. These visual cues intuitively work well, but scientific evidence of their value in games remains scarce. In the case of cueing wayfinding, Steiner and Voruganti (2004) compared a number of different techniques to guide a player, who was instructed to simply explore the surroundings, to a target in a virtual environment. They distinguished between and compared: verbal cues, i.e. an NPC giving directions; written cues, i.e. signposts; a building landmark, i.e. a far off tower; environmental landmarks, i.e. fire and a path landmark, i.e. a glowing trail. All of these led the players to where the experimenters intended, but the glowing trail was significantly better in doing this.

Conversely, the glowing trail and the NPC giving directions were found to be the most explicit means of guiding, whereas the tower and fire were mostly not perceived as guidance at all. Interestingly, and perhaps as a consequence of this study, *Fable 2* also employs a glowing trail (of what they call 'breadcrumbs') to guide a player to the next important spot in a given quest.

Two caveats to the previous examples of guidance cues should be given however. Firstly, while cues are regularly used in games to nudge a player in a certain direction, it is unclear how they would fare in the context of learning. Do cues, for instance, relieve the burden on the selection process and thereby stimulate deeper processing, are they only useful in dictating and automating certain procedural patterns, or will they even inhibit deeper processing and actually make the player a lazy learner? A number of experiments done on the effect of cueing in multimedia learning other than games, support the notion of cueing reducing cognitive load and benefiting learning (cf. Jeung, Chandler & Sweller, 1997; Kalyuga, Chandler & Sweller, 1999). Furthermore, although they did not find that cueing reduced cognitive load, De Koning, Tabbers, Rikers and Paas (2007) reported that, not only did cueing help in learning the cued material in the context of learning from an animation, but even improved learning of the uncued material. The conditions in the experiments of De

Koning and colleagues (2007) and Jeung and colleagues (1997) were however system-paced, which makes it problematic to compare to a game setting. In addition, it should be noted that these results are not uncontroversial, as Tabbers, Martens and Van Merrienboer (2004) only found small significantly positive results on retention tests, and too much guidance in a problem solving context can even lead to lower learning gains (Van Nimwegen & Van Oostendorp, 2009).

Secondly, many games aspire to induce a sense of presence, the sense of psychologically being inside the virtual environment (cf. Slater & Wilbur, 1997). The level of presence-inducing qualities of a game may be linked to the enjoyment of a video game (Tamborini & Skalski, 2006) and, while research on whether it influences task performance is still inconclusive (e.g. Mania & Chalmers, 2001), one can surmise that training fire drill procedures, or in our case a medical procedure in a crisis situation, correctly, is contingent on inducing the same levels of stress as the real world situation. The same applies to games that are used for phobia-alleviation. For a detailed overview on the importance of presence in virtual environments and serious games, we refer to Schuemie, Van der Straaten, Krijn and Van der Mast (2001).

A brightly colored halo around an NPC or a glowing trail, as in the examples above, may certainly be effective, but is not very realistic, and could therefore harm the immersive qualities of a game. In addition, games and their players can be quite competitive and some people may not want to be helped; guidance cues may come off as paternalistic, harming the feeling of autonomy of the player, which in turn leads to lower enjoyment of the game (cf. Przybylski, Ryan & Rigby, 2010). As Steiner and Voruganti (2004) showed that subtle cueing can also be effective, we therefore contend that guiding the player into the right learning direction, at least when it comes to games, should be done subtly and the effect on the engagement of the game should always be tested.

5.3 Code Red Triage

In order to test whether cueing improves learning gains from a serious game, we created a 3D first-person game, Code Red Triage, which trains players in performing a primary (sieve) triage, a technique for medical first responders to swiftly categorize the victims of a mass casualty event in order of urgency of needed medical attention (see chapter 3).

The procedure in the game is performed by means of a menu, with eight buttons portraying necessary as well as irrelevant steps, the information pertaining to the buttons in the middle and the triage categories at the bottom. There was a help button in the top right of the screen which players could press to see a flowchart of the triage procedure on-demand. This image of the procedure can be seen in figure 5.1, the triage elicitation and categorization menu in figure 5.2.

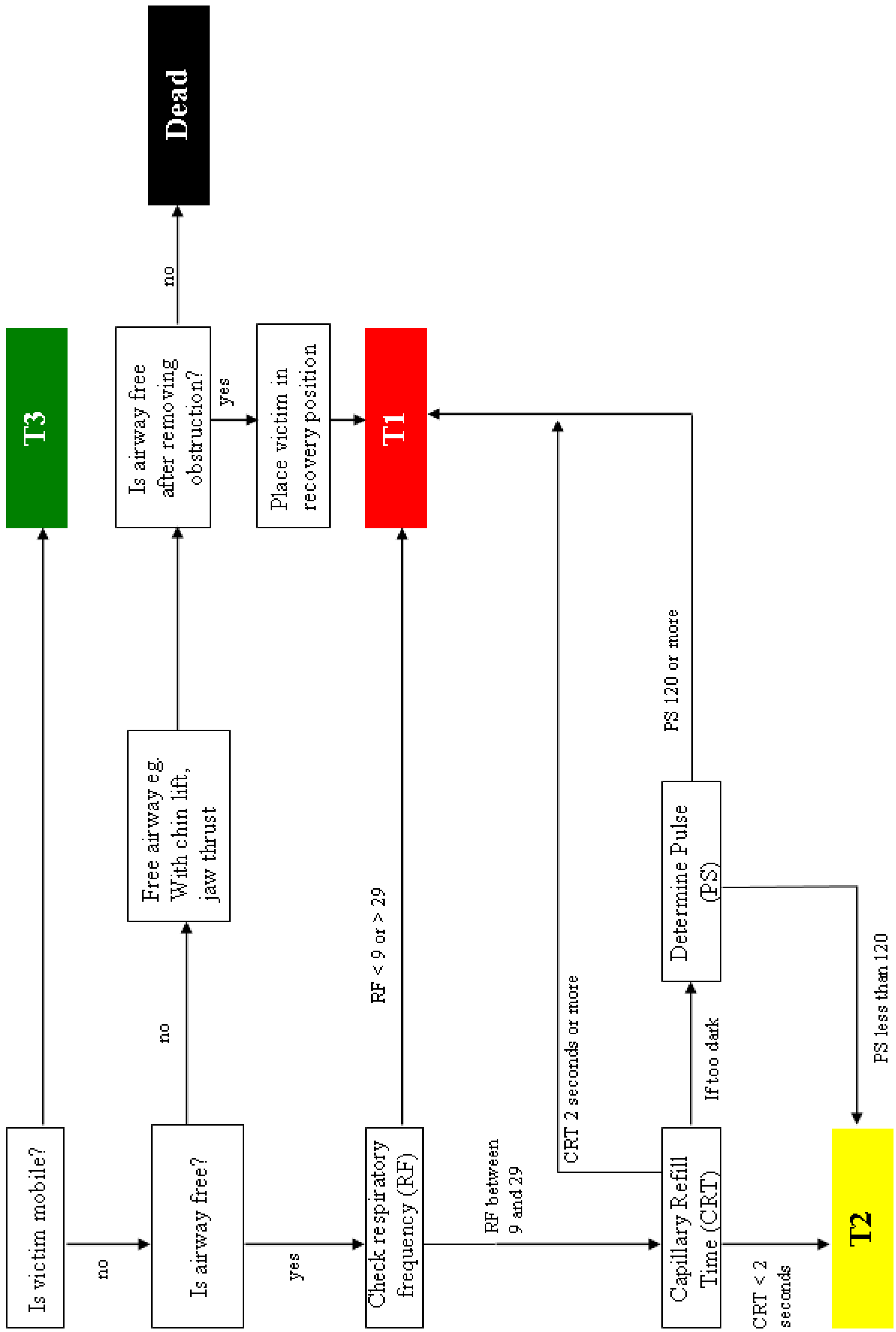


Figure 5.1 – Triage procedure flowchart that could be accessed on-demand in the game

Figure 5.2 – Triage menu in-game

5.4 Auditory and visual cues

In order to test our assertion that subtle cues can improve learning from a serious game while retaining the same level of engagement, we performed an experiment with three conditions: a control condition with no cues, a subtle condition where the cues were presented by sounds, and a more explicit (but still subtle) condition where the cues were visual. There is a possibility that cues serve as an externalization of information that should be internalized (cf. Van Nimwegen & Van Oostendorp, 2009); in our case, if we were to highlight the button that should be pressed next, the player could be conditioned to follow the highlights instead of the content of the button. In order to prevent this, the cues didn't directly state which button should be pressed next, but rather implied this by association. Another reason for this choice is that, as we discussed in chapter 2, cues can be used to aid in selection, but also to aid in organization (signaling). This experiment is designed to help the player in selecting the relevant information; if we were to highlight the buttons in the correct order, the player could use the highlighting pattern to discern the correct organization of the procedure. Conversely, by keeping the cues out of the user interface and in the virtual environment, the player would be nudged towards the right button, instead of the right order. The precise implementation is elaborated on in the method section, 5.6.

While we already discussed examples of the widespread use of visual cues in games in the introduction, auditory cues are less often used. The decision to choose auditory cues lies in Mayer's cognitive theory of multimedia learning (2003) which in turn builds on previous research done by Paivio (1986). According to this, people can process auditory and visual information concurrently. Information in one modality can be offloaded into the other, for instance by turning written text into spoken text, which was done in a virtual learning environment in an experiment by Erlandson, Nelson and Savenye (2010). This resulted in a significant reduction in cognitive load, although no measurable improvement in learning. Nonetheless, it demonstrates the viability of using auditory cues in contemporary games with a high visual information load, and our Code Red Triage where the learning material is wholly visual especially, because they do not add cognitive load in the already heavily taxed visual channel. This notion is further corroborated by an experiment performed by Brown, Newsome and Glinert (1989), who found that under certain conditions when speed is not crucial, auditory cues work just as well as visual cues in a visual search task; it should be noted however that in Code Red Triage, players do operate under time pressure.

Auditory cues may therefore be a good fit for improving learning in serious games that impose a heavy load on the player's visual processing system. We therefore hypothesize that learning from our serious game will improve the most from auditory cues. The visual cueing condition will likely also see an improvement in learning, but at the cost of engagement of the game. Here, improved learning is defined as performing significantly better on knowledge tests than a control group and also having a mental model that is closer to a referent mental model. Closely related to this, we additionally hypothesize that providing auditory cues reduces the reliance of players on additional, flow-breaking help material, where by flow-breaking help material we mean using the procedure button to see the triage procedure flowchart (figure 5.1).

5.5 Mental Models

In order to elicit the way in which information is stored in the mental model of the participants of our experiment, we use Pathfinder software Schvaneveldt (1990). Pathfinder is essentially a computer program that takes a set of concepts and gives these concepts back to the user as a randomized set of pairs. The user has to rate each pair of concepts on how much he or she thinks they are related, and from the resulting matrix of interrelatedness, Pathfinder draws so-called associative networks. These associative networks are (un)directed graphs, that give a pictorial view of how information is represented in the cognitive structure of the participant, and can be compared to another by means of a similarity measure. By calculating the similarity measure of a participant's mental model to a referent 'correct' mental model it is thus possible to effectively calculate how much a participant's mental model has improved (or deteriorated) after playing our game. The ‘correct’ referent mental model we created can be seen in figure 5.3.

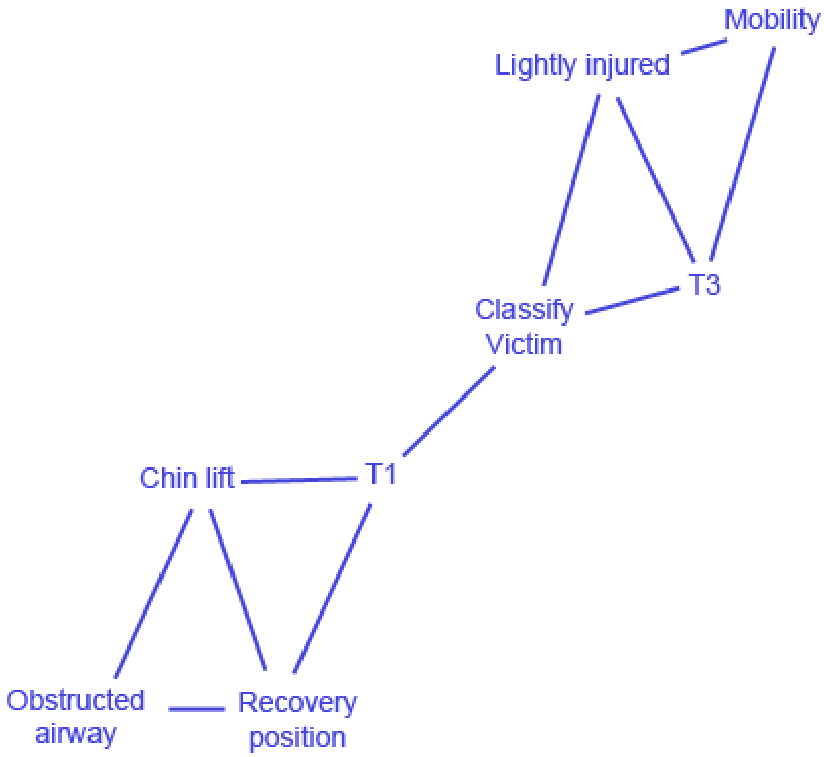


Figure 5.3 – Referent mental model constructed with Pathfinder

It is only a subset of the total triage procedure; in the pilot test in chapter 4 (cf. additionally Van der Spek et al., 2011) we had a larger portion of the procedure, but the number of combinations explodes rapidly with every added concept, making it tiresome for the participant and therefore unreliable. We furthermore deliberately chose to include conceptual items as well as procedural items. If these are connected to each other one gets what Kahler (2002) calls strategic knowledge, and it gives a further indication of how robust the mental model is to when the participant encounters other related problems (see also section 4.1.1.2).

5.6 Experiment

5.6.1 Participants

A total of 41 persons, mostly university students from different studies, participated in this experiment, however 2 were discarded after the experiment; one due to irregularities with the game, the other due to having triaged only 10 victims (out of 19) before the end of the game, therefore Valid *N* is 39. Mean age of the remainder was 22.31 (*SD* = 2.76), and of these, 20 were female and 19 male. We asked the participants to rate their prior game experience by letting them choose one of three options to describe themselves: 'I hardly ever play games', 'I occasionally play games' and 'I'm a gamer'. However, only 4 out of 39 participants considered themselves a gamer, which is too small a number to measure any effects. We therefore chose to combine the second and third choice into a group that has notable prior experience with games. A breakdown of the makeup of the groups can be seen in table 5.1.

	Control	*Auditory cues*	*Visual cues*
No of participants	13	14	12
Male - Female	5 - 8	7 - 7	8 - 4
Game experience*	6 - 7	6 - 8	6 - 6

* *I hardly ever play games – I occasionally play games & I'm a gamer*

Table 5.1 – Breakdown of the participants per condition

5.6.2 Experimental conditions

As was stated previously, the experiment consisted of three different conditions: a control group with no cues, an auditory cueing condition with rather subtle, non-immersion breaking cues and a visual cueing condition with more explicit, possibly immersion breaking clues.

For the auditory cueing condition, we chose procedure cueing sounds that would not seem out of place when one tries to triage victims. In our case these sounds were: a footsteps sound to cue checking the mobility of a victim, a snoring sound to cue checking the airway, a fast breathing sound to cue the respiratory rate and a beating heart sound to cue checking the capillary refill time.

For the visual cueing condition, we let the next step in the procedure be cued by a glowing and pulsating green arrow, pointing at the area of the victim's body pertaining to the check that should be performed. The arrow pointed at the right leg of the victim to cue mobility, at the mouth to cue checking the airway, the breast to cue breathing frequency and the right hand to cue measuring the capillary refill time. See figure 5.4 of a picture of the visual cue in the game.

Figure 5.4 – Visual cue hinting at checking the airway

All three conditions had background music and other sounds, such as screaming, but the moment the player walked up to a victim and pressed the e-button to commence the triage procedure, this sound would be automatically turned down in both conditions to a barely audible level. In both cueing conditions, the mobility cue was shown or heard when the player starts triaging a victim. If the procedure should end there, for instance when the victim is still able to walk and is therefore deemed to be lightly injured, no further sounds or arrow were played or shown. If on the other hand another step in the procedure should be taken, the cue for the next correct button was then played or shown, for all the necessary steps up until the capillary refill time.

In line with the cueing study of Steiner and Voruganti (2004), no information was provided regarding the auditory and visual cues. This means that any effect measured of a cue is in fact a combination of the efficacy of the cue itself, as well as the intuitiveness with which a player can pick up the intention of the cue.

5.6.3 Materials

We used three different tests to measure learning, a pen and paper knowledge test, consisting of ten item multiple choice questions with combined conceptual and procedural items pertaining to the triage procedure; the Pathfinder structural knowledge assessment that we described above; and the in-game score, which was based on whether the participant placed the victim in the right triage category and how well they followed the procedure.

To measure engagement, we used the engagement subset of the ITC-SOPI questionnaire (Lessiter et al., 2001; see also section 4.1.2), a Likert-like questionnaire that's specifically designed to measure all types of presence that any multimedia setting can evoke. One of these types of presence is strongly related to engagement, resulting in a good mix of presence and engagement related questions (e.g. "I would have liked the game to go on longer"). The other factors that constitute presence are 'sense of physical space', 'ecological validity', and 'negative effects'.

5.6.4 Apparatus

The game was made with the Source SDK, a development kit created and distributed freely by Valve Corporation alongside *The Orange Box*, for the creation of modifications ('mods') of Half Life 2 and other games that run on the Source engine. The game was played on a Dell XPS M1730 gaming laptop, with a 17" widescreen monitor, a dual GeForce 8800GTX 3D card and with a large Sony headset. The game ran smoothly on maximum detail and lighting settings with no hiccups in the framerate and only some minor (1 second or less) lag between pressing the triage procedure button and the corresponding menu to pop up. The laptop was set up in a closed laboratory room, with a movable screen on all sides to ensure a minimum of environmental distraction.

5.6.5 Procedure

The participants were sat in the enclosed space behind the computer and started by filling in a short demographics questionnaire. The first test was (1) the Pathfinder exercise to gauge their mental model, after which (2) they completed the knowledge test. Next they received some basic instruction about how to navigate through the game environment.

After this the lights were turned off and they were allowed to play the game. There was no time limit on playing the game until they reached the subway platform, this was done to make the player acquainted with the controls and to measure how fast they could navigate through a 3D environment, to provide us some indication of gaming proficiency. However it quickly became apparent that this was not a good indicator, as experienced gamers seemed more inclined to wander off in different directions out of curiosity or in order to test the environment for glitches than inexperienced gamers, who seemed more inclined to pursue the game objective. We therefore used the self-report questionnaire as was detailed in section 5.6.1.

Upon reaching the platform, a visible timer counted down from 17 minutes, to give the player a sense of urgency; after 17 minutes the game ended. All but 2 participants found all 19 victims. Within the game we tracked the score of the player, which was fed back to the player via a progress bar at

the top of the screen. Every victim classified correctly could award the player a score of 100 points (for a total of 1900 points), but points were deducted from this score for every essential step that was forgotten, taken in the wrong order, or steps that shouldn't have been taken. Additionally, up to 40 points time penalty were deducted; one point for every second longer than a preset time per victim. The (3) total score would then give us an accurate depiction of how well the players performed in the game.
When the game was finished, the lights were switched on again and the participants immediately had to fill in (4) the engagement questionnaire. After this the same procedure was followed as before the game: first (5) the Pathfinder test and then (6) the paper test, but with the items in a different order. On top of this we gave the participants (7) eight victim case descriptions where the participant had to decide on the next step of the procedure (i.e. given this victim, at this moment in the procedure what is the next step in the primary triage), four of which were verbal, and four of which were pictorial, as in the game. This more closely resembles what the participant did in the game, i.e. applying the procedure to a victim case, and is therefore an application test. An overview of the procedure can be seen in figure 5.5.

Figure 5.5 - Procedure

5.7 Results

The means and standard deviations of the pretest and posttest knowledge test and the structural knowledge assessment can be seen in table 5.2. All three conditions significantly improved on their knowledge tests, control condition: $t(12) = -7.9$, $p < 0.001$; auditory cueing condition: $t(13) = -6.6$, $p < 0.001$; and visual cueing condition: $t(11) = -8.6$, $p < 0.001$. The structural knowledge assessment showed similar significant improvements of the posttest over the pretest, control condition: $t(12) = -6.0$, $p < 0.001$; auditory cueing condition: $t(13) = -2.8$, $p < 0.05$; and the visual cueing condition: $t(11) = -3.5$, $p < 0.01$.

	Condition	Pretest	Posttest
Knowledge test	Control	3.0 (1.3)	7.5 (1.6)
	Auditory cues	3.7 (1.2)	6.4 (1.7)
	Visual cues	2.9 (1.2)	7.6 (1.8)
Structural knowledge assessment	Control	0.045 (0.105)	0.308 (0.127)
	Auditory cues	0.037 (0.061)	0.137 (0.124)
	Visual cues	0.039 (0.113)	0.307 (0.232)

Table 5.2 – Means and (standard deviations) of pretest and posttest scores on verbal and structural knowledge

Knowledge test
An ANCOVA with condition (control vs. auditory cueing vs. visual cueing) as fixed factor, the knowledge posttest as dependent variable and controlling for the knowledge pretest as covariate showed a significant effect of cueing $F(2,35) = 3.29$, $p < 0.05$, partial $\eta^2 = 0.16$. A Fisher LSD post-hoc test revealed that this is due to a difference between the control and the auditory cueing condition ($p < 0.05$, $d = 0.87$) and between the auditory and the visual cueing condition ($p < 0.05$, $d = 0.92$). The auditory cueing condition scored lower on the knowledge test than both the visual cueing and the control condition. There was no significant difference between the control and the visual cueing condition. A bar chart of the means can be seen in figure 5.6.

Structural knowledge assessment
An ANCOVA with condition (control vs. auditory cueing vs. visual cueing) as fixed factor, the posttest structural knowledge assessment as dependent variable and controlling for the pretest structural knowledge assessment as covariate showed a significant effect of cueing, $F(2,35) = 4.63$, $p < 0.05$, partial $\eta^2 = 0.21$. A Fisher LSD post-hoc test revealed that this was due to the difference between the control and the auditory cueing condition ($p < 0.05$, $d = 1.0$) and the difference between the auditory and the visual cueing condition ($p < 0.05$, $d = 1.0$). In these cases participants in the auditory cueing condition scored significantly lower than the control and the visual cueing condition respectively. There was no significant difference between the control and the visual cueing condition. A bar chart of the means can be seen in figure 5.7.

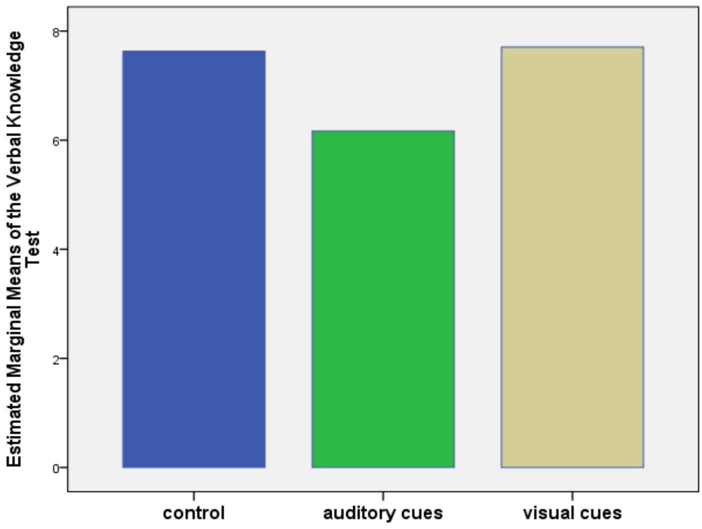


Figure 5.6 - Bar chart of knowledge test scores

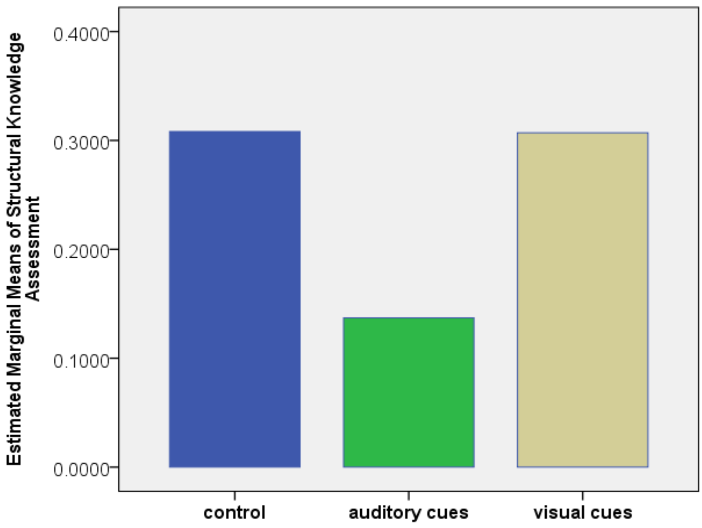


Figure 5.7 - Bar chart of structural knowledge assessment scores

Application test

The application test at the end (verbal and visual items) was not significantly affected by cueing, a MANOVA with cueing as fixed factor and the verbal and visual items as dependent variables: Wilks' $\lambda = 0.87$, $F(4,70) = 1.29$, $p = 0.28$.

In-game score
Cueing did not have a significant effect on the total in-game score at the end of the game [$F(2,31) < 1$], nor did the progression of scores differ per cueing condition over time in the game (i.e. there was no significant interaction effect of cueing on the repeated measures variable victim score, $F(20.94, 324.55) < 1$. See figure 5.8.

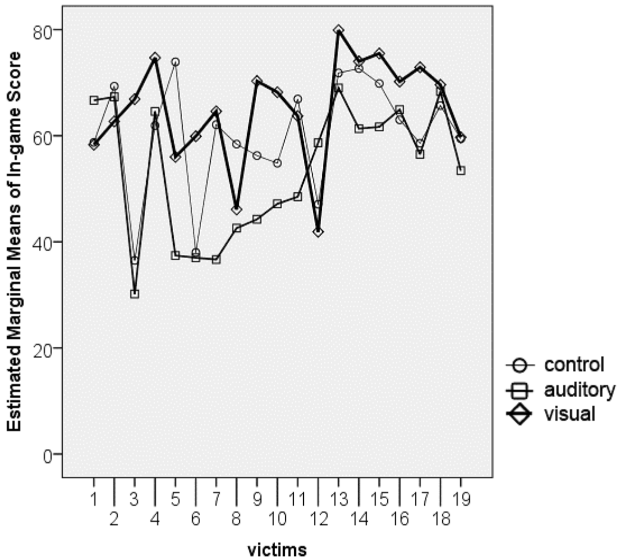


Figure 5.8 – Progression of in-game score over subsequent victims

Engagement
The ITC-SCOPI engagement ratings of the game were unaffected by the cueing condition, $F(2,36) < 1$. The scale was reliable, coefficient $\alpha = 0.82$. Engagement was positively correlated with the pictorial application test [$r(37) = 0.33$, $p < 0.05$] (only), giving some indication that engagement could be important in transfer settings.

Use of in-game help
During the triage portion of the game, the participants were able to access a flow chart of the triage procedure (figure 5.1) as a helpful reference. We had initially hypothesized that in the cueing conditions the need for this help would disappear, as opposed to control condition, which would result in less time spent studying a static picture, and more time in the game. However, no significant difference of cueing was found between the groups, $F(2,36) < 1$. Possibly as a result from this, there was also no significant effect of cueing on the average time spent per victim case, $F(2,36) < 1$.

Moreover, the help function was utilized a lot more than we expected, with an average number of 24, 21 and 20 times for the control, auditory cueing and visual cueing condition respectively. The number of times the triage procedure help was consulted was a significant predictor of the in-game score in the control condition [$b = 0.58$, $t(10) = 2.37$, $p < 0.05$], the auditory cueing condition [b = 0.57, $t(11) = 2.38$, $p < 0.05$], but not for the visual cueing condition [$t(9) = -0.57$, $p = 0.58$].

Game experience

A possible explanation for the lack of predictive power in the visual cueing condition, is that another factor is responsible for a lot more of the variability. Figure 5.9 shows what happens to the respective in-game scores per cueing condition when we distinguish between the participants' prior game experience. In the visual cueing condition, the effect of game experience on the in-game score is significant, $F(1,10) = 5.88$, $p < 0.05$, $d = 1.4$.

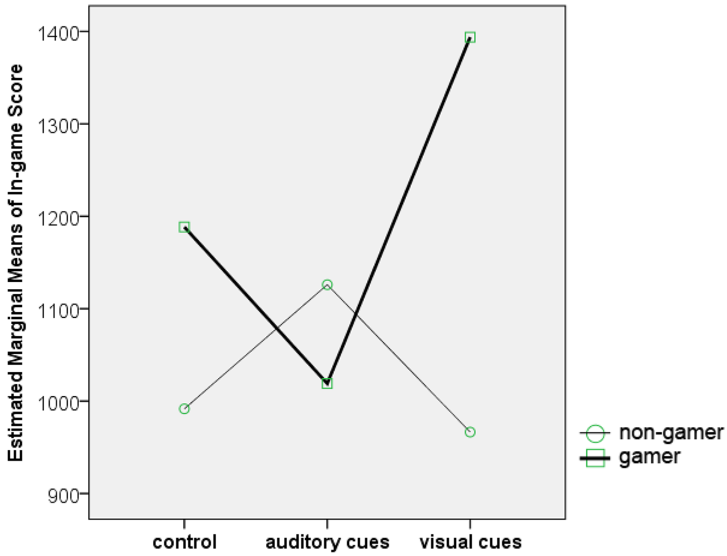


Figure 5.9 – In-game score differentiated for game experience

A similar pattern reveals itself in the graph of the knowledge test, where gamers and non-gamers score almost equally in the control and auditory cueing condition, but a disparity reveals itself in the visual cueing condition. The effect of game experience is not significant, but has a large effect size: $F(1,9) = 2.50$, $p = 0.15$, $d = 1.12$. There is no ostensible effect of game experience on the structural knowledge assessment scores, $F(1,32) < 1$.

If we delete the participants who hardly ever play games and select only the gamer population and subsequently perform an ANOVA on the effect of cueing on in-game score, this is significant [$F(2,18) = 3.89$, $p < 0.05$, $\eta^2 = 0.30$], a post-hoc Tukey HSD test indicates that the visual cueing condition performs significantly better than the auditory cueing condition, $p < 0.05$.

Validity of the measurements

The pictorial application test questions did, however, correlate the strongest of all the measures with the in-game score: $r(37) = 0.73$, $p < 0.001$. The in-game score furthermore was correlated with the knowledge test [$r(37) = 0.45$, $p < 0.01$], and the verbal application test [$r(37) = 0.53$, $p < 0.005$]. These correlations give plausibility to the notion that the in-game score is a valid measure of learning, as well as that the knowledge and application test are valid means of measuring what has transpired in the game.

The in-game score was not significantly correlated with the structural knowledge assessment, however the structural knowledge assessment was correlated with the knowledge test: $r(37) = 0.42$, $p < 0.01$. This could indicate that the in-game score at least is not a very good measure for deep learning (see section 4.1.1.1).

5.8 Conclusion and discussion

Auditory cueing

The conclusion that can be drawn from our experiment is that auditory cues alone, as implemented in our experiment, do not suffice in helping the player to correctly learn a procedural task in a serious game. Contrarily so, both the results from the knowledge test and the structural knowledge assessment were negatively affected in the auditory cueing condition. The application test was not positively or negatively affected by auditory cues.

When the experiment was set up, we were expecting to see either a positive result or no result of the auditory cues and therefore, given the paucity of the participants at hand, chose for a control condition with no cues and two conditions with relevant cues. Learning clearly suffered from the auditory cues; it is likely that the cues weren't recognized (consciously or unconsciously) as such and only distracted the player while they were trying to learn the procedure. However there's a slight possibility that the guidance during the procedure was in fact picked up, but that the system guiding the selection of relevant over irrelevant information prohibited internalization of the procedure, as in Van Nimwegen and Van Oostendorp (2009). Additional research should be performed where relevant auditory cues are compared with distracting or contradictory auditory cues to determine whether poorer performance is due to distraction or a failure to internalize the learning material.

Visual cueing

If the intention of the auditory cues was too obscure and therefore simply distracting, leading to poorer learning of the triage procedure, one would think that the more explicit visual cues would fare better. Although visual cueing led to significantly better scores on the knowledge and the structural knowledge assessment than with auditory cues, there was no difference with the control group. After the experiment was finished, we informally asked some of the participants in the visual cueing condition whether they had noticed the blinking green arrows and understood their purpose. While most recalled seeing the arrows, not many seemed to have paid much attention to them. A few said that they thought the arrow indicated where the victim was injured. Even though this wasn't the intended meaning of the cue, it could still make the cue relevant; if e.g. an arrow pointing at the victim's legs were to indicate an injury, the action in our game that is associated to this the most would be checking the mobility.

This is however contingent on the player of a serious game making that association, which, considering the failure to reject the null hypothesis with regard to the control group, is unlikely to have happened frequently. That it is contingent on the player making the association, could be corroborated by the finding that in the visual cueing condition, performance with respect to the in-game score and post-game learning gains was significantly better for participants with prior game experience as opposed to non-gamers. We think it is plausible that the participants who play games had a better understanding of the meaning of the glowing arrow, because of the ubiquitousness of salient and unrealistic visual cues in contemporary videogames. This in fact mirrors one of the findings of Steiner and Voruganti (2004) when they cued places of interest in a virtual environment; participants that frequently played games were significantly more likely to visit a cued place than those that didn't.

The same effect of game experience curiously did not hold for the structural knowledge assessment, possibly indicating that visual cues will only improve superficial processing of the information; this would require more research however. In section 5.2 we hypothesized that visual cues, which are often deliberately out of place in a certain setting to draw the player's attention, would break the immersion, but no difference in engagement ratings were found between the conditions.

Summarizing, while cues may sometimes be necessary to make a game actually playable, for instance in order to distinguish between friend and foe, we found no evidence that it can be used to improve the selection process of learning. This could however be due to the participants not understanding what the cues meant, which is supported by the fact that participants with game experience seemed to understand them better than those without. Therefore, if one wants to use cues, or perform a follow-up experiment, special care should be taken to apprise the player of the meaning of the cues; they can, however, ostensibly be made highly salient without breaking the immersion of the gamer.

Discussion of the experimental setup

Three threats to the validity of the results can be discerned. Firstly, the small number of participants; twelve to fourteen participants per group may be too little to accurately extricate the effects of such a subtle cueing mechanism. With more participants, some of the inconclusive results, for example the number of times the procedure flowchart was consulted, could become clearer. It is similarly unclear how cueing would influence the in-game score, mental model construction, knowledge tests over longer or shorter time periods. Thirdly, the procedure help button was used a lot more than anticipated; in the control and auditory cueing condition this was even a significant predictor for in-game score. Even though we found no difference in the amount of times the help function was consulted, there is a slim but theoretical possibility that in these two conditions the help button served as the equivalent of the visual cue to learning the material. In addition, with the correct procedure briefly shown before the game and on-demand in the game, one can wonder to what extent learning of the procedure occurs in the dynamic and high cognitive load imposing game environment in the first place, and to what extent from a static picture, rendering the use of cues relatively moot.

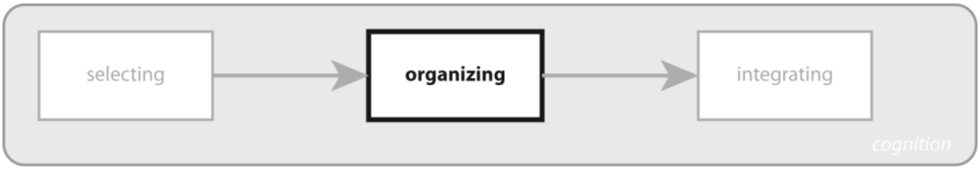
5.9 New version of Code Red Triage

As we mentioned in the previous section, the way Code Red Triage and the experiment was set up until now (except, to complicate matters, in section 4.3, as that experiment was performed chronologically after this one), i.e. giving the procedure beforehand and on-demand, it became unclear exactly what was learned from the triage menu in the game, and what from looking at the flowchart of the triage procedure. In theory, a person with an eidetic memory could have excelled on the posttest, purely by remembering the triage procedure flowchart that is presented beforehand, without even playing the game seriously.

In subsequent experiments we therefore removed the flowchart of the procedure from the entirety of the experiment and changed Code Red Triage to incorporate all the information that is needed to discern the correct triage procedure. Clicking on a check in the triage menu in-game now not only gives the response from the victim to the check, but this is preceded by a few lines that describe what the check or action is about, in which case it is performed and, deliberately vaguely, around which moment in the procedure it is used.

Chapter 6a

Progressive versus variable complexity presentation



Abstract—Contemporary videogames usually progress from simple to more complex, both in terms of the number of options a player has at his disposal, as well as the problems he faces. This game design paradigm may be beneficial for the information organization process in multimedia learning, as it takes into account the limited working memory capacity of the learner, as well as provides a form of signaling of the relevant conceptual links and how they should be connected. However it has never been tested empirically and there may be reasons why other approaches could be better in the case of serious game design. We therefore tested whether the progressive introduction of new options was better than when all the options were available from the start, as well as whether a progressive presentation of problem complexity was better than a variable complexity presentation, when it came to learning and engagement (N = 56). No significant differences in learning gains were found, but having all the options from the start in conjunction with a progressive complexity presentation of problems was considered significantly more enjoyable than other conditions. This is likely due to a greater feeling of autonomy and competence in the player.

6.1 Introduction

For many gamers, the first introduction to the medium of videogames was when they played *Super Mario Bros.* for the Nintendo Entertainment System (NES), a game that held the record for best-selling videogame for over twenty years (Iwata, 2009a). For us, the opening level of the game, where in World 1-1 the eponymous character (Super) Mario encounters a low overhead row of bricks and an advancing enemy called a Goomba, is so iconic that it is one of the first images that come to mind when someone even mentions videogames. It wasn't until recently that the main designer (the videogame equivalent of a movie director) of the game, Shigeru Miyamoto, confessed that this opening setup wasn't coincidental, but created deliberately in this manner (Iwata, 2009b). A novice player with a control pad in hand will try out all the buttons and discover Mario is able to walk left and right and jump. Walking right, the player notices the advancing Goomba and subsequently has two options: walk into the Goomba or try to jump over it. If the player tries the first option, he (or she) will die and will have to start over; if he tries the second, Mario will bump into the low row of bricks and get diverted onto the Goomba, immediately flattening it. By this deliberate level layout, the player will have unconsciously learned, in just a matter of seconds, the central concept around which the game revolves: that stomping on enemies takes them out. In fact, World 1-1 was created last, specifically to teach the novice player most of the key concepts needed later on in the game (Iwata, 2009b).

Games revolve around learning. Learning the actions you can do in a game and learning how and when to do these to resolve problems. Much of what makes a game fun in the first place is learning new things, achieving mastery and the corresponding boost in self-efficacy and feelings of competence (cf. Klimmt & Hartmann, 2006; Przybylski, Rigby & Ryan, 2010). In fact, because games are often long and complex and about learning yet still have to be fun, game designers have become experts of instructional design (Gee, 2005). Gee's argument roughly proceeds as follows: gamers like difficult games, but if the game can't be learned well, it won't sell, whereas games which can, do. In addition, game developers copy what worked in other successful games and thus game design evolves via natural selection into games that have increasingly better instructional design (Gee, 2003). We think this argument is compelling, but it may come as a surprise that there is very little empirical research to substantiate it.

6.2 Organization of information

As we discussed in chapter 2, one of the main activities in learning from multimedia is organizing the incoming information into correct, meaningful and coherent propositions or cognitive structures. For instance, a learner experiences a number of ostensibly separate events, sounds or images, and subsequently has to link these together in a causal manner, or place them in the correct sequential order. In our game, the player can choose from eight different steps in the triage procedure, and has to mentally determine and organize these steps into the order in which they should be taken.

The order in which information is presented will likely influence the way in which the player naturally, at least at first, orders the information. But the rate and manner in which the new information is presented can also influence the organization efficacy of someone engaged in multimedia learning. As we will discuss next, a slower steady rate of information presentation may free up working memory load so that more capacity can be used for the organization process, whereas a quicker or more variable rate of presenting new information may make the incoming information more complex to process, but, perhaps surprisingly, can also lead to a better mental representation and better recall afterwards. As good instructional design is closely related to good game design, this dilemma is related to the game design question of what (we will henceforward call) is the most beneficial 'complexity progression'.

6.3 Complexity progression

6.3.1 Complexity progression in entertainment games

As an example of a central tenet in game design that is used in nearly every game, we refer back to Super Mario Bros. As said previously, at the beginning of the game the player is able to walk left, right, jump and stomp on enemies. Shortly afterwards, Mario will encounter a mushroom, turning him into Super Mario. In this form Mario gains an extra hit point, as well as being able to crash through blocks. Later on in the game Mario will furthermore gain the ability to throw fireballs and temporarily become invincible (and in subsequent games the ability to fly, throw hammers, etc.).

Similarly, the first enemy the player encounters in the game, a Goomba, can only walk left and right and has no attacking ability other than to walk into the player. Further on into the game the player encounters an enemy that can also attack by means of shooting fireballs (the Piranha Plant), then an enemy where something more complex occurs when you stomp on them (the Koopa Troopa) and even later we see the Goomba and Koopa Troopa recurring but with the added ability to fly, thereby increasing their degrees of freedom in movement. Near the end of the game, enemies will often have a number of the previous abilities combined to create all new challenges.

While it is almost never a hundred percent linear affair, grosso modo the progression of complexity in a game proceeds thusly; that the game starts out easy, with the player character only possessing a few abilities and the enemies having simple to deduce characteristics and attack patterns, and the game then becoming gradually more complex as new abilities are added and enemies with extra attack patterns and characteristics encountered. For many other games, most notably those in the genres of what has been termed Legend of

Zelda or Metroidvania (a portmanteau of the names of the videogame series Metroid and Castlevania, who both have similar structures in level design) style games, the newly obtained abilities of the main character are also paramount in defeating the more complex enemies.

That the complexity of the game should gradually increase as the player progresses, based on the assumption that the player simultaneously improves in skill, is one of the main guidelines to create, or keep the player in, a sense of Flow (cf. Chen, 2007; Sweetser & Wyeth, 2005), the pleasurable state during which the player is fully immersed in the game (Csikszentmihalyi, 1990). Gee (2005) furthermore states a number of reasons why this is not only good from a game design standpoint, but also from an instructional design standpoint. First and foremost, because starting small, practicing and automatizing the new information and then moving on to a more difficult puzzle or enemy where new knowledge or skills are added to what the player has learned previously, is what Gee calls a cycle of expertise and the basis with which all expertise is produced. Secondly, because most games incorporate a so-called fish tank or sandbox structure in the tutorial levels, where players are scaffolded and encouraged to form hypotheses and try them out without danger or going too far astray. And thirdly, the player receiving new information, such as a new ability or a hint, just before having to use it to overcome the next puzzle or defeat a 'boss' character, is an example of just-in-time information (Gee, 2005).

While probably true, this design paradigm isn't decisive or all-encompassing. Firstly, it has hardly been tested scientifically in the context of games. It is furthermore plausible that this is the best design paradigm for entertainment games, where the player has to be eased into complex material, learning this without getting stuck. For serious games however, the game experience itself is less important than ultimately reaching the learning goal after gameplay, and there are a few reasons why such simple-to-complex sequencing may not be the best option.

6.3.2 Option complexity: Just-in-time versus just-in-case information presentation

The previous example of the player character gaining new abilities later on in the game to defeat enemies could be seen as the *just-in-time* information presentation of a player's *options*. Just-in-time information presentation means that new information is only presented in a learning environment at the time that it is first needed. In a game setting, the player will not e.g. receive information on how to tackle a problem until this problem is first encountered. This can be contrasted to *just-in-case* information presentation, where the player of a game is given information on how to tackle a problem early on, but only encounters the corresponding problem at a much later time. Just-in-time presentation is said to be better from an instructional design standpoint, because the learner does not have to retain information in working memory that is irrelevant at that time, as is the case with just-in-case information presentation, thus freeing up working memory for more generative processing. This has been shown to improve learning in an experiment by Kester, Kirschner, Van Merrienboer and Baumer (2001).

Just-in-time information presentation can also be a way of guiding the player that has been heavily utilized in games. For instance in the game Legend of Zelda Ocarina of Time, the player frequently has to enter dungeons; enclosed levels of sorts with a 'boss' enemy at the end. In every dungeon the player will encounter a new item, such as a bomb, that is immediately used to solve puzzles (e.g. blow up a rock to proceed) and is key in defeating the boss (e.g. throw the bomb into the boss' mouth). Upon receiving a new item, the more experienced player is thereby cued that this is important in solving subsequent puzzles.

Not having to think about which item from the inventory (i.e. option) to use to defeat the boss because the obvious option was just handed to the player, will likely free up working memory as well as make the organization process easier, because the player is scaffolded via a form of signaling into mentally linking the correct player option to an encountered problem. Therefore it is likely that just-in-time information leads to better information organization.

Conversely, however, it could also be seen as externalization of essential task information when the player is trying to discover which options to use to defeat the boss. That is, if a player is handed the right solution to a problem, he or she may solve that problem, but subsequent organization of information does not receive much conscious thought, leading to less robust models of the information. Externalization has been shown to lead to more superficial learning and less subsequent planning behavior as opposed to when the learner has to discover the complete correct solution himself, with an experiment by Van Nimwegen, Burgos, Van Oostendorp and Schijf (2006).

Presenting options just-in-time for puzzle solving in a game would mitigate the player getting stuck in the game, which could be the aim of entertainment game design, but could potentially lead to less actual comprehension of the underlying complex problem, which would be the ultimate aim of serious games design. Therefore, while there is a good argument to be made for just-in-time presentation, an argument could also be made that a just-in-case design, where the player has both relevant and irrelevant information bits in his or her working memory and has to choose and connect the relevant information bits needed to solve a problem, could actually be a better alternative for serious games.

6.3.3 Problem complexity: progressive versus variable increase in the complexity of presented problems (massed vs spaced learning)

A similar ostensible disparity between the goals of entertainment and serious game design may emerge from the progression from easy to more complex problems itself. In section 6.3.1 we already cited Gee that consecutive small steps where new information is first organized and automatized before the player proceeds to more complex problems, is a good way to gain expertise. This is corroborated by research, e.g. by Moreno and Mayer (2000), who found that for correct organization of information to occur, the amount of instruction should not exceed working memory capacity. Partitioning the information into smaller chunks was subsequently found to lead to better learning (Moreno & Mayer, 2000).

However, here too there may be a reason to doubt whether it is always the superior choice for serious game design. In the case of a game having a gradual complexity increase, the player is likely to encounter more or less the same class of enemies or obstacles a number of times before progressing on to more challenging obstacles. Encountering similar obstacles in succession in learning is called a *massed* presentation approach. The opposite, where obstacle A is presented first, then a dissimilar obstacle B and a dissimilar obstacle C before returning to obstacle A again, is called a *spaced* approach to information presentation.

In an extensive review carried out by Cepeda, Pashler, Vul, Wixted and Rohrer (2006), of the 271 experiments that contrasted massed versus spaced presentations, 259 showed improved learning in the spaced presentation condition. The exact reason for this still seems somewhat unclear; according to Cepeda and colleagues (2006) one likely, and ostensibly the most prevailing, explanation is that the learning of another task in between two spaced instances of a task results in a slightly different context for each time that same task is learned. If the task is encoded into memory with more variable contexts, it will be accompanied by more retrieval cues and thereby become more robust to recall later on, when the context may also be different (Russo, Mammarella & Avons, 2002). However this has been called into question by Dempster (1987) who, over a number of different experiments, also found clear spacing effects, but no effects of encoding variability in the context of what was to be learned. Similar problems with encoding variability as an explanation were put forward by Appleton-Knapp, Bjork and Wickens (2005).

Whatever the underlying explanation, when a serious game requires learning multiple complex problems, encountering these in a spaced instead of a sequential (massed) order may lead to better learning. The main problem with this assertion, however, is that nearly all of the experiments were performed in the context of text memorization. How either massed or spaced presentation approaches fare in the context of learning problem-solving skills is less obvious.

All in all, as we have shown above, even though there are good arguments that using a simple-to-complex order in terms of a player's options and problems encountered leads to better information organization, one can also propose arguments against this notion. In addition, it is unclear how such presentation strategies affect the engagement of a serious game. Consequently, we set out to determine the way in which the game's complex information should be presented, both in terms of problems that need to be solved and the options the player has at his disposal, in order to improve learning and engagement. This leads to an experiment with two variables, problem complexity and option complexity, resulting in four different conditions: (1) just-in-time options and massed problem cases, (2) just-in-case options and massed problem cases, (3) just-in-time options and spaced problem cases, and (4) just-in-case options and spaced problem cases.

6.4 Experiment

6.4.1 Participants

A total of 56 people participated in the experiment, 14 for each of the four conditions. Thirty two participants were male and 24 female, all of them of university level, with an average age of 23.0 (*SD* = 3.5). Of all the participants, 8 reported to play videogames 'rarely to never', 30 occasionally played games, while 18 considered themselves a gamer.

6.4.2 Experimental setup

The test bed for our experiment is a serious game for triage training, called Code Red Triage (cf. additionally chapter 3), which is a total conversion mod of the three-dimensional first person game Half Life 2 Episode 2. Here, the player is a medical first responder that arrives at a train station showing signs of recent panic, where he (or she) receives a call that a terrorist strike has taken place on a subway platform of the station. The player has to navigate to the subway platform and, reaching that, perform a triage, a categorization based on urgency of needed medical attention, on all of the victims that are present.

A mobility (sieve) triage roughly comprises the following steps: (1) check whether the victim is still able to walk, if so the victim is lightly injured, or falls into triage category T3. If not, (2) check whether the airway is obstructed, and if it is, free it; depending on the outcome the victim is either already dead, or has to be placed in the stable recovery position and is in need of immediate medical attention: T1. If the airway was free from the start, (3) check the breathing frequency, if this is too high the victim is in need of immediate medical attention or T1, if the frequency is within bounds, then (4) check the capillary refill time (CRT). If the capillaries take too long to refill, the victim is seriously injured and should be placed in category T1, if this is within bounds, the victim is moderately injured and should be placed in category T2. Sometimes it is too dark to accurately gauge the capillary refill time, in this case additionally (5) the pulse rate should be checked.

Our version of the triage procedure is slightly simplified from the actual one; the player doesn't have to learn the exact numbers pertaining to respiratory or pulse rates, nor know about special cases such as cold weather or children of varying ages. This was done to keep it learnable for novices in a short time period. The participants of the experiment receive almost no information on the triage procedure beforehand, but have to infer the correct procedure by pressing on the buttons in the triage menu and reading the corresponding explanatory text, and see how well they score after placing a victim in a triage category (see figure 6.1). We did not include any prior training of the triage procedure but instead incorporated all the instruction in our game because our experiment is essentially about learning in a game and the corresponding difficulty curve. It is therefore different from a previous version of our game (cf. chapter 5 and Van der Spek, Wouters, Van Oostendorp & Aarnoudse, 2010), and another triage training game by Knight and colleagues (2010), which focused more on practicing the triage procedure.

The complete procedure can be seen represented as a flowchart in figure 6.2. Added to this diagram in light purple is how we ranked the complexity of the different victims. We categorized the victims a priori based on their triage complexity, i.e. the number of steps that are needed to reach the correct diagnosis. In total there were 19 victim cases, four of which had a complexity of 1, five victims were of complexity 3, six (four in one branch, two in the other) of complexity 4 and a remaining four of complexity 5.

Figure 6.1 - Feedback after categorizing a victim in triage category T3, feedback was similar for all conditions

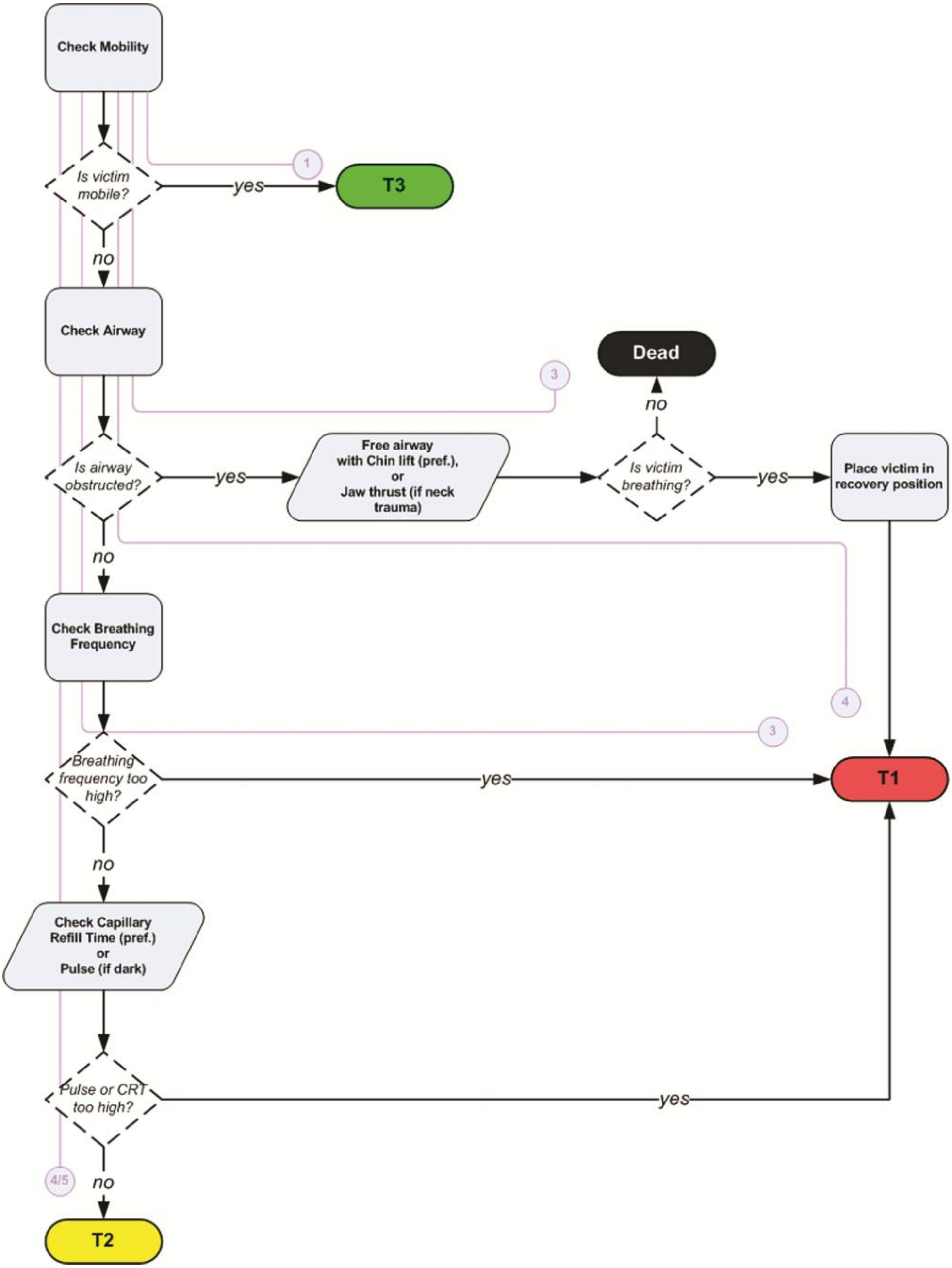


Figure 6.2 - Triage procedure with complexity classes

6.4.3 Experimental conditions

The experiment had the form of a 2x2 design, with the variables problem and option complexity. Example screenshots of which can be seen in figure 6.3. Upon approaching a victim, the participant had to press the E-button to start the triage procedure. An overlay menu then appeared, with buttons to the side that could be used to perform the different checks of the triage procedure (the blue parallelograms in figure 6.2), and the triage category buttons at the bottom.

In section 6.3.1 we discussed how games progressively add options to the player's arsenal in order for the player to tackle progressively more complex problems. In our game, we define the option complexity as the number of procedure buttons that the player can use to check the wellbeing of the victim. We define the problem complexity as the complexity of the victims that need to be triaged.

Just-in-time versus just-in-case procedure buttons

In the just-in-time (JiT) condition, the participant started out with only two buttons for the first victim. New buttons were then gradually added until the fourteenth (of nineteen) victim, when all eight buttons were available. Whenever a button was introduced, it would quickly thereafter be needed to correctly triage a victim. In the just-in-case (JiC) condition, all of the buttons were available from the start.

Progressive complexity (massed) versus high complexity variability (spaced) presentation of victims

In the progressive simple-to-complex complexity condition (massed presentation), the participant first encountered the simple victims, where a single mobility check was enough to determine that they were only lightly injured. The complexity of the victims (see experimental setup and figure 6.2) then progressed stepwise to higher complexity victims, where in the end the participant has to perform five checks to come to the correct victim resolution. This was contrasted with the high complexity variability (or spaced presentation) condition, where the victims had been placed in such a way that the victims of a certain complexity level were spaced apart as far as possible (as could be accommodated by the just-in-time presentation of procedure buttons).

The four conditions are then JiT-Massed, JiT-Spaced, JiC-Massed and JiC-Spaced. The JiT-Massed condition, where both the user interface as well as the victims progress stepwise from simple to complex, seems most in line with the design of contemporary COTS videogames.

Just-in-time presentation (JiT)

Just-in-case presentation (JiC)

Simple-to-complex (massed) victims presentation

High complexity variability (spaced) victims presentation

Figure 6.3 - Experimental conditions

6.4.4 Materials

During this experiment, we tested how the different design paradigms would affect learning, deep comprehension and engagement.

Learning

Learning was divided into three variables. The first measured knowledge pertaining to the triage procedure via a set of eight multiple choice verbal questions in the form of e.g. ‘an accident has happened on a provincial highway, you have already determined that the victim has an unobstructed airway, what is the next step you should take according to the procedure of the primary triage?’ and another eight similar questions, but with a screenshot of a mockup of the game and the possible answers in the form of buttons that can be ticked. Pictorial questions were added because they may be a better reflection of the triage procedure as it is learned in the game (chapter 4). This knowledge test was changed slightly from chapters 4 and 5, where we operated with a distinction between a verbal declarative knowledge or retention test and an application test. We added the pictures to the pretest because it was found that this correlated highly with what the player did in-game and therefore was deemed a relevant measure (see chapter 5). The difference between declarative and procedural knowledge additionally seemed somewhat artificial when it comes to the triage procedure, as we explained in 4.1.1; we therefore combined the two in a single test that measures the participant's knowledge pertaining the triage, both as a procedure and a skill. See Appendix A for the handout with the tests presented to the participants.

The second measure was an assessment of how the information on the triage procedure was stored structurally in the participant's mind, as a measure for deep comprehension. The structural knowledge assessment is performed by having the participant rate concepts pertaining to the triage on their degree of relatedness, which the computer program Pathfinder then transforms into a network, with closely related concepts represented as directly linked nodes and less related concepts that are further away to each other in the network (cf. Trumpower, Shahara & Goldsmith, 2010). Networks can be compared to that of an averaged referent network of advanced learners (Acton, Johnson & Goldsmith, 1994), the resultant similarity score before and after playing the game then giving an indication for the amount of deep learning that has occurred (Ferstl & Kintsch, 1999). We used the same concepts as in section 4.3 and chapter 5, and consequently also the same referent mental model. For an image see figure 5.3.

Thirdly, and slightly separately, we looked at the in-game score as a measure for learning performance. The score was shown to the player for every victim they triaged to not only aid in their learning, but simultaneously also logged to a text-file. The in-game score is based on whether the triage was correct or not, and points are deducted if reaching a conclusion took longer than a preset amount of time, if necessary steps were forgotten and/or unnecessary steps taken, and finally if the steps were done out of order. Figure 6.1 shows how this is fed back to the player.

Engagement

Because the tests for learning are quite extensive (taking on average twelve minutes before and after playing the game), we kept the engagement questionnaire relatively short. We asked the participants to rate how much they enjoyed playing the game on a scale of 1 to 10, and separately to rate how difficult it had been in their experience to get a good score in the game on a scale of 1 to 10 as well.

In addition, we were interested to find out whether the different game designs would affect engagement in terms of the player feeling immersed in the game. To measure this, we used the engagement subquestionnaire of the Independent Television Commission's Sense of Presence Inventory (ITC-SOPI; Lessiter, Freeman, Keogh & Davidoff 2001), which measures engagement in relation to immersiveness via twelve questions, such as 'I would have liked the experience to continue' and 'I had a sense that I had returned from a journey', on five point Likert-scales.

6.4.5 Procedure and apparatus

The experiment started with a demographics questionnaire, where the participant had to rate their game experience (1 – I hardly ever play games, 2 – I occasionally play games, 3 – I'm a gamer). In addition, it had a pre-posttest design when it came to the knowledge tests, to compensate for prior knowledge. The order in which the experiment was performed can be seen in figure 6.4. The experiment took about 40 to 55 minutes per participant, of which in between 12 and 20 minutes were spent playing the game, depending on how long it took the player to navigate and find all the victims.

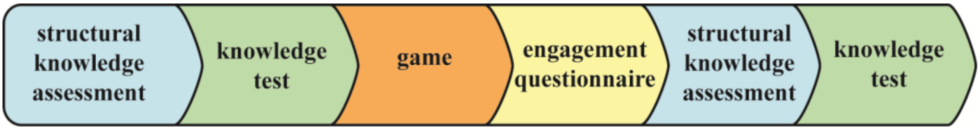


Figure 6.4 – procedure

The experiment was conducted on a single 17" gaming laptop, with all of the graphics settings on maximum. Participants were placed behind a large blue cubicle partitioning screen, and during the game they wore a headset and the lights were dimmed.

6.5 Results

Learning

The means and standard deviations for the knowledge test (scores ranging from 0-16) and structural knowledge assessment can be found in table 6.1 and table 6.2 respectively. Beforehand, we hypothesized that the progressive complexity of options and problems would probably lead to better learning.

However, a full factorial ANCOVA with knowledge posttest as dependent variable, knowledge pretest as covariate and option presentation (JiT vs. JiC) and problem presentation (Massed vs. Space) as factors, shows no significant effects: option presentation main effect $F(1,51) = 2.10$, $p = 0.15$, problem presentation main effect $F(1,51) < 1$ and interaction effect $F(1,51) < 1$.

For the structural knowledge assessment, a full factorial ANCOVA with structural knowledge assessment posttest as dependent variable, pretest as covariate and option presentation and problem presentation as factors also shows no significant effects, both main effects $F(1,51) < 1$ and the interaction effect $F(1,51) = 1.23$, $p = 0.27$, $\eta^2 = 0.02$.

condition	Massed		Spaced	
	pretest	posttest	pretest	posttest
Just-in-time	5.07 (1.77)	8.64 (2.02)	4.62 (1.50)	8.54 (1.76)
Just-in-case	4.86 (1.79)	7.43 (2.38)	5.07 (1.73)	7.64 (2.47)

Table 6.1 – Mean scores (s.d.) on knowledge test

condition	Massed		Spaced	
	pretest	posttest	pretest	posttest
Just-in-time	0.043 (0.075)	0.222 (0.137)	0.029 (0.085)	0.173 (0.139)
Just-in-case	0.061 (0.099)	0.205 (0.156)	0.078 (0.071)	0.249 (0.137)

Table 6.2 – Mean scores (s.d.) on structural knowledge assessment

In-game score
If we look at the sequence of successive scores per victim in the game, a Repeated Measures ANOVA shows a significant between-subjects effect of option presentation $F(1,49) = 4.87$, $p < 0.05$, partial $\eta^2 = 0.09$, in favor of the just-in-time presentation. However, this could be expected; in over half of the victim cases the participants in this just-in-time condition had less buttons to choose from, and therefore less chance of getting penalty points deducted. Perhaps a more relevant insight arises when we look at the graphs of the scores per victim (figure 6.5).

One can clearly see the participants in the just-in-time group score better at the first few victims, when the possibility to err is much smaller. However as the number of procedure buttons (i.e. options) start to increase, this advantage disappears. After the thirteenth victim, differences between option presentation conditions have all but disappeared (before: $F(1,52) = 13.65$, $p < 0.001$, partial $\eta^2 = 0.21$; after: $F(1,49) < 1$). In addition, the difference between the problem presentation conditions is also not significant.

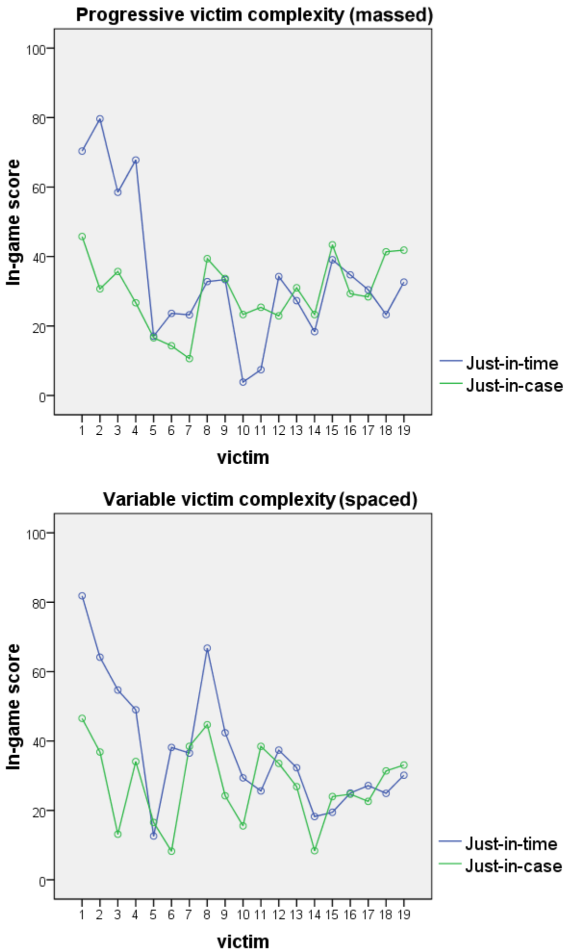


Figure 6.5 - In-game score progression per condition

Engagement

Table 6.3 shows the means and standard deviations for the different conditions on the engagement tests. First the self-report experienced enjoyment and difficulty ratings and lastly the ITC-SOPI engagement subquestionnaire.

	condition	Massed	Spaced
Enjoyment	JiT	6.57 (1.70)	7.07 (0.73)
	JiC	8.00 (1.47)	7.07 (1.14)
Difficulty	JiT	6.93 (2.24)	8.00 (0.96)
	JiC	7.93 (0.92)	7.79 (1.53)
ITC-SOPI	JiT	3.46 (0.53)	3.67 (0.39)
	JiC	3.85 (0.49)	3.60 (0.32)

Table 6.3 – Means (s.d.) pertaining to the affective questionnaires

Engagement as a consequence of a certain game design paradigm may be related to prior expectations and thereby how often the participant plays games at home, we therefore included prior game experience as a covariate. Performing a full factorial ANCOVA with enjoyment as dependent variable, prior game experience as covariate and option presentation and problem presentation as factors, the enjoyment rating of the game shows a significant difference for option presentation main effect: $F(1,51) = 4.19$, $p < 0.05$, $\eta^2 = 0.08$, and a significant difference for interaction effect $F(1,51) = 4.78$, $p < 0.05$, $\eta^2 = 0.08$. Problem presentation did not influence enjoyment, $F(1,51) < 1$. The plot pertaining to the enjoyment rating can be seen in figure 6.6. Here, it is clear to see that the just-in-case option presentation is preferred over the just-in-time presentation, but only when the victims (i.e. problems) are presented in a progressive, massed manner.

Difficulty was not significantly affected. Option presentation main effect: $F(1,51) < 1$; problem presentation main effect: $F(1,51) = 1.33$, $p = 0.26$, $\eta^2 = 0.02$; interaction effect: $F(1,51) = 2.27$, $p = 0.14$, $\eta^2 = 0.04$.

The ITC-SOPI scale was reliable with a Cronbach's alpha coefficient of 0.74. However, even though the resulting plot (figure 6.7) is highly similar to the enjoyment rating, a full factorial ANCOVA with the averaged ITC-SOPI score as dependent variable, option presentation and problem presentation as independent variables and prior game experience as covariate, does not show a significant difference for main effects, $F(1,51) = 2.1$, $p = 0.16$, $\eta^2 = 0.04$ and $F(1,51) < 1$ respectively, nor for interaction effect, $F(1,51) = 3.45$, $p = 0.07$, $\eta^2 = 0.06$. However, given the medium effect size and the similarity with the enjoyment rating, we contend that failing to reject the null hypothesis for the interaction effect is likely due to the small sample size.

If the spaced problem presentation condition is neglected, and an ANCOVA is performed on only the ITC-SOPI score in the massed presentation condition as dependent variable, option presentation as independent variable and prior game experience as covariate, then option presentation is significant, $F(1,25) = 4.58$, $p < 0.05$, $\eta^2 = 0.15$

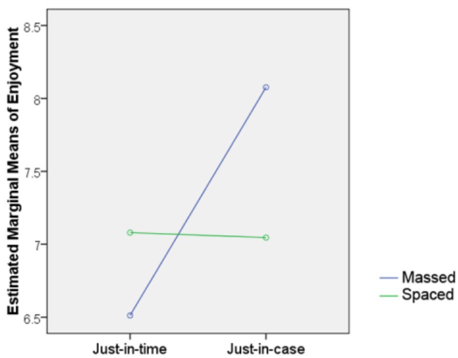


Figure 6.6 - enjoyment rating

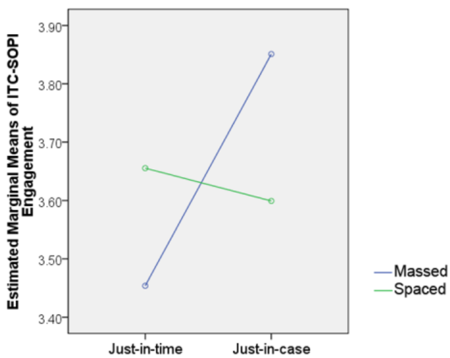


Figure 6.7 - ITC-SOPI

Validity of the measures
As a posteriori evidence for the validity of our measurements, both the structural knowledge assessment test and the knowledge test after the game were significantly correlated with the in-game score, $r(56) = 0.29$, $p < 0.05$; $r(56) = 0.48$, $p < 0.001$ respectively. The structural knowledge assessment was correlated with the pictorial questions of the knowledge test, but not the verbal questions, $r(56) = 0.36$, $p < 0.01$; $r(56) = 0.23$, $p = 0.08$ respectively.

The single item enjoyment of the game rating was significantly correlated with the ITC-SOPI questionnaire $r(56) = 0.41$, $p < 0.01$. In addition, the single item rating of how difficult participants found the game was significantly negatively correlated with the in-game score $r(56) = -0.32$, $p < 0.05$, as well as with the knowledge test $r(56) = -0.27$, $p < 0.05$.

6.6 Conclusion and discussion

We set out to test the assumption that entertainment game design, and specifically the gradual progression from simple to more complex gameplay elements, is also good serious game design, because it aids in the correct organization of information. In entertainment games, this is often in the form of the player character receiving more abilities (the player’s options) over time, and gradually encountering enemies that have more attack patterns (more complex problems), both resulting in more complex interaction possibilities that are taught to the player just in time. In our triage training game, this is analogous to the JiT-Massed condition, or a stepwise increase in the number of procedure buttons at the player’s disposal, with more checks available to determine the severity of the injuries, and a corresponding stepwise increase in the complexity of victims, with more checks needed to determine the triage category.

Perhaps surprisingly, when we contrasted the just-in-time presentation of procedure buttons to a just-in-case approach, and a stepwise massed presentation of the victims to a spaced presentation approach, we found no evidence to support the previous notion with our game. Learning seems to occur in equal amounts regardless of the sequence or rate in which information is presented, as no significant difference was found between the four conditions. This is corroborated further by the in-game scores converging as the player encounters more and more victims. Therefore we can say that simple-to-complex sequencing as is prevalent in entertainment game design, does not improve learning of complex tasks over other complexity presentation design paradigms. There are a number of possible explanations for this apparent lack of effect of complexity presentation on learning gains. It could be that the way complexity is presented in a game is genuinely inconsequential to the mental organization of information.

It may also be possible that the pros and cons pertaining to the different presentation forms as detailed in the Complexity Progression section (section 6.3.3) even each other out. That for instance a massed, progressive complexity presentation of the victims leads to better initial organization of the information pertaining to triaging the victim, but that the spaced presentation of victim cases makes the organization of the information into a mental model more robust for different contexts. Both paradigms could subsequently have been advantageous for the information organization process for different reasons, and the knowledge test fails to discriminate between these.

What we did find was that the participants significantly preferred the version with all the procedure buttons available from the start, at least when it is combined with a stepwise, massed increase in victim complexity presentation. A similar but weaker effect was found on the immersiveness of the game. This would imply that just-in-time information in a serious game, at least when it is used in a simple-to-complex problem presentation, does not lead to better learning and is also less fun than giving the information beforehand.

This can probably be explained by Przybylski and colleagues' research in the motivational pull of videogames (2010), where they found that the feeling of autonomy was the strongest predictor of game enjoyment. Giving players all the options from the start could grant them the illusion that they have control over and may choose how to tackle the problem from number of different ways, even though there is ultimately only one correct way in our game. In addition, the combined preference for a progressive increase in victim complexity may point at a subtle interplay between feelings of autonomy and feelings of competence, which could be heightened if the player doesn't encounter bumps in the learning curve. In another sense, it could also mean that players want to be 'helped' by a progressive increase in the game difficulty, but not overtly so, i.e. when it is combined with an increase in the complexity afforded by the number of abilities.

Some remarks should however be made about the generalizability of these findings. Our game is relatively short, on average lasting approximately fifteen minutes, which it uses to only teach the complex triage procedure, with a relatively high number (nineteen) of cases. The progression of in-game scores showed a convergence over longer time periods, which could indicate a ceiling effect of the different information presentation forms. Therefore, an inability to reject the null hypothesis when it comes to learning may be due to the operationalization of our experiment, in this case the player being immersed in our serious game long enough for a certain amount of learning to occur regardless of information presentation form. Furthermore, while our game revolves solely on learning the triage procedure, in some cases a game designer may want to provide gameplay elements that are simple, intermittent, or inconsequential to the main objectives of the game, in which case just-in-time information may be fully justified, as it does result in the player scoring better initially.

In addition, it is unclear how enjoyment will be affected over longer time periods. Game designer Adams (2006) stressed the importance of introducing new features throughout the game instead of only at the beginning, lest the latter part of the game becomes a grind. As our game is too short to become a grind, introducing new user interface options later on in the game could therefore still improve the enjoyment of the game, as long as it doesn't hinder feelings of autonomy.

6.7 Future work

This was one of the first empirical experiments on information presentation in serious game design to have ever been performed. Many questions still remain. Firstly, the experiment should be replicated in other serious games to rule out operationalization biases, such as the instructional content and time on task. In addition, it is plausible that different game genres and game forms would benefit from different information presentation strategies.

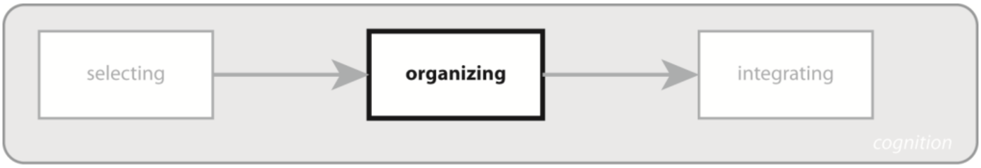
Secondly, in this experiment we contrasted simple to complex sequences to highly complex variants, but other forms exist as well. In the game *Metroid Prime 2*, the player starts the game with the protagonist having all her abilities. The player solves a few puzzles, only to lose the abilities quickly thereafter, and then has to recover these throughout the game. It would be interesting to see how such a backward-chaining like approach to complexity presentation could influence learning.

Thirdly, all of the participants within a certain condition were given the same version of the game, but it could be that different learning styles of the participants would favor different presentation forms. Similarly, a dynamic approach to the information presentation rate, where good players encounter difficult victims earlier as opposed to weaker players, could also lead other enjoyment ratings or learning performance.

Last but not least, these findings could have some impact on the theory of Flow in games and how to engineer this, as the progressive buildup of buttons condition was favored less than the just-in-case approach. Beforehand, starting easy and gradually becoming more complex seemed to be resembling the Flow model the closest, as all the buttons from the start would be too challenging for novices. That participants nonetheless favor the just-in-case condition could be due to an overall greater sense of accomplishment when they finally learn which button to use and when. Therefore, it could be that the Flow graph should start steep and then level off to accommodate a feeling of accomplishment, instead of following a linear path. On the other hand, the in-game score (Figure 6.5) shows that participants in the JiC-Massed condition were the ones that scored most consistently; if in-game score is a correct measure for challenge this would support the linear relation between challenge and skills in Flow theory. The correct path of the optimal Flow state in games is something that needs to be empirically tested.

Chapter 6b

Adapting the complexity presentation to the player's performance



Abstract—Code Red Triage was less efficient than a PowerPoint presentation in engendering learning gains in the player. However, games are continuously assessing the player, and this fact can be used to adapt the complexity of the game to the proficiency level of the player in real time. We performed an experiment with two conditions. In one condition, participants ($N = 14$) played a version of the game that automatically adapted the complexity level of the presented cases based on how well the participant scored previously. Participants ($N = 14$) in the other condition played a control version of the game with no adaptation. Results show that the adapted version of the game is significantly more efficient, and leads to higher learning gains per instructional case, but does not lead to a difference in engagement.

6.8 Adapting the game to the player

We ended chapter 6a by noting that different players may benefit from different speeds at which the victim complexity progresses. Some players may comprehend the instruction pertaining to a set of victim cases quicker than others, making the additional cases redundant, while others do need these additional cases to build a correct mental model of the information. If we look at the in-game score progression of the just-in-case progressive problem complexity condition in chapter 6a in more detail, one can see comparably large standard deviations in the mean scores per victim, indicating that players are in fact at different levels of proficiency at the same moment in the game (figure 6.8).

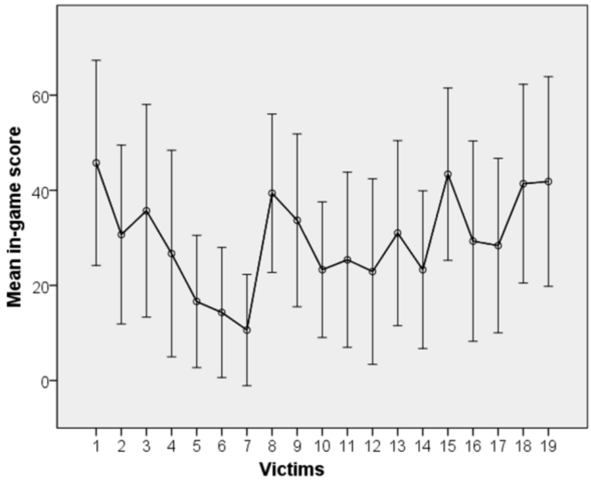


Figure 6.8 – 95% confidence intervals of in-game score per victim

There are a number of reasons why adapting the complexity, and thereby challenge, of the game to the proficiency of the player may be a good idea.

Firstly, as we have already discussed to great extent previously, the working memory of a player can easily be overloaded by the rich multimodal information, as well as the heavier load on selection and organization processes due to the interactive nature, of serious games (see chapters 4 and 5; Kiili, 2005; Moreno & Mayer, 2007). Therefore, slower learners will benefit from a slower pace in information presentation in order to correctly organize all the new information that is coming in. However, correct mental model construction may also be hindered by cognitive *under*load, where the learner is stimulated too little, for instance in the case where a quick learner is confronted with the preset slower paced information presentation. Cognitive underload can lead to (passive) fatigue, which has been shown to result in disengagement from the task and higher distractibility and can subsequently degrade performance (Saxby, Matthews, Hitchcock & Warm, 2007; Paas, Renkl & Sweller, 2004). If a game were to actively prevent the player from becoming cognitively overloaded or underloaded, this may mitigate some of the ostensible problems serious games seem to have in being effective.

Secondly and closely related to this, Csikszentmihalyi (1975) posited that one can experience the feeling of flow, which is a feeling where someone is completely absorbed in an activity to the point of losing self-consciousness while the activity becomes rewarding in its own right, and that this leads to the individual functioning at his or her fullest capacity (Shernoff, Csikszentmihalyi, Schneider & Shernoff, 2003). This is achieved when the provided challenge of an information system is optimally suited to the skills of the user; and as videogames are often stated to be engaging, with players reporting an experience of being completely absorbed in the game, they seem to be ideally suited to produce flow (Sweetser & Wyeth, 2005; Cowley, Charles, Black & Hickey, 2008). Flow has been shown to be positively correlated to learning (Webster, Trevino & Ryan, 1993), therefore, keeping the player in a sense of flow by adjusting the challenge to their skills could improve learning.

Thirdly, in chapter 4 we already noted in our comparison of Code Red Triage with a PowerPoint presentation, that learning the triage procedure in the game condition was less effective than in the PowerPoint presentation even though the participants were given more time to engage in the game. Actual data on the comparable efficiency of other serious games is scarce in the literature; Sitzmann (2011) does note that serious games perform better when learners are allowed unlimited access to the game and when it is embedded in a curriculum. Either way, this could point to an inefficiency of serious games in general, at least when it comes to singularly learning a certain task such as the triage procedure. If quick learners were able to progress in the game at a faster pace, because the game recognizes their proficiency and adapts the game accordingly, this could improve the efficiency of the game.

6.9 Related work

A number of ways to make a game adapt to the proficiency of the player have already been proposed and tested. For instance, Conati and Manske (2009) implemented a pedagogical agent in a game designed to teach children how to factorize numbers. The agent intervened to give advice when the user appeared to be stuck; this intervention was then made adaptive by interpreting the errors the user made, with different levels of user model accuracy. Curiously, this did not lead to better learning, as the students found the new agent to be too obtrusive and got annoyed by its frequent interventions, disregarding the hints that were given. A further discussion on the way adaptation can be implemented in games and the associated challenges can, for instance, be found in Lopes and Bidarra (2011). Here, we will elaborate on two modes of adaptation that are most relevant to our research.

Firstly, one interesting avenue in which a game can be adapted to the player was undertaken by Yun, Shastri, Pavlidis and Deng (2009), who used an infrared camera that was mounted on the TV displaying the game. This camera recorded the faces of the participants while they were playing a game that revolved around shooting enemy robots. Looking at the heat signatures from the supraorbital region of the face, they were able to derive how much apparent stress the game exerted on the player during gameplay. At the same time, the player reported at set intervals whether they found the game too easy, just right or too difficult, and whether they were enjoying the game or would like to quit. This research is relevant to our own for two reasons. One, they discovered that people who found the game too difficult and wanted to quit actually had lower stress levels than when the game was moderately difficult. They argued that this is due to the player becoming disengaged with the game, thereby corroborating the previously made assertion that too high a challenge leads to cognitive overload and is detrimental to the engagement or flow experience. Two, a version of the game where the game automatically adapts to the stress level of the player was shown to lead to higher engagement and better in-game performance (in terms of how many robots were defeated) than in conditions with preset difficulty levels, even for the easy difficulty level.

Another example of how to adapt the game to the player is the entertainment game *The Elder Scrolls 4: Oblivion*. Here, the player roleplays a character in a large and open medieval fantasy world. As the player encounters new locales, performs quests and defeats monsters, his or her character will gradually become stronger and gain better weapons and items (see further Shute, Ventura, Bauer & Zapata-Rivera, 2009). Because the game features an open world for the player to explore freely, this traditionally leads to problems where the player may encounter monsters that are far too strong for his or her avatar to defeat at that point in time. To counter this and provide the optimal experience for everyone, the player's adversaries in the game also progress in power at the same rate as the skill level of the player. Contrarily to what would be expected, many gamers criticized this feature, as it made them feel that their actions were largely inconsequential (Bostan & Ogut, 2009); they were not getting stronger than their enemies and therefore didn't feel like they were mastering the game. As a result, adapting the game to the player too optimally may actually result in less engagement, as we had also witnessed in Conati and Manske (2009).

6.10 An adaptive version Code Red Triage to improve learning efficiency

In the previous section we discussed two different techniques of assessing the player proficiency within the game. The first was a more overt technique, where in real life settings the player would have to install an infrared camera for it to work; the second example featured so-called stealth assessment (Shute et al., 2009). In essence, all games are an assessment device, in that progressing past an obstacle is contingent on acquiring the needed knowledge of how to do so, and the inner game rules rely on a discrete, quantitative operationalization of the problem in order to determine whether the player succeeded. In the case of Code Red Triage, we already have a measure of how well the player is performing in the game, namely the in-game score, which gives an objective measure of whether the player is able to correctly apply the procedure to a given victim case. The player's performance can therefore be seen as an indication of their proficiency level.

We can consequently use this score to adapt the game to the player. If a player scores above a preset threshold, he or she has proven to have a certain level of proficiency and can move on to a more complex victim case. As one can see from figure 6.2, there are six different paths that are taught via the successive victims. Table 6.4 shows these paths in print, with a corresponding tier number, indicating the order in which these are successively taught in the progressive complexity (massed) condition.

While it is technically also possible to introduce adaptivity in the variable complexity (spaced) condition, by deleting all the remaining cases in a certain tier when the player has scored above the threshold for that tier, this could create very different variability encodings in the adaptive group. For example, one participant may consecutively be presented with victim tiers 1-6-1-5, while another participant may have scored high enough at the first tier 1 victim, and encounter victim tiers 1-6-5 consecutively; this could subsequently confound the learning gains. Because of this, we chose the progressive complexity condition for our adaptivity experiment. In addition, deleting successive victims in the just-in-time options condition could lead to differences in the rate with which the options are introduced, which may introduce a second adaptivity variable, something we do not want. As the complexity presentation experiment in chapter 6a furthermore showed little difference in learning gains for the different conditions, we contend that the progressive victim complexity, just-in-case user interface condition is the best condition to compare the efficacy of an adaptive version of the game to a static version.

Tier	Path
1	Mobility
2	Mobility, Airway, Breathing frequency
3	Mobility, Airway, Chin lift
4	Mobility, Airway, Breathing frequency, CRT
5	Mobility, Airway, Jaw thrust, Recovery position
6	Mobility, Airway, Breathing frequency, CRT, pulse

Table 6.4 – The six different paths that have to be taken in order to classify a victim in Code Red Triage

Note from table 6.4 that some of the tiers have equal complexity as defined by the number of steps taken. This order was chosen because it arguably does depict an increase in challenge, as even for the tiers of equal complexity, the first is a continuation of a previously learned path while the other demands taking an alternative path. Regardless of whether this is true, the player has to learn the six paths via the corresponding victim cases. Table 6.5 shows the number of victim cases per tier.

Tier	1	2	3	4	5	6
No of cases	4	3	2	4	2	4

Table 6.5 – Number of victim cases per tier

For the threshold value per tier we chose the (rounded) average score for all the victims in a tier. If a player scores above this threshold value for any given victim, he or she is said to have sufficiently understood the information pertaining to this tier and immediately progresses on to the next one. Contrary to previous versions of the game, where the case information was linked to a physical victim in the game world, the adaptive version of Code Red Triage assigned the case information to a victim dynamically; that is, when a player approaches a victim to perform a triage, the adaptation engine of the game determines which case should be presented together with the victim. This does however lead to the possibility that a game ends because the player has finished the last tier, while he or she still sees victims lying around.

6.11 Experiment

6.11.1 Participants

In total 28 individuals of university-level education, 19 male and 9 female, participated in the experiment, 14 in the adaptive game condition, and 14 in the control condition. Average age was 22.86 with a standard deviation of 5.68. 2 persons hardly ever played games, 16 occasionally played games and 10 considered themselves a gamer. Overall there were too few females and non-gamers to determine any significant underlying correlates.

6.11.2 Experimental conditions

The participants were randomly assigned to one of two conditions, who each received a unique version of the game. One group received the adaptive version of the game, as outlined in section 6.10 above, whereas the other group received a control version of the game, effectively the same version as in the progressive victim complexity, just-in-case information condition of the experiment in chapter 6a.

6.11.3 Hypotheses

We hypothesize that if our game is able to adapt to the proficiency of the player by moving to a higher tier the moment a player has scored sufficiently, this will improve the efficiency thereof. This means that both the knowledge test and the structural knowledge assessment as described in section 6.3.4 will not differ between the two versions of the game, but the participants in the adaptive game condition will have finished the game in a shorter time period.

As a consequence this will also mean that the participant in the adaptive game condition will have learned more efficiently. This will be determined by dividing the post knowledge and structural knowledge scores over the total number of victims that were triaged in the game (and thereby the amount of instructional information received). Another way would be dividing scores over the total time, but reading speed and other factors pertaining to learning style could confound this measure.

The total time a participants spends playing the game is around 17 minutes and this will likely be less in the adaptive condition; therefore, our game is probably too short for the player to get frustrated over an ostensible lack of progression over long time periods as in the example of *The Elder Scrolls: Oblivion*. As such we do hypothesize that the better adjustment of the game challenge to the player's proficiency will lead to a higher engagement in the adaptive game condition.

6.11.4 Materials, apparatus and procedure

The materials, apparatus and the procedure were the same as in the previous experiment, with two minor alterations. Two questions in the knowledge test were changed so that it included at least one verbal and one pictorial question on each tier, and three statements were added to the engagement questionnaire in order to measure a subjective response to the adaptivity: 'at no point did I feel the game was too easy', 'at no point did I feel the game was too difficult' and 'I think I have had sufficient training'.

6.12 Results

Efficiency
The total time in seconds spent triaging in the game until the game over screen, was $M = 589$, $SD = 115$ for the control condition, and $M = 383$, $SD = 107$ for the adaptive condition. The effect of condition on the time spent was significant, $F(1,26) = 24.10$, $p < 0.001$, $d = 1.86$. The total number of victims triaged before the game over screen was displayed, was $M = 18.64$, $SD = 0.75$ in the control condition and $M = 10.71$, $SD = 2.09$ in the adaptive condition. This effect too was significant, $F(1,26) = 178.66$, $p < 0.001$, $d = 5.05$.

	Condition	Pretest	Posttest
Knowledge test	Control	5.57 (2.65)	10.57 (3.48)
	Adaptive	5.86 (1.79)	10.57 (2.44)
Structural knowledge assessment	Control	0.043 (0.098)	0.281 (0.071)
	Adaptive	0.101 (0.071)	0.271 (0.185)

Table 6.6 – Mean knowledge test and structural assessment scores (s.d.) for the control and adaptive condition

The pretest and posttest scores on the knowledge test and the structural knowledge assessment can be seen in table 6.6. An ANCOVA with pretest as covariate, condition as fixed factor and posttest as dependent variable, shows no significant effect of condition on the knowledge test [$F(1,25) < 1$], nor on the structural knowledge assessment [$F(1,25) < 1$].

Learning per unit of instruction
Dividing the posttest scores of the participants by the amount of victim cases triaged, gives us an indication of how much the participant has learned per unit of instruction, and whether this would be higher in a game that adapts the information presentation to the player's proficiency. Indeed, an ANCOVA with pretest as covariate, condition as fixed factor and posttest divided by the total number of victims triaged as dependent variable shows that condition had a significant effect on both the knowledge test [$F(1,25) = 21.98$, $p < 0.001$, $d = 1.81$] and the structural knowledge assessment [$F(1,25) = 5.05$, $p < 0.05$, $d = 0.89$]. The means and standard deviations of these tests are printed in table 6.7.

	Condition	Posttest per victim
Knowledge test	Control	0.57 (0.19)
	Adaptive	1.02 (0.30)
Structural knowledge assessment	Control	0.015 (0.004)
	Adaptive	0.028 (0.019)

Table 6.7 – Mean knowledge test and structural assessment scores (s.d.) per victim

In-game score
The total in-game score was significantly higher for the control condition ($M = 777.7$, $SD = 321.2$) than the adaptive condition ($M = 316.4$, $SD = 107.8$), $F(1,26) = 25.95$, $p < 0.001$, however this more or less follows from the previous result that participants triaged significantly less victims in the adaptive condition. Looking at the progression of in-game scores in figure 6.9, one can see that the line depicting the scores of the participants in the adaptive game

condition stays relatively level, at least until the 9th victim, after which there may be too few participants left (indicated by the number next to the marker) to make accurate statements. Conversely, the control condition first peaks at the easy lightly injured victims, before dropping steeply and then gradually climbing again over the remainder of the game. A Repeated-Measures ANOVA with the average score of victims 1 to 9 as repeated measures and condition as fixed factor, shows a significant effect of condition; Mauchly's test $\chi(35) = 34.02$, $p > 0.05$, $F(8,192) = 6.74$, $p < 0.001$.

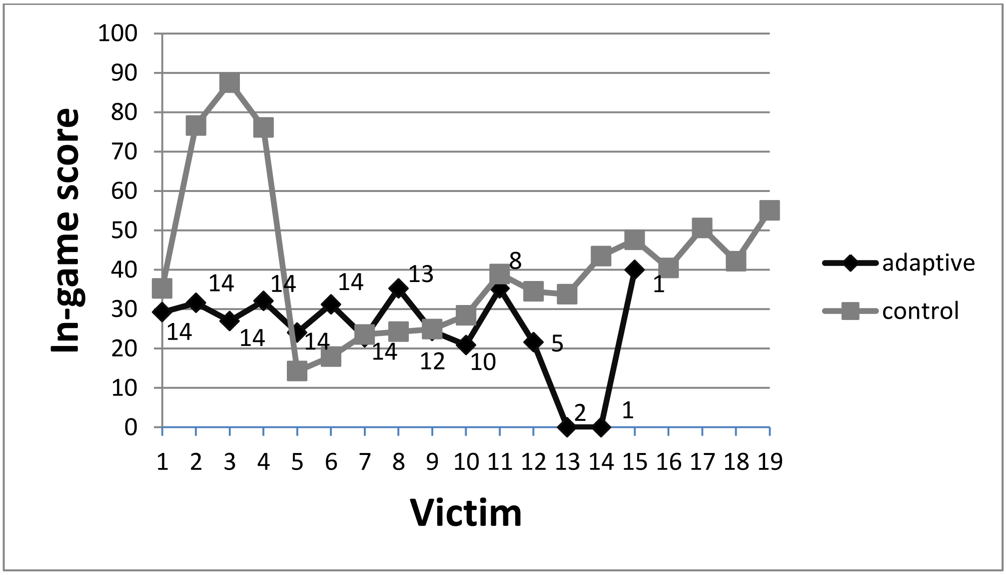


Figure 6.9 – In-game score progression per condition

Engagement

Reliability of the ITC-SOPI Engagement questionnaire was relatively poor, coefficient $\alpha = 0.59$. Scores on the statement 'I lost track of time' had a low item-total correlation; deleting this item yields a slightly better alpha of 0.64. The scores for the means and standard deviations of all the affective questionnaires are printed in table 6.8.

	Control	*Adaptive*
ITC-SOPI Engagement (1-5)	3.63 (0.33)	3.66 (0.45)
'at no point too easy' (1-5)	3.21 (1.37)	4.07 (1.21)
'at no point too hard (1-5)	2.57 (0.94)	2.93 (1.14)
'had enough training' (1-5)	2.29 (1.20)	2.07 (1.21)
Subjective Enjoyment (1-10)	7.43 (0.94)	7.14 (0.66)
Subjective Difficulty (1-10)	7.07 (1.49)	7.57 (1.70)

Table 6.8 – Means (s.d.) pertaining to the affective questionnaires

An ANOVA showed no significant effect of condition on the ITC-SOPI Engagement questionnaire, $F(1,26) < 1$. Nor was there an effect of condition on the subjective ratings of whether the participant enjoyed the game [$F(1,26) < 1$] or whether they found it difficult to reach a good score [$F(1,26) < 1$]. A MANOVA with the three added Likert-scale questions ‘at no point did I find the game too easy’, ‘at no point did I find the game too hard’ and ‘I think I have had sufficient training’ as multiple dependent variables, also shows no significant effect of condition, Wilks’ $\lambda = 0.83$, $F(3,24) = 1.62$, $p = 0.21$.

6.13 Conclusion and discussion

In chapter 4 we noted that Code Red Triage was less efficient than a more traditional PowerPoint presentation. In chapter 6a we contrasted different versions of complexity presentation to determine which version would lead to superior information organization, and in 6b we saw that even in the massed progressive complexity case, there was a comparatively large variability in the in-game score, indicating multiple levels of proficiency. We subsequently set out to discover if Code Red Triage could be made more efficient by letting players, who have shown to have organized the information well by getting a higher score, progress faster through the game via an adaptation system. We had three hypotheses concerning the adaptive game version; one, that participants would finish the game sooner, while gaining the same amount of knowledge; two, that individual learners learn more efficiently in the adaptive version because the complexity progression is better aligned to their proficiency level; and three, that this better alignment between proficiency and challenge would lead to higher engagement.

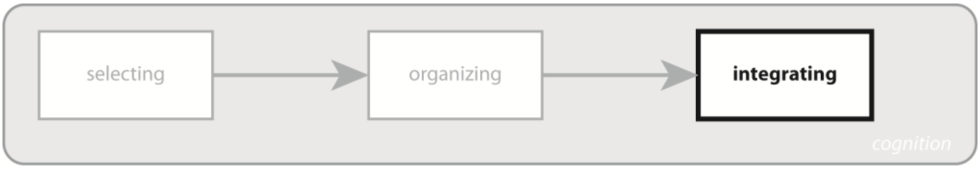
Both hypotheses one and two were confirmed; participants in the adaptive game version finished the game significantly quicker, while posttest knowledge test and structural knowledge assessment scores were equal. In addition, the participants also learned significantly more per victim case than in the control condition. There are two possible explanations for this effect. It could be that the moment a participant grasps the procedural path pertaining to a certain tier, the information presented with the following victim cases in that tier is redundant, at least to a point that it does not enhance the participant’s mental model greatly anymore. It could also be that the better adjusted challenge leads to a higher cognitive interest and therefore more active organization of the information; or both explanations at the same time.

In order to determine whether the adaptive condition not only made information organization more efficient, but even improved it, another experiment should be set up where the participants in the adaptive and control conditions have an equal amount of time. The cognitive interest explanation is repudiated somewhat by the rejection of the third hypothesis though, which stated that a better adjusted challenge would lead to higher engagement. For this too we propose two explanations, apart from the fact that adaptation evidently does not influence engagement.

It was already mentioned previously that in single-blind experiments as our own, the participants lack knowledge of the other condition and thereby reference material. The intervention itself may be too small next to all the other determinants of engagement, such as the setting, world, expectations, control interface, et cetera, to show up as a difference on the rating scale. Another explanation could lie in the difference in in-game score progression that was shown in figure 6.9. While the adaptive condition shows that participants had an approximately even level of challenge throughout, after a first spike and dip the participants in the control group consistently improved their scores. The possibly higher level of self-efficacy at the end of the game in the control condition may have offset the engagement inherent in a better adjusted game, as in the adaptive condition.

Chapter 7

Introducing surprising events to improve learning



Abstract--Serious games show great potential, but many fail to live up to that potential. One way to improve serious game design could be to introduce surprising events linked to the instructional material, which stimulate the active integration thereof into previous knowledge structures. We modified our serious game for triage training, called Code Red Triage, to include three surprising events pertaining to key moments in the triage procedure. Forty-one participants were divided into two groups: one group played a version of the game with the surprising events and the other group played a control version. A pre-posttest design showed no significant difference in engagement and superficial learning, but did show the participants in the surprising events condition obtained significantly superior knowledge structures, indicating that surprising events in a serious game foster deeper learning.

Parts of this chapter are based on:

Van der Spek, E.D., Van Oostendorp, H. & Meyer, J-J. Ch. (in press). Surprising Events Stimulate Deep Learning in a Serious Game. *British Journal of Educational Technology*.

7.1 Introduction

In the experiment where we contrasted our own game Code Red Triage to a PowerPoint presentation of the same material, we discovered that the more traditional instruction outperformed the game, at least when it comes to acquired declarative and procedural knowledge directly after the intervention, and the game also did not seem more engaging than the PowerPoint (see chapter 4). We furthermore found that the experienced cognitive load was significantly higher for the serious game and significantly negatively correlated with post intervention test scores. The cognitive load imposed by the inherently interactive nature of games on the selection and organization processes of multimedia learning therefore seemed to have an adverse effect on the efficacy of serious games, something which mirrors the outcome of research done on the interactivity of other multimedia learning environments (Moreno & Valdez, 2005). We therefore sought ways to improve the selection and organization process by introducing attentional cues (chapter 5) to improve the selection process and investigating the ideal complexity progression (chapters 6a and 6b) to ameliorate the effectiveness and efficiency of the organization process.

However, apart from reducing the cognitive load and thereby taking away obstructions that hinder learning, serious games also have a relatively unique characteristic that may be used to improve the effectiveness thereof, in that they can stimulate the integration process and thereby more active learning.

7.2 Improving learning with serious games by stimulating the integration process

Most games are not simply a multimodal representation of interactive information with some form of corrective feedback, but regularly distinguish themselves by having a narrative, either implied by the progression of obstacles and environments a player faces in a spatial narrative, or in the more traditional sense of storylines or plots (cf. Jenkins, 2004). Narrative events are known to be able to enhance learning in a number of different ways.

Hoeken and Van Vliet (2000) identified a number of different story techniques that can be used to enhance comprehension of a text: suspense, curiosity and surprise. Of these, the surprising event proved to be the most effective; it improved comprehension of the text as well as increasing appreciation for the story. For a way in which curiosity can be fostered in a serious game, potentially benefitting learning, see Wouters, Van Oostendorp, Boonekamp and Van der Spek (2011).

Experiments in text processing done by Zwaan and colleagues (1995) have shown that a reader comprehends a narrative on basis of a situation model—a reader's mental model of the story in terms of the protagonist, time, space, causality and intentionality, which is updated on one or more of these dimensions whenever an event takes place. They furthermore showed that important concepts in the story were linked together stronger when the corresponding events shared a number of values on the dimensions (e.g. if events happen in the same location at roughly the same time).

When a person plays a game, he or she will invariably create a situation model of the events and underlying rules of the game. If this initial situation or mental model turns out to be erroneous however, new corrective information will not necessarily lead to a significant update of the mental model; learners are known to cling to previous preconceptions almost to a fault (De Jong & Van Joolingen, 1998; Van Oostendorp & Bonebakker, 1999).

The underlying assumption of why surprising events improve comprehension of a story is that a surprising event is by its nature unexpected and therefore not (intuitively) a logical consequence of what preceded it; the player has to reevaluate his or her preconceptions and determine if they have missed something (Kintsch, 1980), thereby forcing a retrieval (or activation) and update of a preexisting mental or situation model (Zwaan, Langston & Graesser, 1995). A surprising event, therefore, stimulates the integration process in learning from multimedia.

From a neurological standpoint, a novel stimulus such as a surprising event will lead to a higher activation of a number of brain areas, releasing neuromodulators and noradrenaline, which subsequently results in more attention to the stimulus and better memory formation than with a familiar stimulus (Ranganath & Rainer, 2003).

Graesser, Chipman, Leeming and Biedenbach (2009) noted that complex learning and engagement in serious games are often at odds with each other, where, in line with the popular proverb 'no pain no gain', deep learning leads to less engagement. Deep learning here is the cognitive process in which the learner actively tries to make sense of new information, instead of only memorizing it; during this active knowledge construction, the learner builds a mental representation of the instructional material by using prior knowledge extensively (Moreno & Mayer, 2007). This concept of a mental representation or mental model is closely related to the situation model of Zwaan and colleagues (1995), and therefore we hypothesize that a surprising event forces an update of the situation model and subsequently leads to improved deep learning (Graesser et al., 2009).

If introducing surprising events can lead to a higher appreciation of a game, heighten cognitive interest neurologically and improve the integration of new information with prior knowledge into a coherent knowledge structure, while still being relevant in the context of the game world, then this game adaptation may mitigate the ostensible tradeoff between engagement and deep learning.

Given these considerations, we contend that narratives in games could be used to aid comprehension, anecdotal evidence of which has already been found by Dickey (2010), who noted that narrative helped sustain the player's motivation and cognitive interest while playing a 'whodunit' game.

Conversely however, in an experiment done by McQuiggan, Rowe, Lee and Lester (2008), it was found that adding a substantial narrative in the learning environment Crystal Island, though increasing feelings of presence in a virtual environment, did not lead to better learning. In fact, of three conditions, a PowerPoint narration, a version of the game with minimal narrative and a version of the game with a scenario and fully fleshed out back-stories to the characters, it was a case of less is more, where the PowerPoint presentation resulted in higher learning gains than the minimal narrative condition, which in turn showed higher learning gains than the full narrative condition. The researchers contended that the reason for this could be that the added narrative overloaded the player's cognitive system. In addition it is unclear how much of the characters' personalities and back-stories were directly related to the learning material or was simply superfluous information, meaning the players were perhaps learning more than the task that was tested required. Corroborating this notion, they noticed in a follow-up experiment that some of the story elements and locations led to off-task behavior, where the player spent time discovering things that were unrelated to the learning task, which in turn was negatively correlated with learning gains (Rowe, McQuiggan, Robison & Lester, 2009). Therefore, it is important that narrative events introduced in a serious game should not lead to irrelevant, off-task behavior.

The hypotheses we subsequently tested with this experiment is that implementing surprising events in a serious game (1) improves learning, possibly (2) deep learning (i.e. improved knowledge structures), and (3) leads to a higher enjoyment of the game.

7.3 Surprising events in triage training

To test whether introducing surprising events in a serious game can improve learning and engagement in a serious game (as well as other hypotheses, cf. Van der Spek et al., 2011), we created a 3D triage training game called Code Red Triage. The game is a First Person total modification of *Half Life 2 Episode Two*, where the player roleplays an ambulance first responder at the scene of a mass casualty event. The player arrives at a train station and receives an incoming call of a terrorist strike on a subway platform, with first reports indicating between 10 and 30 victims (in actuality there are 19). The player then has to find the way to the subway platform and upon arriving there categorize all the victims according to the urgency with which medical attention should be provided, so that when the other ambulance crews arrive, the most heavily injured victims will be treated first.

This is the full extent of the 'plot'; Code Red Triage operates a light narrative, which is mostly in the form of environmental storytelling (Jenkins, 2004), as opposed to more narrative-heavy learning environments such as that of McQuiggan and colleagues (2008). The player experiences a progression from a bright train station with minor traces of panic, through winding corridors to a moodily lighted subway platform with increasing signs of carnage that is built to evoke the feeling of what a player would expect the aftermath of a terrorist strike to look like. In addition, the succession of victim cases the player has to categorize could also be seen as a narrative of sorts.

The procedure with which the victims are categorized is called a triage; in our version of the game we teach the mobility sieve triage procedure (Hodgetts, 2002). Here, the medical first responder should roughly perform the following checks: determine whether the victim is still mobile, whether the airway is obstructed, measure the breathing frequency and the blood circulation. According to the outcome of these checks the victim should be categorized in one of four categories: highly urgent (T1), moderately urgent (T2), non-urgent (T3) and dead. For our experiment, it is also important to know that if the airway is obstructed it needs to be freed by either a chin lift in normal conditions, or a jaw thrust if the victim has a neck trauma. In addition, the blood circulation is ideally gauged with the capillary refill time check, but this is only applicable if there is enough light; in darkness, the player should check the pulse rate of the victim. The procedure has been slightly simplified from the real one in order to keep it learnable for laymen in a short time period.

Beforehand the player will only be told about the four different triage categories; the procedure with which to arrive at one of the four categories is for the player to find out by playing the game. When the player arrives at the subway platform he (or she, but for brevity purposes we will refer to the player as he from now on) will encounter the nineteen victims walking, standing, sitting or lying down in a roughly sequential manner. Pressing the e-button while standing in front of a victim will open up the triage menu screen (see figure 7.1). Here, the player can choose from a number of options, such as check the airway or perform the chin lift; pressing the corresponding button will give information on what the option entails in the triage situation, and what this means for the current victim. E.g. 'You determine whether the airway is free and the victim is able to breathe freely, if the airway is obstructed, it needs to be freed. / You do not have to check the airway of this victim, because the victim is still able to walk.'.

After the player has appraised the severity of the injuries, he can choose one of the four different categories. He will be awarded a score for how well he performed the triage, based on whether he clicked the correct buttons, whether this was done in the right order, the correct triage category was chosen and in what time frame. Briefly after triaging the victim, he sees how well he performed expressed in text, and the score is added to the total score at the top of the screen. The player ultimately has to discern the correct procedure by the information given by the button presses and the feedback presented to them after every triage. Previous experiments have shown that the score is a relevant measure for performance of the triage (see chapters 4, 5, 6).

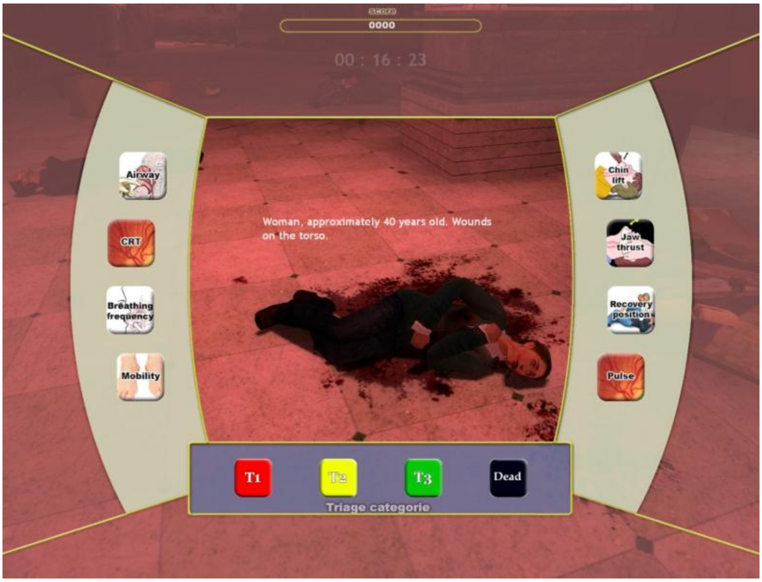


Figure 7.1 - Triage screen in game

After careful consideration of the game and the triage procedure, we decided that the moments most suitable to the introduction of a surprising narrative event, were (1) the mobility check, where being mobile means the victim is lightly injured; (2) the choice between a chin lift without neck trauma or a jaw thrust when the victim does have a neck trauma; and (3) the choice between the capillary refill test and measuring the pulse rate based on lighting conditions. These were key moments in which new information to some extent contrasted the previous notions of the player. E.g. in the case of the neck trauma event the player had been using the chin lift, but needed to switch to the jaw thrust, as using a chin lift on a patient with a neck trauma would lead to serious injuries. The surprising events, which will be explained further in the Experimental conditions section, preceded these key moments and should furthermore be designed to not lead to irrelevant off-task behavior.

7.4 Experiment

7.4.1 Participants

In total 41 persons participated in the experiment, but two participants did not answer all of the post-game questions, resulting in a different valid N for that test. The participants were all university students following a masters-level game design course; participation in the experiment was a compulsory prerequisite to pass the exam, but the students were told beforehand that they would receive a coupon worth € 7.50, in order to counterbalance this possible demotivation effect.

The average age of the participants was 23.3 (*SD* = 1.89), with 36 male and 5 female. Two women were in the control condition, three in the experimental condition. As the participants were students enrolled in a game design course, most of them had an interest in gaming. Out of 41, 2 hardly played videogames, 21 occasionally played games, and 18 would label themselves as ‘a gamer’.

7.4.2 Experimental conditions

In order to test the hypothesis whether surprising events can foster learning, we created a condition with three surprising events, pertaining to three different choices/steps in the triage procedure. These steps, as well as the corresponding events, are furthermore unrelated to each other, so as to prevent any interaction effects. This Surprising Event (SE) condition is then contrasted to a control condition, which does not have these surprising events.

Both conditions of the game are based on a version of the game where the instruction is built up in such a way that the game starts off easy and gradually becomes more difficult. In practice, this means that the player begins with only two buttons, or steps in the procedure, to choose from for the first victim: a mobility check and investigating the airway. From one victim to the next, buttons will gradually be added until the thirteenth victim, when all eight buttons are available for the remaining six victims (see figure 7.1). A similar buildup of difficulty can be seen in the victim cases, where the first few victims will only require a mobility check to determine that they are lightly injured, but subsequent victims will progressively require more procedural steps to arrive at a triage decision.

The three surprising events are a mobility event, a neck trauma event and a light failure event. All surprising events are in the form of a short in-game cut scene of two to four seconds and occur only once before a new concept is taught. See figure 7.2 for a timeline of screenshots of the different events.

Mobility event

The mobility event takes place during the first encounter with a victim, in the hallway before the player goes down the stairs towards the subway platform. In the SE condition, the victim runs towards the player screaming, and cowers down in front of him. In the control condition, the victim walks around in circles at a leisurely pace. The purpose of the first few victims is to teach that if a victim is able to walk, then that victim is lightly injured according to the triage procedure.

The SE condition is more eventful; it adds sound (the victim is mute in the control condition), the player experiences a cut scene for the first time, and the victim running towards the player adds immediacy. There are a total of five consecutive victims that are lightly injured, for which only the mobility has to be checked.

Neck trauma event
In the neck trauma event in the SE condition, the thirteenth victim in the game is standing upright, back turned towards the player. The moment the player approaches the victim, the cut scene starts. Here, the camera starts shaking and a rumbling sound is played. A large piece of ceiling debris comes crumbling down and falls on the victim's head, who produces a startled scream before collapsing to the ground. This surprising event leads to a neck trauma in the victim, which in turn means that the player should perform a jaw thrust instead of a chin lift. In the control condition, the victim is already lying down on the floor unconscious, with the suspicion of a neck trauma introduced by the text. There are two consecutive victims that suffer neck trauma.

Light failure event
The final notion that we wanted to highlight with a surprising event, was that the capillary refill check is preferable unless there's insufficient lighting, in which case the player has to check the pulse rate. As is noticeable from figure 7.2, the subway platform area of the game looks different in the two conditions of the game. In the control condition, the emergency lighting was turned on, presumably (to the player) because the blast impact caused a power outage. At the outset all the victims are bathed in enough red light to perform a capillary refill check, but the last few victims are placed in shadows, blocking the possibility to check their blood circulation with the capillary refill time. In the SE condition the normal incandescent lighting is still on upon reaching the platform. Right before reaching the fourteenth victim the surprising event occurs; a cut-scene of an explosion in a powerbox leading to a temporary blackout. Two seconds later the emergency lights jitter on, accompanied by a loud mechanical whizzing sound. After this it becomes apparent to the player that it is too dark to perform a capillary refill check. Four victims need their pulse rate checked.

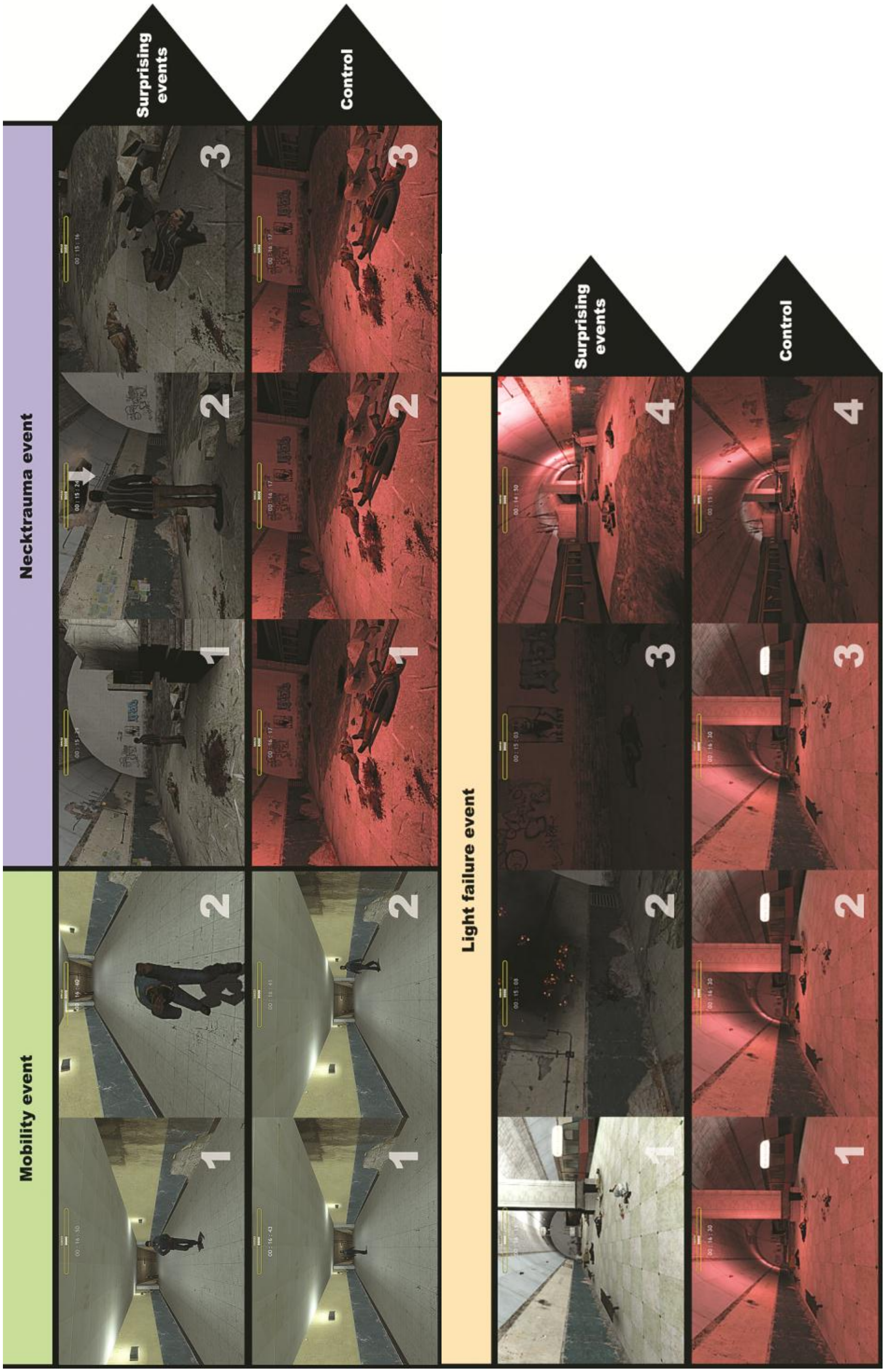


Figure 7.2 – The three surprising events and control condition equivalent in Code Red Triage

7.4.3 Materials

In order to test our hypotheses that surprising events in games foster learning and engagement, we used a number of different instruments. As was mentioned earlier, the game records how well the player performs the victim triages while playing the game. While this in-game score may not give an accurate depiction of how well the player is learning—triaging a victim incorrectly leads to a lower score but may still give the player insight into the procedure, it does tell us whether the experimental intervention leads to the player exhibiting desired playing behavior, i.e. play the game as it is meant to be played.

To measure how well participants have learned the triage procedure, we used a pre-posttest design with a knowledge test (see Appendix A). The sixteen questions on this test were divided into eight verbal questions and eight pictorial questions. The pictorial questions were rendered mockups of victims shown together with a short description and the participant then has to choose the next step in the triage procedure, which is more in line with the game (see chapter 5). All of these sixteen questions were four-item multiple choice questions.

The eight questions in either the verbal and pictorial part of the knowledge test consisted of three questions that were related to the three surprising events. Here, the participant was presented with a victim case with certain injuries, and had to decide upon the next step in the triage procedure. Because they were in the form of case resolutions and did not ask about the material pertaining to the surprising event outright, we named these *indirect questions*. The remaining five questions were about other, unrelated parts of the procedure. We called these *'unrelated' questions*.

In addition to these sixteen questions, we added five *direct questions* (one for mobility event, two for neck trauma and lighting event) asking outright e.g. ‘if it is dark and you want to check a victim’s blood circulation, then according to the procedure of the primary triage you should use..’, again with four possible answers. We only asked these questions post intervention, as we felt they would direct the participant into attending the related instruction if they had come up in a pretest knowledge test too much.

Summarizing, there were six (three verbal and three pictorial) indirect questions of the instructional material pertaining to the surprising events, five direct questions, and ten questions unrelated to the surprising events.

In addition to gauging the effect of surprising events in a game narrative on a more superficial declarative level, we were also interested in the effect on a deeper level of understanding (‘deep learning’). In order to test this, Trumpower, Sharara and Goldsmith (2010) among others propose a structural assessment of knowledge, where, in our case, participants rate concepts pertaining to the triage procedure on their degree of relatedness. The software program Pathfinder can then be used to construct distance networks called PFnets from this data, which in turn can be averaged and used to compare groups of participants to a reference expert network (cf. Schvaneveldt, 1985).

This Structural Knowledge Assessment (SKA) technique is able to distinguish between experts and novices and to predict transfer performance on decision-making tasks (Day, Arthur Jr. & Gettman, 1994), which is of key importance to assessing training games such as ours. Furthermore, we've found it was able to discern the effects of game design on knowledge construction in a previous experiment with our game (the attentional cueing experiment in chapter 5).

An example of the resultant network can be seen in figure 7.3. As the participants have to rate each combination of pairs, the number of possible combinations and thereby ratings grows quickly with each new concept added. Therefore, we only chose a subset of the triage procedure that incorporates the information pertaining to the surprising events. The network in figure 7.3 was derived from averaging the PFnets of two persons with extensive textbook knowledge of the triage procedure, and serves as the referent network against which the participants are compared. Concepts that are closely related have direct links, while less related concepts are placed further away in the network, via intermediate links. We find that, grosso modo, novices have more diffuse networks, whereas the expert has a cleaner network with 'classify victim' as the central node, linked to the main steps in the triage procedure and a cluster comprising chin lift, jaw thrust and blocked airway with neck trauma.

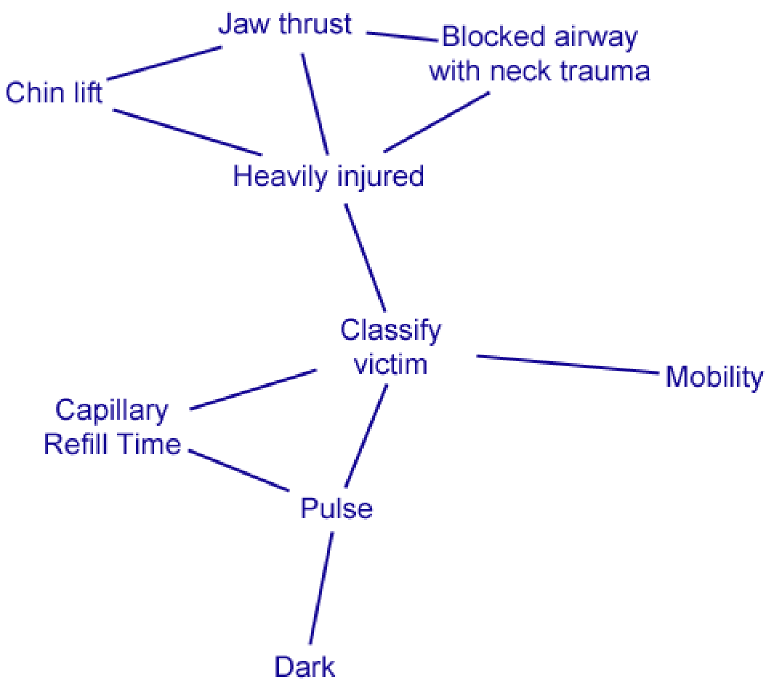


Figure 7.3 - Referent conceptual network for the triage procedure

The last test we used was for measuring the subjective emotional response to the game. We asked them to rate how difficult they found the game and how much they enjoyed the game, on a score of 1 to 10. In addition, we used the engagement subset of the Independent Television Commission's Sense of Presence Inventory (ITC-SOPI: Lessiter, Freeman, Keogh & Davidoff, 2001). By this time the participant is subjected to quite a number of tests, so we used only this subset, as it gives a mix of presence and engagement in a twelve item Likert-scale.

7.4.4 Procedure and apparatus

The participants were sat behind a desk with a 17" widescreen laptop computer, which was closed off from the rest of the room by a cubicle partitioning wall. They first filled in a small demographics questionnaire and were the presented the different tests in the order as can be seen in figure 7.4.

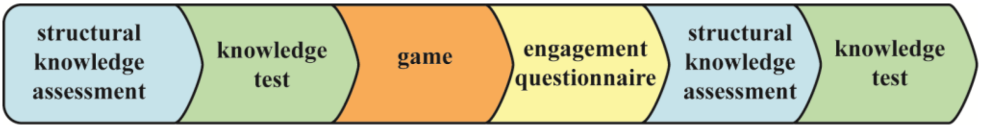


Figure 7.4 – procedure of the experiment

7.5 Results

Table 7.1 shows the percentages of correct answers on the different paper knowledge tests: indirect questions pertaining to the surprising events, direct questions pertaining to the surprising events, the aggregate average of these questions and finally the scores on the questions unrelated to the surprising events. As we explained in section 7.4.3, the direct questions were only presented after the game was played; therefore we used the pretest indirect questions as a covariate for both the posttest indirect questions and the posttest direct questions. For this reason they appear twice in table 7.1; the indirect questions pretest scores have been colored grey for the direct questions, to reflect that they are the same as above.

Multiple MANCOVAs were performed with direct and indirect posttest questions as multiple dependent variables, corresponding indirect pretest score as covariate and condition as fixed factor. No significant effects of condition were reached for either the questions pertaining to mobility (Wilks $\lambda = 0.96$, $F < 1$), pertaining to the chin lift / jaw thrust distinction (Wilks $\lambda = 0.97$, $F < 1$) and the blood circulation (Wilks $\lambda = 0.90$, $F(2,37) = 2.0$, $p = 0.21$), although the latter does show a trend for the direct questions: $F(1,38) = 3.45$, $p = 0.07$, multivariate η^2 is 1 minus $\lambda = 0.10$. The aggregate scores (the aggregates of the direct and indirect questions), as well as the unrelated questions also were not significantly ($p > 0.05$) different with condition as factor (ANCOVA $F < 1$).

		Control		Surprising Events	
		Pretest	*Posttest*	*Pretest*	*Posttest*
Indirect questions	Mobility	17.50 (24.47)	22.50 (41.28)	7.14 (17.93)	26.19 (43.64)
	Chin lift / Jaw thrust	25.00 (30.35)	60.00 (34.79)	28.57 (29.88)	71.43 (40.53)
	Blood circulation	37.50 (22.21)	85.00 (23.51)	50.00 (35.36)	90.48 (20.12)
Direct questions	Mobility	17.50 (24.47)	30.00 (47.01)	7.14 (17.93)	28.57 (46.29)
	Chin lift / Jaw thrust	25.00 (30.35)	72.50 (34.31)	28.57 (29.88)	73.81 (40.68)
	Blood circulation	37.50 (22.21)	90.00 (26.16)	50.00 (35.36)	100 (0)
Aggregate average		26.67 (17.44)	60.00 (20.34)	28.57 (19.82)	65.08 (21.35)
Unrelated questions		29.00 (11.19)	53.00 (14.90)	30.95 (16.40)	52.00 (17.35)

Table 7.1 - Means (and s.d.) of the test scores in percentages

	Control		Surprising Events	
	Pretest	*Posttest*	*Pretest*	*Posttest*
Structural knowledge assessment	0.055 (0.087)	0.176 (0.114)	0.071 (0.091)	0.280 (0.158)

Table 7.2 - Means (and s.d.) of the structural knowledge assessment (SKA) similarity scores

Note that, even though none of the effects are significant, the participants in the experimental condition with the surprising events have almost consistently scored higher for all of the different tests.

Table 7.2 shows the means and standard deviations for the SKA, after one posttest outlier was deleted from the control group. An ANCOVA on the SKA test, with SKA posttest as dependent variable, SKA pretest as covariate and condition as fixed factor shows a significant difference for condition $F(1,37) = 5.47$, $p < 0.05$, $\eta^2 = 0.12$. Figure 7.5 shows a bar chart of the difference in means and standard deviations. Participants in the experimental condition scored higher on the Structural Knowledge Assessment test than subjects in the control condition.

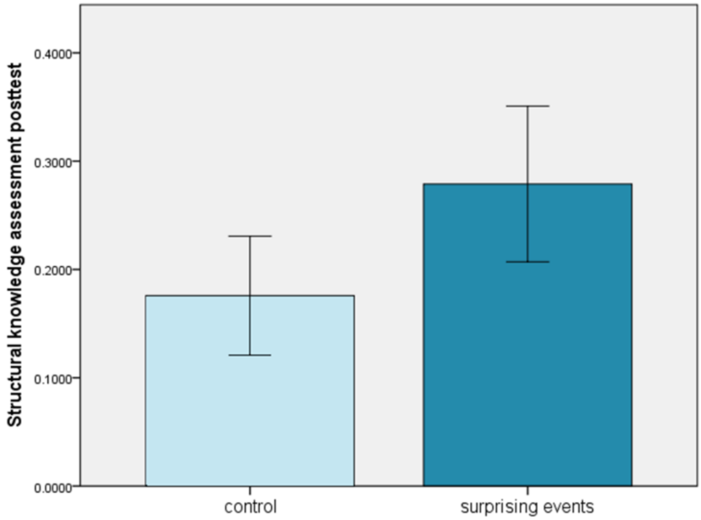


Figure 7.5 - Posttest structural knowledge assessment similarity scores

Table 7.3 shows the results for the in-game score, subjective difficulty rating, subjective enjoyment rating and the ITC Sense of Presence Inventory Engagement mean. Overall the SE condition seemed to score a little bit better during the game, but none of these statistics were significant, $F(1,39) < 1$ for all of the ANOVAs and the reliability of the ITC-SOPI scale was poor with coefficient $\alpha = 0.53$. Individually, the Engagement subquestion 'I can vividly remember parts of the experience' seems to indicate a trend in favor of the surprising events condition ($F(1,39) = 3.55$, $p = 0.07$, $d = 0.59$).

	Control	Surprising events
In-game score	718.35 (214.99)	754.33 (228.23)
Subjective difficulty	7.45 (1.15)	7.38 (1.36)
Subjective enjoyment	6.50 (1.57)	6.43 (1.47)
ITC-SOPI Engagement	3.32 (0.29)	3.29 (0.43)

Table 7.3 – Means (and s.d.) of the in-game score and extra test items

If we take a deeper look at the in-game score, and investigate the score for the triage of the first two victims immediately after the surprising event occurs (the first victim to 'learn' how the event changes the game rules, and the second to apply this new knowledge), as a measure of immediate understanding of the game rules, then we see the same trend. The participants in the SE condition scored higher on average than those of the control condition after every event, however this is not significant, $F(1,39) < 1$.

Finally, we include a number of correlations that arise from the data. The self-reported scale where participants had to rate how much of a gamer they see themselves is positively correlated with the posttest pictorial knowledge test ($r(39) = 0.37$, $p < 0.05$), but not the verbal test, possibly indicating a relationship between aptness in processing visual information and playing games, or willingness to play games.

In the SE condition, the SKA posttest is positively correlated with the in-game score ($r(19) = 0.47$, $p < 0.05$) and with the ITC-SOPI Engagement score ($r(19) = 0.45$, $p < 0.05$), however the engagement and in-game score are uncorrelated. None of these scores are significantly correlated in the control condition. As the SKA scores are higher in the SE condition, this means that knowledge construction is integral to a good in-game score and the feeling of being drawn into a game world, but only at a sufficient level of understanding.

7.6 Conclusion and discussion

7.6.1 On learning

We set out to determine whether surprising events in a game narrative can be used to improve the integration of new information into prior knowledge structures, and subsequently learning. Our experiment has shown that a relatively easy adaptation to an existing serious game can already stimulate deeper learning, without harming the engaging characteristics of the game. The surprising events improved the way the newfound knowledge is integrated into the player's mental model, as was evidenced by a structural knowledge assessment. This is likely because the surprising event precipitated an update of the mental model, where the mental model is retrieved from the long term memory and the new information has to be integrated. Therefore, introducing surprising events is a good way of increasing the effectiveness of a serious game.

Curiously, the surprising events did not lead to better scores on the paper knowledge tests. On face value, this simply implies that a change in the game narrative only stimulates deeper knowledge construction, and not more superficial (verbal) knowledge acquisition. A possible explanation for this may be that the narrative is interpreted on a deeper, semantic level, and any changes therein will therefore lead to a comprehension that differs on a semantic level and not necessarily on other levels. However, there are a number of different provisos to be made.

When looking at the results of the paper knowledge tests in more detail, the hundred percent correctness of the direct questions on the capillary refill time and pulse rate distinction in the SE condition is remarkable, but because the control condition scored highly on this question as well, the difference was not significant. It may well be that the control condition of the game already was too successful as a learning material to notice a significant difference, and that introducing surprising events may benefit other serious games even on a knowledge acquisition level. This notion is repudiated by the fact that no difference in in-game score was found directly after an event, although we already argued that in-game score may not fully reflect actual learning. It does, however bring us to another possible threat to the validity.

Our game lasted around twenty minutes; it may additionally be that three surprising events of a few seconds is too small an intervention to induce more substantial knowledge acquisition effects. In addition, we also did not test for knowledge retention over longer periods. Previously, we found that the knowledge acquired from a game is generally retained longer than that from the more traditional verbal instruction (see chapter 4), which may be due to the eventful nature of games. Therefore introducing surprising events could lead to even better retention. However, this needs to be researched.

Another interesting avenue for future research comes from neuropsychology, which shows that inducing stress or arousal in a person after (instead of before) something has been learned, leads to better recall (cf. McGaugh, 2006). Already many entertainment games first teach a player a new technique in a relatively quiet part of the game, and then have the player use this in a stressful situation such as a boss fight. If this was found to improve learning, it could be worthwhile to discern how much of this change can be attributed to practicing the new skill, or due to increased arousal levels.

7.6.2 On engagement

There was no significant difference in engagement and subjective difficulty and enjoyment between the two conditions of the game. One can therefore say that introducing a few surprising events does not improve the engagement of a serious game. Naturally the same provisions concerning the time of the instruction and the number and nature of the events apply to the subjective enjoyment of the game as well. However, here too a number of additional remarks can be made on the validity of this finding.

Firstly, the participants only saw one version of the game and therefore had to rate the engaging qualities without being able to compare the two versions. For most participants it furthermore was their first real encounter with a serious training game of this type. The problem is that without a clear referent, it becomes unclear what the participant's prior expectations are. When people gain more experience with serious games, they may start to demand more interesting gameplay experiences than they do now.

Conversely, it could also be that a serious game that has a stimulating narrative is more fun in direct comparison to a condition without surprising events, but that this raises the expectations of the player to that of entertainment games, which is a hurdle too high for many financially challenged serious games developers. This is a question that needs to be answered with more longitudinal research where participants get to experience multiple serious games.

A final remark concerns the use of Likert-scale tests with laymen. In contrast to e.g. the participants of the experiments in chapter 5 and 6a, which had a more diverse background of university studies, all the participants in this experiment were students in computer and information science, and as such had little experience with experimental testing and Likert scales. This may, to some extent, explain why the reliability of the scale was markedly lower than previously. In addition, although the participants were encouraged to be as 'painfully honest as possible', most refrained from using the extreme ends of the rating scale. With a 5-point scale, that leaves little discriminative power. One person in the experiment actually physically jumped the moment the lights in the game went out with an explosion. Afterwards he rated 'I can vividly remember parts of the experience' and 'I experienced emotions while playing' with a neutral score (3), which may well be a case of Dutch reservedness. As a result of these remarks, we contend that the question pertaining to engagement is not satisfactorily answered by our experiment and needs further research.

Chapter 8

General conclusion and discussion

8.1 Main conclusions

The main research question of this dissertation was:

Can we discover game design techniques based on cognitive principles that improve the efficacy of serious games, without harming the engaging qualities of the game?

This question was motivated by a failure of some serious games to reach their learning goals and an ostensible lack of systematic empirical research when it came to the cognitive effects of game design decisions. In order to determine whether game design techniques influence the potential of games for learning as well as its engaging qualities, and subsequently whether these game design techniques can be used to make serious games more effective, we created a serious game called Code Red Triage.

This serious game, a total conversion mod of the popular game Half-Life 2—which in turn is based on the Source Engine, was designed to let the player take on the role of a medical first responder, and teach him or her the mobility sieve triage procedure, a procedure used for categorizing the victims of a mass-casualty event in order of the urgency with which medical attention is needed. In the first of four game design experiments (chapter 5), the game was partially embedded in the triage training; that is, participants received an explanation of the triage and briefly saw the procedure before playing the game, and were able to retrieve the triage procedure in-game as well. This was shown to possibly influence the results in two of the three conditions of that first experiment, which meant that the effects of our game design interventions on learning could have been muddled. For the remaining three experiments we therefore changed the setup of the experiment and the game so that the whole of the triage instruction was contained within the game.

We subscribed to Mayer's (1999; 2005; 2008) theory of multimedia learning, which states that incoming audiovisual information in a game setting is *selected*, *organized* and then *integrated* into a mental model before it is stored in long-term memory (see chapter 2). These three cognitive processes for the greater part take place in a person's working memory, which only has a limited capacity and can be overloaded (Baddeley, 2002). While it's not a direct proof of overload, in chapter 4 we did find that the cognitive load experienced in the game was significantly higher than when the same instruction was presented as a PowerPoint, even though the participants in the game condition had slightly longer to engage in the material, and that this cognitive load was negatively correlated with learning gains. Therefore, learning could hypothetically be improved by better adjusting the game design to the cognitive system of the player, i.e. mitigating the cognitive load pertaining to, or optimizing, the selection, organization and integration process.

First, we tested whether the selection process could be ameliorated by using cues (De Koning et al., 2007). Auditory and visual cues were implemented that indirectly guided the player in choosing the correct next step in the triage procedure (chapter 5). The visual guidance cues condition, where the next step was hinted at by showing a pulsating bright green arrow that points at the limb or body area of the victim indicating where the next check should be performed, showed similar learning gains and improved knowledge structures as the control condition with no cues. The auditory cueing condition, where the next correct step in the procedure was hinted at by playing a sound that could be connoted to the check that should be performed, even scored significantly worse than the control condition on the structural knowledge assessment, and arguably also on the knowledge test. There was a significant effect in the visual cueing condition of game experience on learning gains, where participants with prior game experience scored higher. This could indicate that cueing only works when the purpose of the cue is clear, in this case because the gamer population has prior experience with visual cues in nearly all contemporary videogames.

Secondly, in chapter 6a we tested whether the organization process could be optimized by determining the best way in which the complexity of the game's obstacles (in our game victim cases) should be presented in the level design. In videogames, the games obstacles or problems generally progress from simple to complex, as do the options of the player to overcome these problems; these are oftentimes presented in a just-in-time manner. We contrasted two variables, problem presentation and options presentation. Problem presentation entailed the way the victim (i.e. problem) cases were presented: either in a massed, progressively more complex manner; or with a spaced, high variability in the complexity (Cepeda et al., 2006). Options presentation comprised the way the new triage checks or options were introduced: either all from the beginning (just-in-case), or progressively building up from when they were first needed (just-in-time) (Kester et al., 2001).

Interestingly, we found no significant difference in learning gains, indicating that the level and interface design in terms of complexity presentation is ostensibly inconsequential to the learning process, at least when it comes to learning a single relatively simple procedure in a relatively short serious game. It is unclear whether this is because there is actually no effect, or because the player encountered a sufficient amount of cases to gain the same basic level op understanding of the triage procedure in every version of the game.

To build on this further, in chapter 6b we set out to see if the presentation mode of complexity, instead of engendering more effective information organization, could be harnessed to engender more efficient learning. As games continuously assess how well the player is performing (Shute et al., 2009), they can adapt themselves to the proficiency of the player. The progressive victim complexity, just-in-case interface version of the game was fitted with an adaptation engine (Lopes & Bidarra, 2011) that skipped consecutive victims of the same tier if the participant had managed to score above a certain threshold. This turned out to significantly increase the efficiency of the game, by cutting out on average almost half of the number of victims, while leading to the same learning gains as in the control condition.

Fourth and last, in chapter 7 we performed an experiment where we introduced surprising events into the game narrative. This was hypothesized to force an update, and thereby evaluation, of an existing mental model, which in turn stimulates integrating the new information more effectively (Zwaan et al., 1995). The surprising events were shown to significantly improve mental model construction as evidenced by the structural knowledge assessment, which appears to confirm the hypothesis. However it did not lead to improved scores on the knowledge test; this is possibly due to a ceiling effect.

Perhaps somewhat surprisingly, nearly none of the game design interventions had an effect on the affective questionnaires, that is, reported engagement, enjoyment and difficulty appraisals of the game. This was actually what we set out to do in order to construe useful game design guidelines; that the guidelines should improve learning without harming the engaging qualities of the game. Only the condition with a progressive victim complexity and just-in-case user interface (chapter 6a) was enjoyed significantly more than the other complexity presentation conditions in the respective experiment. This was likely due to a higher sense of autonomy and possibly competence.

8.2 Cognition-based guidelines to improve serious game design

As a result of the research outlined in this dissertation, the following cognition-based design guidelines can be propounded to improve the efficacy of serious games.

1. Using attentional cues does not lead to better learning, at least for novice gamers or when the purpose of the cue is unclear. Ambiguous auditory cues even lead to worse learning. Visual cues may have a positive effect on experienced gamers.

2. The way in which the complexity of the problems and the options that are needed to resolve these problems are presented in the game is inconsequential to learning. However, a progressive increase in the complexity of problems, together with the player having all the options beforehand available, leads to a more enjoyable game.

3. Adapting the game to the player's performance, as demonstrated by the in-game score, significantly improves the efficiency of a serious game.

4. Introducing surprising events pertaining to the instructional material in a game significantly improves deeper learning.

8.3 General limitations of the research

The research in this dissertation was undertaken to see if we could find cognition-based design guidelines to improve learning in a serious game. The cognitive correlates of game design decisions had until then, to our knowledge, never been tested systematically and empirically. For this research we therefore tested whether design techniques can positively impact learning; and of the three cognitive processes inherent in mental model construction, selection, organization and integration, targeting which will have the highest potential of ameliorating learning. We created a single serious game with a singular instructional goal, to see which of the game design techniques benefit learning the most, however herein lies the largest limitation of the research. All of the effects described in this dissertation are effects on the learning of an explicit procedure, in a 3D virtual environment, within a single short session. Generalization to other game settings, game type and learning goals should be performed with caution (cf. Shapiro & Pena, 2009).

In addition, all our participants were university students, nearly all of whom were enrolled in the same university as the experiment took place; some were additionally highly familiar with the building, which could influence the affective response to the laboratory setting. Aside from the possible bias the choice for ostensibly intelligent participants could have exerted on the way the user's cognition works in relation to the design guidelines, it could also be that the experimental setting influenced the appraisal of the engagement of the game, as the motivation to play the game is different than leisure. On the other hand, this problem is mitigated by the fact that most serious games are not played for leisure either, but embedded in a compulsory curriculum.

After all of our experiments, the exact role of cognitive load is also still relatively unclear. In the comparison of Code Red Triage to the more traditional PowerPoint presentation, we discovered that experienced cognitive load was negatively correlated with learning. This would mean that decreasing cognitive load would lead to higher learning gains. Conversely, it is for instance highly unlikely that the adaptive condition in the experiment of chapter 6b would lead to lower cognitive load (if anything it would be higher), yet it did lead to significantly more efficient learning. Apparently, a higher cognitive load is not negative in itself; whether it is positive or negative depends on the capacity of the learners. If the cognitive load is still within boundaries, increasing cognitive load could have positive effects (Paas et al., 2004).

That there was no difference in the appraisal of the game's difficulty in this or all game design experiments could hint at four explanations for the apparent lack of the interplay between managing cognitive load through game design and learning. One, that the relationship only works one way; a higher cognitive load deteriorates learning gains, but higher learning gains don't necessarily entail a lower cognitive load. Two, while the previous explanation is ostensibly plausible, it could also be possible that the difficulty appraisal question in our post-test questionnaire failed to ascertain differences correctly.

The participants had to give a single appraisal on the difficulty over a sometimes 20 minute game, it could be that some participants remembered the easy parts, some the more difficult parts, some tried to estimate how well they did from the last overall score, while others found it too hard to give an accurate appraisal over such a long time period and therefore simply went for the politically correct answer (cf. De Jong, 2010). Thirdly, Grootjen, Neerincx and Veltman (2006) state that a higher cognitive task load imposed by the system may be counteracted by the user investing more effort. In the case of serious games especially, a more difficult game may stimulate some gamers to invest more effort, while other gamers become agitated by this and invest less effort. We did not measure preference for difficulty in games, but this could potentially confound the subjective difficulty rating. Perhaps there is even an argument to be made that this difference in difficulty preference may in part explain the lower reliability of the ITC-SOPI scale in the experiment of chapter 6a, in which one of the conditions presumably was more cognitively demanding. Fourthly, the appraisal only gauged overall cognitive load; in the terms of Cognitive Load Theory (cf. Kirschner, 2002), it could be that our interventions decreased extraneous cognitive load, while simultaneously increasing germane cognitive load. However we did not have a metric to distinguish between these types of cognitive load—if in fact they can even be accurately distinguished (also see De Jong, 2010), nor is it likely that one of the experiments that saw a marked improvement of learning, the surprising events experiment, reduced extraneous cognitive load. In retrospect it is unfortunate that the cueing experiment lacked a difficulty appraisal, making it difficult to determine whether the participants in the auditory cueing condition, which resulted in poorer learning, suffered from a higher cognitive load.

The reason we only asked participants to appraise their difficulty in reaching a good game score at the end of game play, was that if we did ask them at set points during the game, this could abate the player's immersion into the game. The arguably already flimsy illusion that the participant is engaged in playing a real game could be shattered by a popup prompting them to step out of the game and back into the real world, thereby transforming from a medical first responder doing an important job back into a participant participating in an experiment with no clearly defined (to the participant) goal. Incidentally, this may also be the reason for a lack of difference in the enjoyment scores in the comparison experiment with the PowerPoint presentation, where we did employ a popup halfway through the game to measure cognitive load.

It should however be noted that the standard deviations of the difficulty appraisals were, for the most part, relatively small. We therefore hazard the opinion that the first explanation, which stated that our game design techniques influence learning, but not overall cognitive load (or at least do not push it past capacity boundaries), is still the most probable one.

8.4 Implications and Future research

One of our reasons to target the selection, organization and integration processes of mental model construction during gameplay, was to exploratorily quantify, e.g. via effect sizes, which of the type of techniques would be most effective in improving learning with serious games. However this is defeated slightly by the fact that only the surprising events experiment, and thereby stimulating integration, led to an unequivocal improvement of the effectiveness (note: not efficiency) of our serious game (chapter 7). Stimulating knowledge integration is subsequently ostensibly the best choice to improve learning with a serious game. The implication for game developers therefore is that they should create a surprising event around the time a new piece of instruction is introduced that is important for understanding an element of gameplay. For game researchers it shows that other techniques to stimulate the integration of information into a mental model, such as introducing contradictory information or the use of feedback, are promising avenues for future research. In addition, the article of Hoeken and Van Vliet (2000) that was used as the basis of our surprising events experiment, also experimented on the narrative techniques of suspense and curiosity, which was shown to improve comprehension of a text due to heightened cognitive interest. The relationship of cognitive interest on mental model construction has been largely left untouched by our research; the way in which this potentially impacts information selection, organization and integration may lead to even more game design guidelines.

That some experiments did not lead to significant learning improvements does not mean that the pertaining cognitive processes are not worth investigating. In fact, given the relative effectiveness of using cues in other multimedia instruction (e.g. De Koning et al., 2007; Van Oostendorp et al., 2008), there is ample reason to pursue additional research lines pertaining to cueing. For instance, how will cues fare if their intention is communicated to the player beforehand, so that they can be used by the player unambiguously, and how does this relate to more implicit cues? In addition, cues can be used for selecting as well as better organizing the information; finally, it is unclear which, or a possible combination thereof, is more advantageous.

That we found no difference in learning when it came to different complexity presentation strategies in game design (chapter 6a) was unexpected but an intriguing result. We contend that it is probable that this has to do with the brevity of the game and the singular procedure that was learned in the game. More longitudinal research both in terms of the length of the game and the retention over longer time periods would be needed to reach more definitive conclusions. In addition, while some of the case resolutions require a path in the procedure that incorporates a previous path (e.g. for both a T3 and a T2 victim, one has to start by checking the victim's mobility), they could also all be seen as distinct paths to reach a resolution. It could therefore be worthwhile to test the different complexity presentation strategies on more complex problems with many interdependent constituencies.

While essentially only the surprising events made the game a more effective learning tool in that a better mental model was formed from the learning material than a control group, we did find that adapting the challenge to the performance of the player led to a marked increase in the efficiency of the game, where close to half of the amount of instruction could be cut down (chapter 6b). There are many opportunities for more research that can be done as a result of this that could improve this further. For instance, we briefly discussed an experiment where a camera was used to measure the stress a player experienced. While setting up a camera in front of the player is only lightly intrusive, it would arguably be better to find out if there are measures of stress and engagement that can be embedded stealthily in the gameplay itself. These measures could both be more reliable than post intervention questionnaires, and create opportunities for adaptation for every game without the need for additional materials.

It is also interesting to see if the improved efficiency due to adaptation holds true for other serious games and other learning goals. If this is the case, this could have important implications for both serious and entertainment game developers. Serious game developers can implement an adaptation engine to improve the efficiency of their serious games; it would then be interesting to ascertain if the inefficiency of serious games in general, as compared with traditional forms of education, can be wholly mitigated by adapting the game to the player. For entertainment games, the introductory tutorial stage sometimes is a necessary evil in order to teach the player the basics of the gameplay; this stage can't be very challenging or the absolute novice will get frustrated early on, but others may dislike the slow pace, especially on a second playthrough. On a higher level, tutorial stages necessitate a plodding narrative, where the story should also reflect the growth in skills the player and his or her avatar have achieved. This can be detrimental to the art of storytelling in games, as almost every game story necessarily becomes a variant of the Hero's Journey (cf. Dunniway, 2000). While adaptation in learning the gameplay basics may not resolve this completely, it could go a long way in mitigating it.

With this research we used Mayer's cognitive theory of multimedia learning as a framework for our experiments, but other frameworks may also lead to interesting observations, for instance from a standpoint of cognitive load and cognitive load theory. Similarly, we focused on learning the triage procedure and how a player creates a mental model from this, but there are many different goals attainable with serious games; from training motor skills to communicative skills and attitudinal change. It would be interesting to see how game design decisions could impact these, and whether this could lead to guidelines as well.

It is furthermore our conviction that all games, whether entertainment or serious, revolve around learning, from the gameplay basics in the beginning, to every obstacle the player faces and needs to learn to overcome. While we found very little effects of our design guidelines on the engaging qualities of our game, it is plausible that the moment it became clear to the player that it was an educational instead of an entertainment game, he or she stopped perceiving it as an entertaining experience per se and rated its engaging merits (or lack thereof) accordingly. It would be interesting to determine how the design guidelines fare in more big-budget entertainment games, and when they are explored more in situ as a commercial off the shelf entertainment product. This is of course contingent on game companies willing to spend R&D money and publish the results, even the ones that retain the null-hypothesis. In any case more crossover between game companies and universities is highly needed, and could lead to valuable new insights in the science of game design.

Finally, while we did not set out to find proof for Mayer's cognitive theory of multimedia learning, in testing guidelines that should target individual cognitive processes from Mayer's theory, we did find effects on the total learning gains after playing our game. Targeting the selection process with auditory guidance cues led to worse post game knowledge structures, targeting the organization process with adaptivity in the complexity presentation led to more efficient learning and targeting the integration process by introduction of surprising events led to improved post game knowledge structures. This shows the viability of using Mayer's theory as a framework for game design interventions and eo ipso the viability of adjusting a serious game's design to the cognitive processes of the player, for the purpose of improving learning from that game. It would also encourage additional research to determine if they can really be considered separate processes when it comes to learning with games, i.e. if targeting the selection of information via game design would only improve the selection process.

References

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86*(2), 303-311. doi:10.1037/0022-0663.86.2.303

Adams, E. (2006). The Designer's Notebook: 'Bad Game Designer, No Twinkie!' VII. *Gamasutra: Features*. Retrieved from: http://www.gamasutra.com/view/feature/2754/the_designers_notebook_bad_game_.php

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory & Cognition*, *9*(4), 422-433. doi:10.3758/BF03197568

Appleton-Knapp, S. L., Bjork, R. A., & Wickens, Th. D. (2005). Examining the spacing effect in advertising: encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research, 32*(2), 266-276. doi:10.1086/432236

Ayres, P., & Paas, F. (2007). Making instructional animations more effective: A cognitive load approach. *Applied Cognitive Psychology*, *700*, 695-700. doi:10.1002/acp

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist*, *7*(2), 85-97. Hogrefe & Huber Publishers. doi:10.1027//1016-9040.7.2.85

Barab, S., Warren S., & Ingram-Goble, A. (2006). *Academic play space: Designing games for education*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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 for the Behavioral and Social Sciences.

Bos, J. E., De Vries, S. C., Van Emmerik, M. L., & Groen, E. L. (2010). The effect of internal and external fields of view on visually induced motion sickness. *Applied ergonomics*, *41*(4), 516-21. doi:10.1016/j.apergo.2009.11.007

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, *16*(2), 153-184. doi:10.1016/0364-0213(92)90022-M

Bostan, B., & Ogut, S. (2009). Game challenges and difficulty levels: lessons learned From RPGs. In *International Simulation and Gaming Association Conference*. Retrieved from http://www.silentblade.com/presentations/Bostan_Ogut_Full_Paper.pdf

Brown, M. L., Newsome, S. L., & Glinert, E. P. (1989). An experiment into the use of auditory cues to reduce visual workload. In K. Bice, & C. Lewis (Eds.) *Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind (CHI '89)* (pp. 339-346). New York: ACM. doi:10.1145/67449.67515

Brunken, R., Plass, J., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53-61. doi:10.1207/S15326985EP3801_7

Burke, J. W., McNeill, M., Charles, D., Morrow, P., Crosbie, J., & McDonough, S. (2009). Serious games for upper limb rehabilitation following stroke. In *2009 Conference in Games and Virtual Worlds for Serious Applications* (pp. 103-110). IEEE. doi:10.1109/VS-GAMES.2009.17

Castlevania [Computer software]. Tokyo, Japan: Konami.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380. doi:10.1037/0033-2909.132.3.354

Chen, C. (1998). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers*, *10*(2), 107-128. doi:10.1016/S0953-5438(98)00015-0

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM, 50*(4), 31-34. doi:10.1145/1232743.1232769

Clark, J., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149-210. Springer Netherlands. doi:10.1007/BF01320076

Clark, R., Yates, K., Early, S., & Moulton, K. (2010). An analysis of the failure of electronic media and discovery-based learning: Evidence for the performance benefits of guided training methods. In K. H. Silber & R. Foshay (Eds.), *Handbook of Training and Improving Workplace Performance, Volume I: Instructional Design and Training Delivery* (pp. 263-297). Somerset, NJ: Wiley.

Conati C., & Manske, M. (2009). Evaluating adaptive feedback in an educational computer game. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Vilhjálmsson (Eds.), *Lecture Notes in Computer Science, 5773, Intelligent Virtual Agents* (pp. 146-158). Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-04380-2_18

Conati, C., & Zhao, X. (2004). Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game. In *Proceedings of the 9th international conference on Intelligent user interfaces* (pp. 6-13). New York, NY, USA: ACM. doi:10.1145/964442.964446

Cook, M. P. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, *90*(6), 1073-1091. doi:10.1002/sce.20164

Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment 6*(2), article 20. doi:10.1145/1371216.1371223

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety*. San Francisco, CA: Jossey-Bass.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper and Row.

Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (1988). *Optimal experience: Psychological studies in flow in consciousness*. Cambridge: CUP.

Day, E. A., Arthur Jr., W., & Gettman, D. (2001). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology*, *86*(5), 1022-1033. doi:10.1037/0022-0663.86.2.303

De Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*(2), 105-113. Lawrence Erlbaum Associates, Inc.

De Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179. doi:10.2307/1170753

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, *38*(2), 105-134. doi:10.1007/s11251-009-9110-0

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Applied Cognitive Psychology, 21*(6), 731–746. doi:10.1002/acp.1346

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction*, *20*(2), 111-122. doi:DOI: 10.1016/j.learninstruc.2009.02.010

Dede, C., Clarke, J., Ketelhut, D. J., Nelson, B., & Bowman, C. (2005, April). *Students' motivation and learning of science in a multi-user virtual environment*. Paper presented at the meeting of the American Educational Research Association, Montréal, Quebec.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162–170.

Dickey, M. D. (2010). Murder on Grimm Isle: The impact of game narrative design in an educational game-based learning environment. *British Journal of Educational Technology*, *42*(3), 456-469. doi:10.1111/j.1467-8535.2009.01032.x

Dondlinger, M. J. (2007). Educational video game design: A review of the literature. *Journal of Applied Educational Technology 4*(1), 21-31.

Driscoll, M. P. (2005). *Psychology of Learning for Instruction,* 3rd edition. Boston: Pearson Education.

Dunniway, T. (2000). Using the Hero's Journey in Games. *Gamasutra Features*. Retrieved from: http://www.gamasutra.com/view/feature/3118/using_the_heros_journey_in_games.php

Entwistle, N., & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education*, *19*(2), 169-194. doi:10.1007/BF00137106

Erlandson, B. E., Nelson, B. C., & Savenye, W. C. (2010). Collaboration modality, cognitive load, and science inquiry learning in virtual inquiry environments. *Educational Technology Research and Development*, *58*(6), 693-710. doi:10.1007/s11423-010-9152-7

Fable 2 [Computer software]. Redmond, WA: Microsoft Game Studios.

Ferstl, E. C., & Kintsch, W. (1999). Learning from text: Structural knowledge assessment in the study of discourse comprehension. In H. Van Oostendorp & S. R. Goldman (Eds.) *The construction of mental representations during reading* (pp. 247-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Full Spectrum Warrior [Computer software]. Agoura Hills, CA: THQ.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*(4), 441-467. doi:10.1177/1046878102238607

Garzotto, F. (2007). Investigating the educational effectiveness of multiplayer online games for children. In *Proceedings of the 6th international conference on Interaction design and children - IDC '07* (pp. 29-36). New York, NY: ACM Press. doi:10.1145/1297277.1297284

Gee, J. P. (2003). *What Video Games Have to Teach us About Learning and Literacy*. New York: Palgrave/Macmillan.

Gee, J. P. (2005). Learning by design: good video games as learning machines. *E-learning and Digital Media*, *2*(1), 5-16.

Gee, J. P., & Shaffer, D. W. (2010). *Looking where the light is bad: Video games and the future of assessment* (Epistemic Games Group Working Paper No. 2010-02). Madison: University of Wisconsin-Madison.

Gillan, D. J., & Cooke, N. M. (2000). Using pathfinder networks to analyze procedural knowledge in interactions with advanced technology. In E. Salas (Ed.), *Human–technology interaction in complex systems* (Vol. 10). Greenwich, CT: JAI.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, *83*, 88–96.

Gomez, R. L., Hadfield, O. D., & Housner, L. D. (1996). Conceptual maps and simulated teaching episodes as indicators of competence in teaching elementary mathematics. *Journal of Educational Psychology*, *88*(3), 572-585. doi:10.1037//0022-0663.88.3.572

Gonzalvo, P., Canas, J. J., & Baja, M.-T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86*, 601-618. doi:10.1037/0022-0663.86.4.601

Graesser, A., Chipman, P., Leeming, F., & Biedenbach, S. (2009). Deep learning, emotions and games. In U. Ritterfield, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects*. New York, NY: Routledge.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371-395. doi:10.1037/0033-295X.101.3.371

Greitzer, F. L., Kuchar, O. A., & Huston, K. (2007). Cognitive science implications for enhancing training effectiveness in a serious gaming context. *Journal on Educational Resources in Computing (JERIC)*, *7*(3), article 2. doi:10.1145/1281320.1281322

Grootjen, M., Neerincx, M. A., & Veltman, J. A. (2006). Cognitive task load in a naval ship control centre: from identification to prediction. *Ergonomics*, *49*(12-13), 1238-1264. doi:10.1080/00140130600612705

Half-Life 2 [Computer software]. Bellevue, WA: Valve Corporation.

Half Life 2: Episode Two [Computer software]. Bellevue, WA: Valve Corporation.

Hays, R. T. (2005). *Effectiveness of instructional games: A literature review and discussion.* Technical Report 2005-004, Naval Air Warfare Center Training Systems Division, Orlando, Florida.

Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*(5), 1084-1102. doi:10.1037/0278-7393.18.5.1084

Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and Instruction, 21*(4), 325-360. doi:10.1207/s1532690xci2104_1

Herrington, J., Oliver, R., & Reeves, T. C. (2003). Patterns of engagement in authentic online learning environments. *Australian Journal of Educational Technology*, *19*(1), 59-71.

Hodgetts, T. J. (2003). Major Incident Medical Training: A Systematic International Approach. *International Journal of Disaster Medicine*, *1*(1), 13-20. doi:10.1080/15031430310013429

Hoeken, H., & Vliet, M. V. (2000). Suspense, curiosity , and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics*, *27*(4), 277-286. doi:10.1016/S0304-422X(99)00021-2

Hustinx, P., Meeuwis, D., & Hermans, R. (2004). *Geneeskundig Management bij Grootschalige Incidenten*. Utrecht: De Tijdstroom.

IJsselsteijn, W. A. (2003). Presence in the past: what can we learn from media history? In G. Riva, F. Davide, & W. A. IJsselsteijn (Eds.), *Being There: Concepts, effects and measurement of user presence in synthetic environments* (pp. 17-40). Amsterdam: IOS Press.

Iwata, S. (2009a, July). *Iwata Asks: Wii Sports Resort*. Retrieved from http://us.wii.com/iwata_asks/wiisportsresort/vol1_page1.jsp

Iwata, S. (2009b, December). *Iwata Asks: New Super Mario Bros*. Wii. Retrieved from http://us.wii.com/iwata_asks/nsmb/vol2_page1.jsp

Jarvis, S., & De Freitas, S. (2009). Evaluation of an immersive learning programme to support triage training. In Rebolledo-Mendez, G., Liarokapis, F. & De Freitas, S. (Eds.), *IEEE Virtual Worlds for Serious Applications, First International Conference* (p117-122). Los Alamitos, CA: IEEE Computer Society. doi:10.1109/VS-GAMES.2009.31

Jenkins, H. (2004). Game design as narrative architecture. In N. Wardrip-Fruin & P. Harrigan (Eds.), *First Person: New Media as Story, Performance, Game*. Cambridge, MA: MIT Press.

Jeung, H.-J., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology, 17*(3), 329-343. doi:10.1080/0144341970170307

Johnson-Laird, P.N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, UK: Cambridge University Press.

Jonassen, D. H. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research and Development*, *39*(3), 5-14. doi:10.1007/BF02296434

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural Knowledge: Techniques for Representing, Conveying, and Acquiring Structural Knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahler, S.E. (2002). *A comparison of knowledge acquisition methods for the elicitation of procedural mental models* (Doctoral dissertation, North Carolina State University). Retrieved from http://repository.lib.ncsu.edu/ir/handle/1840.16/4900

Kallinen, K., Salminen, M., Ravaja, N., Kedzior, R., & Sääksjärvi, M. (2007). Presence and emotion in computer game players during 1st person vs. 3rd person playing view: evidence from self-report, eye-tracking, and facial muscle activity data. In *Proceedings of Presence 2007* (pp. 187-190). Barcelona, Spain: International Society for Presence Research.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*(4), 351–371. doi:10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology, 1,* 5-23.

Kato, P. M., Cole, S. W., Bradlyn, A. S., & Pollock, B. H. (2008). A video game improves behavioral outcomes in adolescents and young adults with cancer: A randomized trial. *Pediatrics*, *122*(2), 305-317. doi:10.1542/peds.2007-3134

Kester, L., Kirschner, P. A., Van Merrienboer, J. J. G., & Baumer, A. (2001). Just-in-time information presentation and the acquisition of complex cognitive skills. *Computers in Human Behavior*, *17*(4), 373-391. doi:10.1016/S0747-5632(01)00011-5

Khaled, R., Barr, P., Biddle, R., Fischer, R., & Noble, J. (2009). Game design strategies for collectivist persuasion. In *Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games - Sandbox '09* (p. 31). New York, New York, USA: ACM Press. doi:10.1145/1581073.1581078.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education*, *8*(1), 13-24. doi:10.1016/j.iheduc.2004.12.001.

Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, *9*(1-3), 87-98. doi:10.1016/0304-422X(80)90013-3

Kintsch W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*. 294-303. doi:10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York: Cambridge University Press.

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*(1), 1-10. doi:10.1016/S0959-4752(01)00014-7

Klimmt, C., & Hartmann, T. (2006). Effectance, self-efficacy, and the motivation to play video games. In P. Vorderer & J. Bryant (Eds.), *Playing video games: Motives, responses, and consequences* (p. 133–145). Mahwah, NJ: Lawrence Erlbaum Associates.

Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., De Freitas, S., et al. (2010). Serious gaming technology in major incident triage training: a pragmatic controlled trial. *Resuscitation*, *81*(9), 1175-1179. doi:10.1016/j.resuscitation.2010.03.042

Kraiger, K., Salas, E., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Factors*, *37*, 804–816.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences 9*(2), 75-82.doi:10.1016/j.tics.2004.12.004

Legend of Zelda [Computer software]. Kyoto, Japan: Nintendo.

Legend of Zelda Ocarina of Time [Computer software]. Kyoto, Japan: Nintendo.

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. doi:10.1162/105474601300343612

Lopes R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: a survey. *IEEE Transactions on Computational Intelligence and AI in Games 3*(2), 85-99. doi:10.1109/TCIAIG.2011.2152841

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education*, *14*(2), 225-244. doi:10.1007/BF03172967

Lowe, R. K. (2003). Animation and learning: selective processing of information in dynamic graphics. *Learning and Instruction 13*(2), 157-176. doi:10.1016/S0959-4752(02)00018-X

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 369-333. doi:10.1207/s15516709cog0504_2

Mania, K., & Chalmers, A. (2001). The effects of levels of immersion on memory and presence in virtual environments: A reality centered approach. *CyberPsychology & Behavior, 4*(2), 247-264. doi:10.1089/109493101300117938

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology, 93*(2)*,* 377-389. doi:10.1037/0022-0663.93.2.377

Mayer, R. E. (1999). Designing instruction for constructivist learning. In C. M. Reigeluth (Ed.), *Instructional-design theories and models, vol. 2* (pp. 141-159). Mahwah, NJ: Erlbaum.

Mayer, R. E. (2001). *Multimedia Learning*. New York: Cambridge University Press.

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, *13*(2), 125-139. doi:10.1016/S0959-4752(02)00016-6

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

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, *63*(8), 760-769. doi:10.1037/0003-066X.63.8.760

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*(2), 312-320. doi:10.1037/0022-0663.90.2.312.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43-52. doi:10.1207/S15326985EP3801_6

McGaugh, J. L. (2006). Make mild moments memorable: add a little arousal. *Trends in cognitive sciences*, *10*(8), 345-347. doi:10.1016/j.tics.2006.06.001

McGrath, D., and Hill, D. (2004). UnrealTriage: A game-based simulation for emergency response. *The Huntsville Simulation Conference.*

McQuiggan, S., Rowe, J., Lee, S., & Lester, J. (2008). Story-based learning: The impact of narrative on learning experiences and outcomes. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 530-539). Berlin, Germany: Springer Berlin Heidelberg. doi:10.1007/978-3-540-69132-7_56

Metroid [Computer software]. Kyoto, Japan: Nintendo.

Metroid Prime 2 [Computer software]. Kyoto, Japan: Nintendo.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications*. London: Springer-Verlag.

Mirror's Edge [Computer software]. Redwood City, CA: Electronic Arts.

Mishra, P., & Foster, A. (2007). The claims of games: A comprehensive review and directions for future research. In R. Carlsen (Ed.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2007* (pp. 2227-2232). Chesapeake, VA: AACE.

Moreno, R., & Mayer, R. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*(2), 358-368. doi:10.1037/0022-0663.91.2.358

Moreno, R. & Mayer, R. E. (2000). A learner-centered approach to multimedia explanations: Deriving instructional design principles from cognitive theory. *Interactive Multimedia Electronic Journal of Computer Enhanced Learning*, *2*. Retrieved from http://imej.wfu.edu

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*(3), 309-326. doi:10.1007/s10648-007-9047-2

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177-213. doi:10.1207/S1532690XCI1902_02

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development*, *53*(3), 35-45. doi:10.1007/BF02504796

Nelson, B. & Erlandson, B. (2008). Managing cognitive load in educational multi-user virtual environments: The River City case study. *Educational Technology Research and Development, 56*(5-6), 619-641. doi:10.1007/s11423-007-9082-1

Nunez, D., & Blake, E. (2003). A direct comparison of presence levels in text-based and graphics-based virtual environments. In *Proceedings of the 2nd international conference on computer graphics, virtual reality, visualisation and interaction in Africa - AFRIGRAPH '03* (pp. 53-56). New York, New York, USA: ACM Press. doi:10.1145/602330.602340

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist 38*(1), 1-4. doi:10.1207/S15326985EP3801_1

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science, 32*(1), 1-8. doi:10.1023/B:TRUC.0000021806.17516.d0

Paivio, A. (1986). *Mental Representations: A Dual-Coding Approach,* New York: Oxford University Press.

Pathfinder [Computer software].

Prensky, M. (2003). Digital game-based learning. *Computers in Entertainment 1*(1), 1-4. doi:10.1145/950566.950596

Prensky, M. (2007). *Digital Game-Based Learning*. St. Paul, MN: Paragon House.

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, *14*(2), 154-166. doi:10.1037/a0019440.

Przybylski, A. K., Ryan, R. M., & Rigby, C. S. (2009). The motivating role of violence in video games. *Personality and Social Psychology Bulletin, 35*(2), 243-259. doi:10.1177/0146167208327216

Radvansky, G. A., & Copeland, D. E. (2001). Working memory and situation model updating. *Memory & Cognition, 29*(8), 1073-1080. doi:10.3758/BF03206375

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature reviews. Neuroscience*, *4*(3), 193-202. doi:10.1038/nrn1052

Ratan, R., & Ritterfeld, U. (2009). Classifying serious games. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects*. New York/London: Routledge.

Rauterberg, M., Gomez-Martin, M., Gomez-Martin, P., & Gonzalez-Calero, P. (2004). Game-driven intelligent tutoring systems. In M. Rauterberg (Ed.) *Entertainment Computing – ICEC 2004, Springer Lecture Notes in Computer Science* (Vol. 3166, pp. 21-39). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-28643-1_14

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., et al. (2007). Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, *10*(1), 45-56.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344-360. doi:10.1007/s11031-006-9051-8

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009). Off-task behavior in narrative-centered learning environments. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. Graesser (Eds.), *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (p. 99–106). Amsterdam: IOS Press.

Russo, R., Mammarella, N., & Avons, S. E. (2002). Toward a unified account of spacing effects in explicit cued-memory tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 819-829. doi:10.1037/0278-7393.28.5.819

Saxby, D. J., Matthews, G., Hitchcock, E., & Warm, J. S. (2007). Development of active and passive fatigue manipulations using a driving simulator. In *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting* (pp. 1237-1241). Santa Monica, CA: Human Factors and Ergonomics Society.

Schuemie, J. S., Van der Straaten, P., Krijn, M., & Van der Mast, C. A. P. G. (2001). *CyberPsychology & Behavior, 4*(2), 183-201. Doi:10.1089/109493101300117884

Schvaneveldt, R. W. (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization.* Norwood, NJ: Ablex.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1985). *Pathfinder: Scaling with network structures* (Memorandum in Computer and Cognitive Science, MCCS-85-9, Computing Research Laboratory). Las Cruses: New Mexico State University.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers & Education*, *46*(3), 223-234. doi:10.1016/j.compedu.2005.11.003

Shaffer, D. W., Squire, K., Halverson, K. R., & Gee, J. P. (2005). Video games and the future of learning. *Phi Kappan Delta*, *87*, 104-111.

Shapiro, M.A. & Pena, J. (2009). Generalizability and validity in digital game research. In U. Ritterfield, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects*. New York, NY: Routledge.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education*, *49*(4), 413-430. doi:10.1007/s10734-004-9448-9

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18*(2), 158-176. doi:10.1521/scpq.18.2.158.21860

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* (pp. 295-321). New York, NY: Routledge.

SimCity 4 [Computer software]. Redwood City, CA: Electronic Arts.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, 489-528. doi:10.1111/j.1744-6570.2011.01190.x

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE) - Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 6*(6), 603-616.

Smets, N. J. J. M., Abbing, M. S., Neerincx, M., Lindenberg, J., & van Oostendorp, H. (2010). Game-based evaluation of personalized support for astronauts in long duration missions. *Acta Astronautica*, *66*(5-6), 810-820.

Source SDK [Computer software]. Bellevue, WA: Valve Corporation.

Squire, K., Barnett, M., Grant, J. M., & Higginbotham, T. (2004). Electromagnetism supercharged!: Learning physics with digital simulation games. In *Proceedings of the 6th international conference on Learning sciences* (pp. 513–520). International Society of the Learning Sciences.

Steiner, K. E., & Voruganti, L. (2004). A comparison of guidance cues in desktop virtual environments. *Virtual Reality*, *7*(3-4), 140-147.

Super Mario Bros [Computer software]. Kyoto, Japan: Nintendo.

Super Mario Galaxy 2 [Computer software]. Kyoto, Japan: Nintendo.

Sweetser P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, *3*(3), 1-24. doi:10.1145/1077246.1077253

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology*, *10*(3), 251-296. doi:10.1023/A:1022193728205

Tabbers, H. K., Martens, R. L., & Van Merriënboer, J. J. G. (2004), Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74*(1), 71–81. doi:10.1348/000709904322848824

Tamborini, R., & Skalski, P. (2006). The role of presence in the experience of electronic games. In P. Vorderer, & J. Bryant (Eds.), Playing *video games: Motives, responses, and consequences* (pp. 225-240). Mahwah, NJ: Lawrence Erlbaum Associates.

The Elder Scrolls 4: Oblivion [Computer software]. Rockville, MD: Bethesda.

The Orange Box [Computer software]. Bellevue, WA: Valve Corporation.

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.

Trumpower, D. L., Sharara, H., & Goldsmith, T. E. (2010). Specificity of structural assessment of knowledge. *Journal of Technology, Learning, and Assessment*, *8*(5). Retrieved from http://www.jtla.org

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*(4), 247-262. doi:10.1006/ijhc.2002.1017

UnrealEngine 2 [Computer software]. Cary, NC: Epic Games.

Van der Spek, E. D., Van Oostendorp, H., Wouters, P., & Aarnoudse, L. (2010). Attentional cueing in serious games. In K. Debattista, M. Dickey, A. Proenca, & L. P. Santos (Eds.), *Second International Conference on Games and Virtual Worlds for Serious Applications VS-GAMES 2010* (pp. 119-125). Los Alamitos, CA: IEEE.doi: 10.1109/VS-GAMES.2010.8

Van der Spek, E. D., Wouters, P., & Van Oostendorp, H. (2011). Code Red: Triage or COgnition-based DEsign Rules Enhancing Decisionmaking TRaining In A Game Environment. *British Journal of Educational Technology*, *42*(3), 441-455. doi:10.1111/j.1467-8535.2009.01021.x

Van Dijk, V., Van der Spek, E. D., Van Oostendorp, H. & Erkens, G. (2011). Serious games versus traditional instruction; knowledge construction, retention and cognitive load. *Unpublished manuscript, Utrecht University*.

Van Nimwegen, C., Burgos, D., Van Oostendorp, H., & Schijf, H. J. M. (2006). The paradox of the assisted user: guidance can be counterproductive. In R. Grinter, Th. Rodden, P. Aoki. E. Cutrell, R. Jeffries, & G. Olson (Eds.), *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI '06).* (pp. 917-926). New York: ACM. doi:10.1145/1124772.1124908

Van Nimwegen, C., & Van Oostendorp, H. (2009). The questionable impact of an assisting interface on performance in transfer situations. *International Journal of Industrial Ergonomics, 39*(3), 501-508. doi:10.1016/j.ergon.2008.10.008

Van Oostendorp, H., & Beijersbergen, M.J. (2007). Animated diagrams: their effect on understanding, confidence and mental effort. In *Proceedings of the 12th European Association of Learning and Instruction (EARLI) Conference*. Budapest, Hungary.

Van Oostendorp, H., Beijersbergen, M. J., & Solaimani, S. (2008). Conditions for learning from animations. In *Proceedings of the 8th international conference on International conference for the learning sciences - Volume 2* (ICLS'08) (pp. 438-445), International Society of the Learning Sciences.

Van Oostendorp, H., & Bonebakker, C. (1999) Difficulties in updating mental representations during reading news reports. In H. van Oostendorp & S.R. Goldman (Eds.), *The Construction of Mental Representations during Reading*. Mahwah, NJ: Lawrence Erlbaum Associates.

Van Oostendorp, H., Otero, J., & Campanario, J. (2002). Conditions of updating during reading. In M. Louwerse, & W. van Peer (Eds.), *Thematics: Interdisciplinary Studies*. Amsterdam/New York: John Benjamins Publishing Company.

Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC medical research methodology*, *5*, 35-46. doi:10.1186/1471-2288-5-35

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*, 229–243.

Webster, J., Trevino, L. K., & Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. *Computers in Human Behavior, 9*(4), 411-426. doi:10.1016/0747-5632(93)90032-N

Wenden, A. L. (1998). Metacognitive knowledge and language learning. *Applied Linguistics, 19*(4), 515-537. doi:10.1093/applin/19.4.515

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators & Virtual Environments*, *7*(3), 225-240. doi:10.1162/105474698565686

Wong, W. L., Shen, C., Nocera, L., Carriazo, E., Tang, F., Bugga, S., et al. (2007). Serious video game effectiveness. In *Proceedings of the international conference on Advances in computer entertainment technology - ACE ’07* (p. 49). New York, New York, USA: ACM Press. doi:10.1145/1255047.1255057

Wouters, P., Van der Spek, E. D., & Van Oostendorp, H. (2009). Current practices in serious game research: a review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games-based Learning Advancements for Multisensory Human Computer Interfaces: Techniques and Effective Practices* (p. 232–255). Hershey, PA: IGI Global. doi:10.4018/978-1-60566-360-9.ch014

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, OnlineFirst*, 1-23. doi:10.1007/s11423-010-9183-0

Wouters, P., Van Oostendorp, H., Boonekamp, R., & Van der Spek, E. D. (2011). The role of Game Discourse Analysis and curiosity in creating engaging and effective serious games by implementing a back story and foreshadowing. *Interacting with Computers*, *23*(4), 329-336. Elsevier B.V. doi:10.1016/j.intcom.2011.05.001

Yun, C., Shastri, D., Pavlidis, I., & Deng, Z. (2009). O' game, can you feel my frustration?: Improving user's gaming experience via stresscam. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI '09)* (pp. 2195-2204). New York: ACM. doi:10.1145/1518701.1519036

Zwaan, R. A., Langston, M. C. & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292-297. doi:10.1111/j.1467-9280.1995.tb00513.x

Appendix A

Experiment handout

Name	
Address (streetname + nr)	
City (countrycode)	

Gender: Male / Female (delete as applicable)

Age in years:

For the next questions, tick the box most applicable to your situation:

How would you describe yourself?
□ I (almost) never play videogames
□ I occasionally play videogames
□ I'm a gamer

How much time, on average, do you spend playing videogames weekly?
□ none
□ 0 – 2 hours
□ 2 – 4 hours
□ 4 – 6 hours
□ more than 6 hours

Do you suffer from color vision impairment?   Yes / No

Pre-knowledge test: Encircle the letter with the correct answer

1. During the primary triage of a victim, one first checks:
a. Whether the airway is free
b. If the victim is able to walk
c. If the victim has external wounds
d. The heart rate

2. A victim whose breathing frequency is too high will, according to the primary triage procedure, likely fall in the following triage category:
a. T1
b. T2
c. T3
d. Dead

3. On a Friday afternoon a severe accident took place on a brightly lit provincial highway. You want to check the blood circulation of a victim. What is the best method according to the procedure of the primary triage in this case:
a. The capillary refill time
b. Measuring the blood pressure
c. Measuring the pulse rate
d. None of these methods are suitable for this case

4. A victim has a free airway. The airway was already free without any intervention from you. What is the next step according to the procedure of the primary triage:
a. Checking the capillary refill time
b. Checking the breathing frequency
c. Placing the victim in the stable recovery position
d. Measuring the blood pressure

5. While triaging a mature victim with a blocked airway, you have a strong suspicion of neck trauma. What is the best treatment according to the primary triage:
a. Apply a chin lift
b. Place the victim in a stable recovery position
c. Apply a jaw thrust
d. Apply a neck brace

6. Every triage category has its own color. The color for triage category T2 is:
a. Blue
b. Red
c. Green
d. Yellow

7. Study the following statements:

   I. A victim whose blocked airway has been freed but still doesn't breathe, falls into triage category T1
   II. The CRT is measured via the nail bed

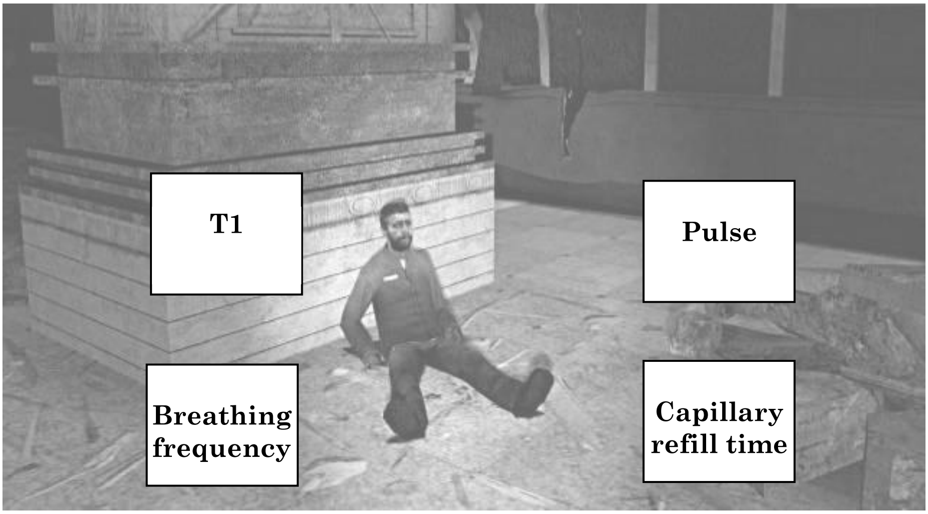
Which of these statements is correct:
a. Both I and II are correct
b. Both I and II are incorrect
c. Only I is correct
d. Only II is correct

8. At the site of a multi-vehicle collision, you encounter a victim with two broken legs. The victim can breathe freely, has a normal breathing frequency and a quick capillary refill time. In which triage category would you categorize the victim:
a. T1
b. T2
c. T3
d. Dead

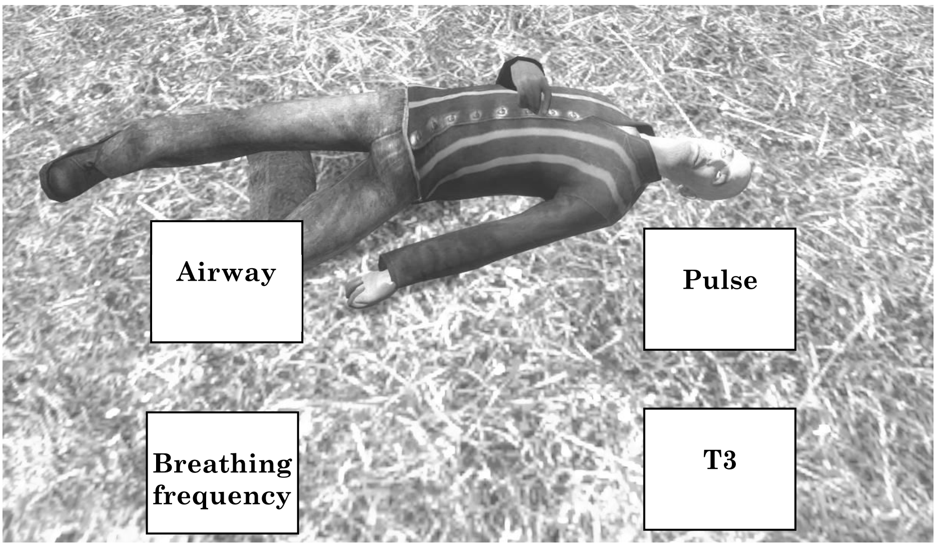
Visual items:

For each of the following items, check the box with the action that should be performed next, according to the procedure of the primary triage.

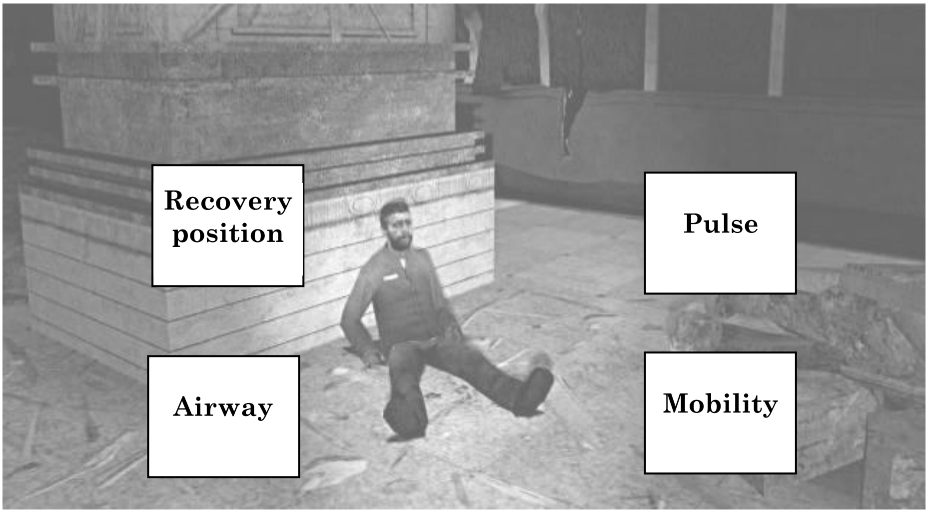
9. Male, 27 years old. It's very dark. The man can't walk and has a free airway.



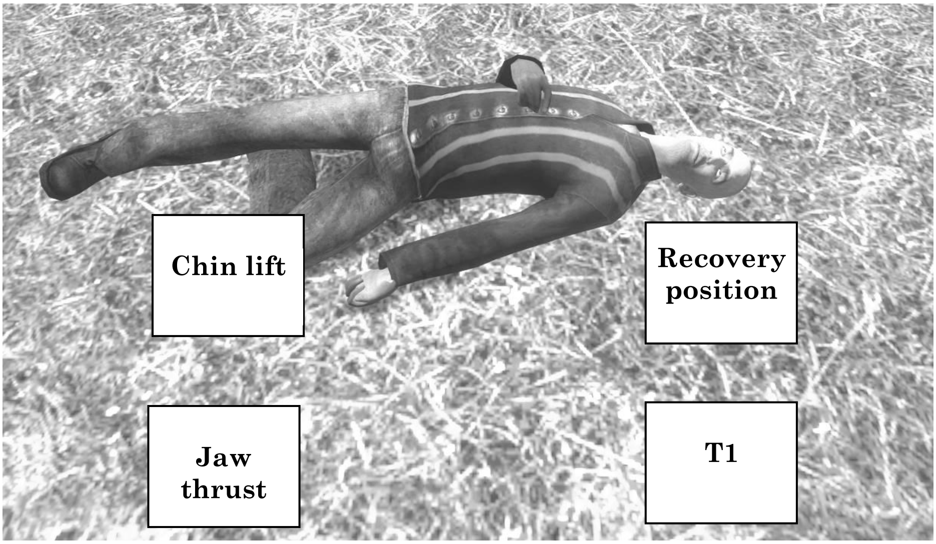
10. Male, 46 years old. There's enough daylight for a triage. The man is slightly dizzy but says he's still able to walk.



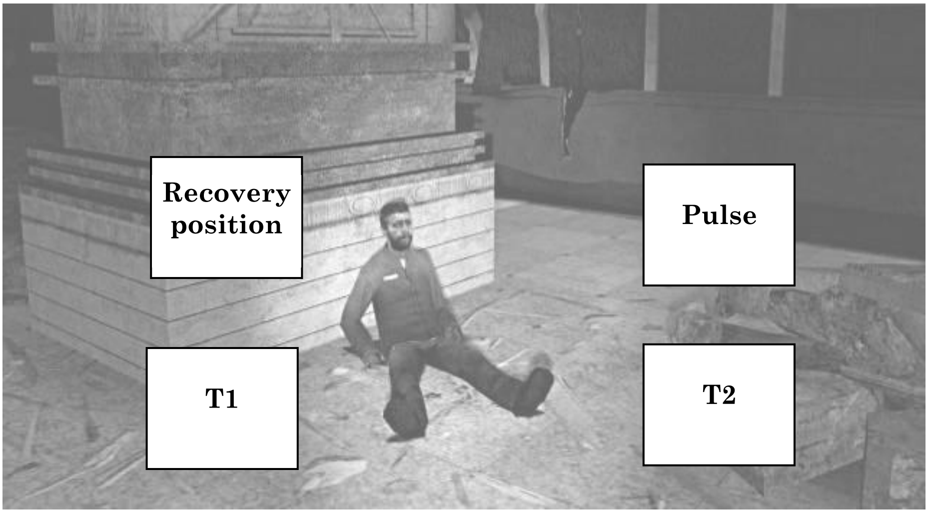
11. Male, 27 years old. Enough light.



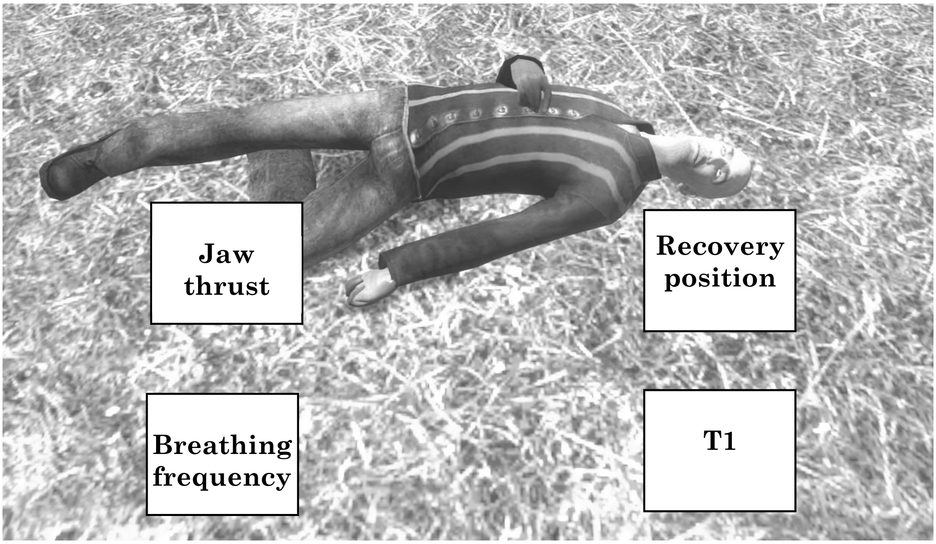
12. Male, 46 years old. There's enough daylight for the triage procedure. The man indicated that he's unable to walk. You also discovered a blocked airway.



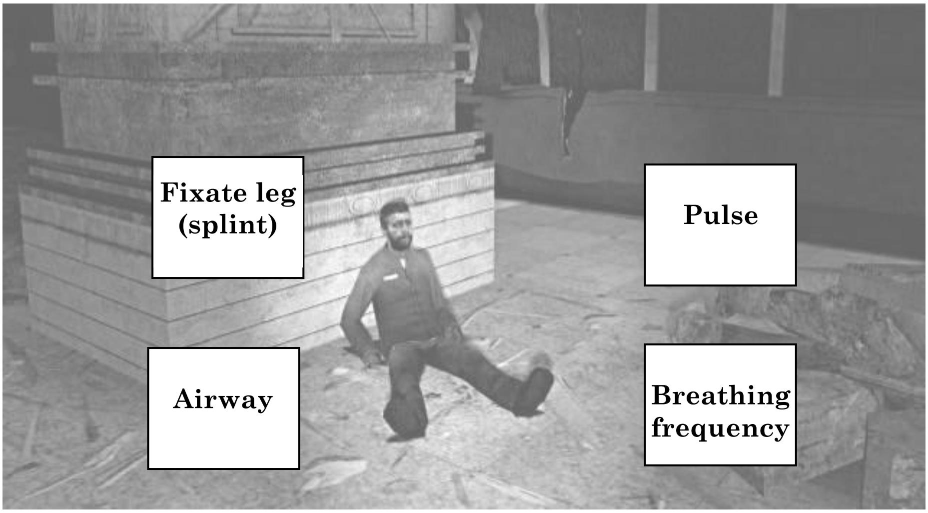
13. Male, 40 years old. Measuring the capillary refill time just failed because it was too dark.



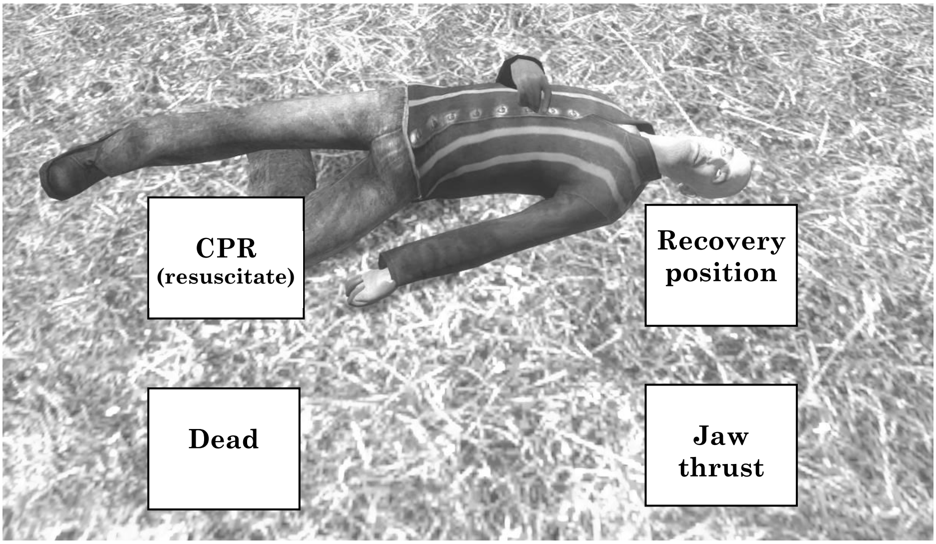
14. Male, 51 years old. There's enough daylight for a triage. The airway was blocked, but you just freed it, and the victim now breathes normally.



15. Male, 27 years old. Enough light. The victim is unable to walk due to a bone fracture in the leg.



16. Male, 30 years old. There's enough daylight for a triage. You discovered a blocked airway, but after freeing this, the victim didn't start breathing.



The triage of victims

When a large accident has happened, creating many injured victims, the medical first responder that first arrives at the scene has to determine which victims are in need of medical attention most urgently. During this categorization, commonly known as the primary triage, the victims have to be categorized into one of four triage categories. These categories are **T1**: very urgent; **T2**: reasonably urgent; **T3**: not urgent; and **Dead**. The primary triage takes place at the spot where the victims are found. Furthermore, this has to be done as quickly as possible.

Instructions on the game Code Red Triage

In the game Code Red Triage, you will learn the procedure of the primary triage. The procedure is the same for all victims, and always starts with the same step. It will be your duty to try to figure out how the procedure goes.

The goal of the game is to figure out the procedure of the primary triage and subsequently practice it. It's important that you try to perform the procedure as quickly and correctly as possible.

The game starts off in the hall of a trainstation. There, you will receive information on where to find the victims. To get to the victims you will first have to pass through the corridors of a subway station. You will get instructions on this in the game. This part is intended to let you practice navigating in the game, you can use the following keys:

Key	
W	**move forward**
A	**move left**
D	**move right**
S	**move backward**

You can control your direction with the mouse. While walking through the corridors, a new area has to be loaded twice; this will take a few seconds.

You will have 17 minutes to categorize all the victims in a triage category, starting from the moment you encounter the first victim. **When you reach a victim, you can start the triage procedure by pressing the *e-button.*** An interface will then appear, with a number of buttons on either side (see figure below). In the beginning you will only have a few buttons, but more will be introduced later on. You will see some introductory information on the victim first (this is not part of the triage). With the buttons you can then proceed to get extra information (this is). Every button will return general information on what it means, and the corresponding information specific to that victim.
For instance, the button 'Airway' could, as general information, return 'You determine if the airway is free and the victim can breathe freely, the victim-

specific information could then be 'the victim is able to breathe freely'. In the interface these two bits of information are separated by a blank space. The general information is the same for every victim, but does give hints on how the triage procedure goes.

Because measuring some of the values (such as the pulse) takes some time in real life, this will not appear immediately but requires you to wait a short period of time.

Not all buttons are always necessary to categorize the victims in a triage category. Which button is applicable depends on the triage procedure and the state the victim is in. Time is a crucial factor with the primary triage. Selecting a button that is unnecessary will cost you time. Therefore, it is important to discern the correct procedure and subsequently only use the buttons that are necessary.

When you think you've gathered enough information, you can select the correct triage category with the four buttons on the bottom of the screen. The victim will then turn the same hue as the triage category (a T3 victim will thus turn green), after which the triage screen closes and you can move through the virtual environment again.

Shortly after closing the triage screen, you will see the score of your triage classification. For placing the victim in the right triage category you will be awarded 100 points, but points can be deducted for:

a. Following the wrong sequence of steps
b. Forgetting necessary steps
c. Taking unnecessary steps, and/or
d. Surpassing the allotted time.

Every time u categorize a victim, you will see your score and how well you did on these items.

In addition, the total score at the top of the screen will change. You can then proceed to the next victim.

So remember to think before clicking a button, to get a high score! And also remember: sometimes you need different buttons than before, **but the sequence always remains the same and you always have to start with the same button!**

If the instructions on the game are clear, you can ask the experiment leader to start the game. If you still have questions you are free to ask them.

Questionnaire on how you experienced the game

(Please be as honest as possible! It's science, not politics)

On a scale of 1 to 10, rate how difficult you found it to achieve a good score in the game.

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

On a scale of 1 to 10, rate how much you enjoyed the game.

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

The following questionnaire consists of a number of statements, which are followed by a score of 1 to 5. Here you should rate to which degree you agree with the statement, where 1 stands for completely disagree, and 5 stands for completely agree. Tick the box corresponding to the score you would give.

		Strongly disagree	*Disagree*	*Neither agree nor disagree*	*agree*	*Strongly agree*
1.	I felt sad that my experience was over	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
2.	I had a sense that I had returned from a journey	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
3.	I would have liked the experience to continue	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
4.	I vividly remember some parts of the experience	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
5.	I felt myself being ‘drawn in’.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
6.	I felt involved in the game world	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
7.	I lost track of time	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
8.	I enjoyed myself.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
9.	My experience was intense	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
10.	I paid more attention to the displayed environment than I did to my own thoughts (e.g., daydreams etc.)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
11.	I responded emotionally	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
12.	The content appealed to me	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Post-knowledge test

NB the post-knowledge test was the same as the pre-knowledge test, but with the questions presented in a different order.

And finally these short questions:

1. The first step in the triage procedure is:
a. Mobility
b. Airway
c. Pulse
d. CRT

2. If a victim's airway is blocked and he/she does not have neck trauma, you must:
a. Put him/her in a stable recovery position
b. Apply a chin lift
c. Apply a jaw thrust
d. Insert a breathing tube

3. If a victim's airway is blocked and he/she does have neck trauma, you must:
a. Put him/her in a stable recovery position
b. Apply a chin lift
c. Apply a jaw thrust
d. Insert a breathing tube

4. If it is light, the preferred method of measuring blood circulation is:
a. Capillary refill time
b. Blood pressure
c. Pulse rate
d. Thermometer

5. If it is dark, the preferred method of measuring blood circulation is:
a. Capillary refill time
b. Blood pressure
c. Pulse rate
d. Thermometer

This concludes our experiment, thank you for your cooperation!

Summaries

English summary

Games for learning, so-called serious games, have a great potential to engender learning in an engaging manner, but corroborating evidence of their efficacy and engagement has so far been underwhelming (chapter 1). If serious games demonstrate problems in becoming efficacious it is also unclear why this is the case because of a lack of systematic empirical research in what constitutes good serious game design, and the manner in which instruction should best be integrated in the game design. Furthermore, good instructional design may often be at odds with good game design, where deeper learning could lead to less engagement (Graesser, Chipman, Leeman & Biedenbach, 2009). We consequently set out to test a number of serious game design principles to see if implementing or altering these would improve learning without harming the engaging qualities of a serious game.

We used Mayer's cognitive theory of multimedia learning (1999; 2005; chapter 2), which states that learning from multimedia involves *selecting* relevant from irrelevant information, *organizing* the information into a coherent knowledge structure and *integrating* this structure into prior cognitive knowledge structures, both as a theory of learning in a game and as a framework to base our experiments on. Experiments were performed on game design principles that would hypothetically improve the cognitive processes responsible for the selection, organization and integration of information, and thereby the efficacy of the serious game. Learning from a serious game was assessed with three methods: a paper knowledge test that measured the declarative and procedural knowledge that was gained, a structural knowledge assessment that measured how well the knowledge was organized structurally and which was indicative of deep learning, and the in-game score. Engagement was measured with a subset of the ITC-SOPI Engagement questionnaire (Lessiter, Freeman, Keogh & Davidoff, 2001).

In order to test our game design techniques we created a game called *Code Red Triage* (chapter 3), which taught and trained players how to perform a triage procedure; categorizing the victims of a mass casualty incident according to urgency of needed medical attention. The game, a total conversion mod of *Half Life 2*, let the player take on the role of a medical first responder that arrived in a train station; here, the player had to navigate towards a subway platform where an explosion had taken place and nineteen victims needed to be triaged. The game was a 3D first person game played from the eyes of the protagonist. When a player arrived at a victim, a triage menu was conjured up where the player could choose from eight different triage action buttons and categorize the victim into one of four triage categories. Subsequently, the game gave feedback in the form of a game score and short statements whether they forgot an action, took one that was unnecessary and whether it was done in the right order. From this the player had to infer the correct procedure of the triage, by selecting the right buttons pertaining to a certain victim case, organizing the actions into the correct order and integrating this ordered information with knowledge that was attained by triaging previous victims. Code Red Triage was evaluated in a pilot with medical experts and laymen and found to be successful in conveying the triage procedure to both groups. As the learning gains were larger for laymen, this made them a suitable group for testing our experimental interventions.

When we compared our game with a static PowerPoint presentation of the same instructional information (chapter 4), we discovered that the game condition exerted a higher cognitive load on the user than the PowerPoint, likely due to the user having to actively select and organize the information, and that this in turn was negatively correlated with the learning gains after engaging in the instructional material. This difference between conditions does disappear after a week.

Four game design principles were subsequently empirically tested to determine whether implementing these could improve learning while retaining the same engaging qualities of a game. These were *guidance cues* to improve the selection process, the manner of *complexity presentation* to improve the effectiveness and *adaptation to the player's performance* to improve the efficiency of the organization process, and the *introduction of surprising events* to improve the integration process of multimedia learning.

In the first experiment (chapter 5) the selection process was targeted, where a player of a serious game has to mentally select the relevant information while playing a game and discard the irrelevant information, by the introduction of guidance cues. There were two different forms of guidance cues that were tested, namely auditory cues and visual cues, which were compared with a control group. There were a number of reasons why we contrasted these different types of cues. Firstly, because visual cues, while proven to be effective in other instructional multimedia (De Koning, Tabbers, Rikers & Paas, 2007) and therefore a likely candidate for improving the cognitive selection process, could also come off as paternalistic in a game. As games are frequently enjoyed because they offer the player a sense of autonomy on how to tackle problems, visual cues may subsequently lead to lower enjoyment. The salient, anomalous nature of visual cues could furthermore decrease feelings of presence, which can lead to less enjoyment and also interfere with the learning goals. Secondly, the use of auditory instead of visual cues could offload the guidance information into another sensory channel and should therefore further mitigate problems with cognitive load (Mayer, 1999).

Contrary to our expectations, the auditory cueing condition led to significantly worse learning, compared with the control and visual cueing condition. The results do not conclusively confirm whether this is due to the learner picking up the auditory cues correctly, but subsequently failing to internalize the procedure; or whether the auditory cues are not picked up correctly and therefore distract the learner during learner. The visual cueing condition performed significantly better than the auditory condition, but not the control condition. A significant effect of prior game experience was noted in the visual cueing condition on the in-game score and post-game learning gains, where participants that play games performed better than those that didn't. As this did not manifest itself in the other conditions, we can say, with some degree of confidence, that guidance cues in a game only work when the player has prior experience with games or the intention of the cues is clear to the player. Engagement was not affected by either of the type of cues.

In the second experiment (chapter 6a), we investigated the organization process by varying the way in which the information complexity was presented and measuring its subsequent effect on learning and engagement. In entertainment games it is common practice that games gradually become more complex, both in terms of the obstacles or problems a player faces and the options which the player has at his or her disposal to resolve these. As entertainment games have embedded instruction as well, and are still fun to play, Gee (2005) argues that this game design paradigm is a good way to engender learning from a game. From an instructional science standpoint, it can be considered a just-in-time information presentation of the options, and a progressive, massed presentation of the problems. Such a progressive increase in complexity should mitigate the load on the working memory and improve information organization. There are, however, arguments to be made why an opposite approach could also lead to better learning. A spaced presentation of units of instruction, where for instance problem A is presented first, then a dissimilar problem B is presented before problem A is presented again, has been shown to lead to better learning. Similarly, just-in-time presentation may leave the organization process mainly external, leading to less active processing or internalization.

We contrasted two variables, option complexity and problem complexity, which was operationalized in our game as the number of buttons presented to the player and the complexity of the victim cases, for a total of four conditions. No significant effect of the variables was found on any of the measures for learning. However, the just-in-case option presentation with a progressive, massed presentation of problems was enjoyed significantly more than the other conditions. This likely corroborates the work of Przybylski, Rigby and Ryan (2010), in that autonomy and competence may be important for enjoyment of a game.

In another experiment revolving around the presentation of complexity and consequently the organization of information (chapter 6b), we investigated whether the efficiency of serious games could be improved by adapting the complexity presentation to the performance of the player. A version of Code Red Triage with a progressive increase in complexity was altered to automatically proceed to a more complex victim when the player scored above a certain threshold in the game. This was compared to a control condition with no adaptivity.

The two conditions showed a similar effectiveness in engendering learning, as measured in total learning gains, but the participants in the adaptive learning condition did so with significantly less victims triaged and in a shorter time period. Consequently, the adaptive version of the game showed a significantly higher learning gain per victim, both in terms of the knowledge test and structural knowledge assessment, leading to an overall higher efficiency of the serious game. The adaptivity engine did not impact the affective responses to the game such as engagement or difficulty rating.

In the fourth and final experiment (chapter 7), we tested whether the integration process of learning with serious games could be stimulated by introducing surprising events in the game narrative. The underlying assumption was that people regularly overlook new information that deviates from what was understood previously. Because a surprising event constitutes an experience that is by its nature unexpected and therefore unpredicted by a person's prior mental model, this should lead to a better update of the mental model. Due to the surprising event, the mental model is activated from the long-term memory stronger, which in turn can facilitate better integration of the new information into the existing knowledge structure.

In line with the already existing environmental narrative of Code Red Triage, we implemented three surprising events in the game. These events were presented at the moment a player had to learn new information that to some extent contrasted prior notions of the player. The participants in the surprising events condition had significantly superior knowledge structures, as assessed by a structural knowledge assessment, after playing the game compared with the control condition, confirming that a relatively simple intervention can lead to deeper learning and more effective serious games. It did not lead to better scores on the knowledge test with indirect questions, nor on the test with questions directly related to the surprising events, but this could be due to a ceiling effect. It also did not have an effect on the reported engagement.

Concluding (chapter 8), we set out to discover if the efficacy of serious games could be ameliorated by discerning problems in the constituent parts of learning with a serious game and systematically target these for improvement, thereby coming to a set of design guidelines for serious game developers. Of the four experiments, two succeeded in markedly improving the efficacy of our serious game. From this we can therefore conclude that it is advisable to do two things. Firstly, include an adaptation engine based on an in-game performance measure, to speed up learning in order to improve the efficiency of a game. Secondly, introduce surprising events around key moments when something should be learned that contrasts prior notions, in order to stimulate deeper learning. Using auditory cues seems ill-advised, although more research is needed to state this conclusively.

We set out to provide guidelines that improved learning while simultaneously not harming the engaging qualities of the game. This greatly succeeded. Only one of our experiments showed a significant effect of the intervention on engagement ratings. The validity is compounded slightly however by the fact that the participants lacked a clear referent to judge their version of the game by. The results from the second experiment seem to corroborate the findings of Przybylski and colleagues (2010), in that games that engender a higher sense of autonomy are enjoyed more; or vice versa, that games that take autonomy away will be enjoyed less.

Finally, we did not set out to find proof for Mayer's cognitive theory of multimedia learning. However, in testing guidelines that should target individual cognitive processes from Mayer's theory, we did find effects on the total learning gains after playing our game. Targeting the selection process with auditory guidance cues led to worse post game knowledge structures, targeting the organization process with adaptivity in the complexity presentation led to more efficient learning and targeting the integration process by introduction of surprising events led to improved post game knowledge structures. This shows the viability of using Mayer's theory as a framework for game design interventions. It would also encourage additional research to determine if they can really be considered separate processes.

Nederlandse samenvatting

Games die bedoeld zijn om iets te leren, zogenaamde serious games, hebben veel potentie om mensen iets leren en dit ook nog eens leuk en immersief te maken, maar ondersteunend bewijs voor zowel de doeltreffendheid met betrekking tot leren als het leuke en immersieve is vooralsnog weinig overtuigend (hoofdstuk 1). Van de serious games die moeite hebben om doeltreffend te worden, is het daarnaast onduidelijk waardoor dit komt, vanwege een gebrek aan empirisch onderzoek naar het design van serious games en de wijze waarop instructie hierin verweven kan worden. Daarbovenop kan een goed instructioneel ontwerp van een serious game vaak tegelijkertijd een slecht game design betekenen, aangezien dieper leren kan resulteren in minder uitdaging (Graesser, Chipman, Leeman & Biedenbach, 2009). Ons doel was daarom om verscheidene game design principes te testen, en om uit te vinden of het implementeren of veranderen hiervan tot betere leerresultaten zou leiden, zonder dat het de serious game minder leuk of immersief maakt.

Voor het onderzoek hebben we gebruik gemaakt van Mayer's cognitieve theorie van multimedia leren (Mayer, 1999; 2005; hoofdstuk 2), zowel als theorie voor het leren in een game, als een raamwerk om onze experimenten aan op te hangen. Deze theorie stelt dat het leren van multimedia het *selecteren* van relevante over irrelevante informatie, het *organiseren* van de informatie in een coherente kennisstructuur en het *integreren* van de structuur in reeds eerder opgeslagen kennisstructuren behelst.. Experimenten zijn uitgevoerd op de game design principes die hypothetisch de cognitieve processen verbeteren die zorgen voor de selectie, organisatie en integratie van informatie, en zodoende ook de doeltreffendheid van de game zouden moeten verbeteren. Het leren van de game werd getoetst via drie methodes: een papieren kennistest die de opgedane declaratieve en procedurele kennis meette, de structurelekennistoets die meette hoe goed de kennis structureel georganiseerd was, en de in-game score. De mate waarin het leuk en immersief werd bevonden werd gemeten via een subset van de ITC-SOPI vragenlijst (Lessiter, Freeman, Keogh & Davidoff, 2001).

Om de game design technieken te toetsen hebben we een game gemaakt, genaamd Code Red Triage (hoofdstuk 3), welke als doel heeft om de speler de triage procedure te leren; het categoriseren van slachtoffers bij een grootschalig ongeluk naar gelang de snelheid waarmee ze medische hulp nodig hebben. Het spel, een zogenaamde 'total conversion mod' van Half-Life 2, zet de speler in de rol van ambulancebroeder die aankomt op een treinstation. Van hier moest de speler de weg zoeken naar een metroperron waar een explosie heeft plaatsgevonden en negentien slachtoffers getriëerd dienen te worden. De game was 3D en werd getoond vanuit de ogen van de hoofdrolspeler. Zodra de speler bij een slachtoffer arriveerde kwam een triagescherm tevoorschijn waar hij of zij uit acht verschillende triage-acties kon kiezen en het slachtoffer in een van vier triage categorieën in kon delen. De game gaf vervolgens feedback over de keuze van de speler in de vorm van een in-game score en korte opmerkingen over of men een actie was vergeten, een overbodige actie had gedaan, en of de acties in de juiste volgorde waren gedaan. Aan de hand hiervan moest de speler herleiden hoe de triage procedure ging, door de juiste acties te selecteren die behoren bij een bepaalde slachtoffercasus, deze in de goede volgorde te organiseren en de geordende informatie vervolgens weer te integreren met informatie die was verkregen tijdens het triëren van vorige slachtoffers. Code Red Triage was geëvalueerd in een pilot met medische experts en leken, waaruit bleek dat de game succesvol was in het overbrengen van de triage procedure aan beide groepen. Omdat de leerwinsten groter waren in de groep met de leken, zijn de experimentele interventies hieropvolgend getest met leken.

Toen we onze game vergeleken met een PowerPoint presentatie met hetzelfde instructionele materiaal (hoofdstuk 4), bleek dat de game een hogere cognitieve belasting uitoefende op de gebruiker dan de PowerPoint, waarschijnlijk doordat de gebruiker zelf actief de informatie moest selecteren en organiseren, en dat dit vervolgens negatief gecorreleerd was met de leerwinsten na interactie met het instructionele materiaal. Dit verschil tussen de condities verdween na een week.

Vier game design principes zijn vervolgens empirisch getest om te bepalen of ze het leren van een game kunnen verbeteren, zonder de leukheid en immersiviteit van de game te schaden. Dit waren *guidance cues* om het selectieproces te verbeteren, de manier waarop *presentatie van complexiteit* plaatsvindt om de effectiviteit en *adaptatie aan de speler's prestaties* om de efficientie van het organisatieproces te verbeteren, en de *introductie van verrassende gebeurtenissen* om het integratieproces te stimuleren tijdens het leren van serious games.

In het eerste experiment (hoofdstuk 5), werd getracht het selectieproces te verbeteren, waarbij een speler tijdens het spelen van een game mentaal de relevante informatie moet selecteren en irrelevante informatie moet negeren, met behulp van guidance cues. Twee verschillende vormen van guidance cues werden getest, namelijk auditieve cues en visuele cues, welke werden vergeleken met een controle groep. Er waren een aantal redenen waarom we deze verschillende vormen van cues contrasteerden. Ten eerste, omdat visuele cues, hoewel aangetoond is in andere instructionele multimedia (De Koning, Tabbers, Rikers & Paas, 2007) dat deze effectief waren en daarom een geschikte kandidaat waren om het cognitieve selectieproces te verbeteren, ook als paternalistisch over kunnen komen in een game. Aangezien games vaak leuk gevonden worden omdat het de speler een gevoel van autonomie verschaft om problemen aan te pakken, kunnen visuele cues daarom tot een minder leuke game-ervaring leiden. Het opvallende en onnatuurlijke karakter van visuele cues kunnen daarnaast gevoelens van ‘presence’ verminderen, en zodoende zowel de game minder leuk maken als interfereren met de leerdoelen. Ten tweede, omdat het gebruik van auditieve cues in plaats van visuele cues de ondersteunende informatie via een ander zintuiglijk kanaal aanbiedt en zodoende het visuele systeem minder cognitief belast (Mayer, 1999).

In tegenstelling tot onze verwachtingen, leidden de auditieve cues tot significant slechter leren in vergelijking met de controle en visuele cueing conditie. Uit de resultaten kan niet met zekerheid afgeleid worden of dit komt doordat de speler de cues correct oppikte maar als gevolg hiervan de stof niet goed internaliseerde, of dat de cues incorrect zijn opgepikt en daarom de speler afleidde. De visuele cueing conditie presteerde significant beter dan de auditieve conditie, maar niet beter dan de controle conditie. Een significant effect van eerdere game-ervaring werd gevonden in de visuele conditie op de in-game score en de post-game leerwinsten, waarbij participanten met eerdere game-ervaring beter presteerden dan participanten zonder ervaring. Aangezien dit effect niet plaatsvond in de andere condities, kunnen we, met enige zekerheid, zeggen dat (visuele) guidance cues in een game alleen werken wanneer de speler eerdere game-ervaring heeft of de intentie van de cues duidelijk is voor de speler. Plezier en immersiviteit werden niet beïnvloed door de cues.

In het tweede experiment (hoofdstuk 6a) hebben we het organisatieproces onderzocht door de manier te variëren waarop de complexiteit van informatie werd gepresenteerd, en het effect van deze variatie te meten op leren en plezier en immersiviteit. In entertainment games is het gebruikelijk dat games gradueel complexer worden, zowel in termen van de obstakels die de speler tegenkomt, als de opties die hij of zij ter beschikking heeft om deze te overwinnen. Aangezien entertainment games ook ingebedde instructie bevatten en nog steeds leuk zijn om te spelen, betoogt Gee (2005) dat dit game design paradigma een goede manier is om mensen te laten leren in een game. Vanuit het oogpunt van de instructiewetenschap kan dit gezien worden als 'just-in-time' informatie presentatie van de opties en een progressieve, 'massed', presentatie van problemen. Zo’n progressieve toename van complexiteit zou problemen met de belasting van het werkgeheugen kunnen verhelpen en de informatie-organisatie verbeteren. Er zijn echter ook argumenten te bedenken waarom een tegenovergestelde aanpak ook tot beter leren zou kunnen leiden. Voor een 'spaced' presentatie van instructie-eenheden waar, bijvoorbeeld, probleem A als eerste wordt gepresenteerd, dan een ander probleem B, voordat vervolgens probleem A weer ter sprake komt, is aangetoond dat het tot beter leren kan leiden. Just-in-time informatie kan daarnaast de organisatie van informatie geëxternaliseerd laten, hetgeen weer leidt tot minder actieve verwerking of internalisatie.

We contrasteerden twee variabelen, optiecomplexiteit en probleemcomplexiteit, geoperationaliseerd in onze game als de hoeveelheid knoppen die aan de speler gepresenteerd worden en de complexiteit van de slachtoffercasussen; in totaal vier condities. Op geen enkele kennistoets werd een significant effect van de variabelen gevonden. Daarentegen werd de just-in-case optiepresentatie in combinatie met de progressieve massed probleempresentatie wel significant leuker gevonden dan de andere condities. Deze bevinding onderbouwt waarschijnlijk eerder werk van Przybylski, Rigby en Ryan (2010), in die zin dat autonomie en competentie belangrijk zijn voor de leukheid van een game.

In een ander experiment met betrekking tot de presentatie van complexiteit en derhalve de organisatie van informatie (hoofdstuk 6b), hebben we onderzocht of de efficiëntie van serious games verbeterd kan worden door de complexiteitspresentatie te adapteren aan de prestaties van de speler. Een versie van Code Red Triage met een progressieve toename van complexiteit werd aangepast om automatisch naar een complexere slachtofferklasse te gaan wanneer de speler boven een zekere grens scoorde in de game. Dit werd vergeleken met een controle conditie zonder adaptiviteit.

In de twee condities werd een vergelijkbare effectiviteit in leren gevonden, gemeten aan totale leerwinsten, maar de participanten in de adaptieve conditie hadden significant minder slachtoffers en totale tijd nodig om dit resultaat te bereiken. Als gevolg hiervan hadden de participanten in de adaptieve conditie een hogere leerwinst per slachtoffer, zowel in termen van de kennistest als de structurelekennistoets, hetgeen leidt tot een hogere efficiëntie van de game over het geheel genomen. Adaptiviteit had geen invloed op de affectieve beoordeling van de spelers, zoals hoe leuk, immersief of moeilijk ze het vonden.

In het vierde en laatste experiment (hoofdstuk 7) hebben we onderzocht of het integratieproces van leren met serious games gestimuleerd kan worden door de introductie van verrassende gebeurtenissen in het narratief van de game. De onderliggende aanname hiervan was dat mensen nieuwe informatie die afwijkt van wat eerder bekend was dikwijls over het hoofd zien. Doordat een verrassende gebeurtenis door haar aard onverwacht is en daarom niet voorspeld kan worden door het eerder opgebouwde mentale model van de speler, zou dit tot een betere update van het mentale model moeten leiden. Het bestaande mentale model wordt onder invloed van de verrassende gebeurtenis sterker vanuit het geheugen geactiveerd. Dit kan op zijn beurt leiden tot betere integratie van de nieuwe informatie in de bestaande kennisstructuur.

Drie verrassende gebeurtenissen werden geïmplementeerd in het reeds bestaande omgevingsnarratief van Code Red Triage. Deze gebeurtenissen werden gepresenteerd op het moment dat de speler nieuwe informatie moest leren die in enige mate eerdere preconcepties tegensprak. De participanten in de ‘verrassende gebeurtenissen’ conditie ontwikkelden significant betere kennisstructuren, zoals bleek uit de structurelekennistoets, na het spelen van de game in vergelijking met de controle conditie, hetgeen bevestigt dat een relatief simpele interventie al kan leiden tot dieper leren en effectievere serious games. Het leidde niet tot betere scores op de kennistest met indirecte vragen, noch op de test met vragen die direct gerelateerd waren aan de verrassende gebeurtenissen, maar dit kan komen door een zogeheten plafond effect. Het had ook geen effect op hoe leuk of immersief de speler het vond.

Concluderend (hoofdstuk 8), dit onderzoek was gestart om te onderzoeken of de doeltreffendheid van serious games verbeterd kan worden door problemen te bepalen in de verschillende onderdelen van cognitief leren met serious games, en deze systematisch proberen te verbeteren, uitkomend op een set van richtlijnen voor seriousgameontwikkelaars. Van de vier experimenten bleken er twee succesvol in het significant verbeteren van de doeltreffendheid van onze serious game. Hieruit kunnen we concluderen dat het aanbeveling verdient om twee dingen te doen. Ten eerste, dat een adaptatiemechanisme gebaseerd op de prestaties van de speler in de game geïmplementeerd dient te worden om de efficiëntie van het leren in de game te verbeteren. Ten tweede, dat verrassende gebeurtenissen geïntroduceerd dienen te worden, op die momenten dat men iets moet leren dat contrasteert met eerdere preconcepties, om dieper leren te stimuleren. Het gebruik van auditieve cues lijkt afgeraden te worden, alhoewel meer onderzoek nodig is om dit beslissend te kunnen stellen.

De richtlijnen dienden daarnaast geen negatieve invloed op de leukheid en immersiviteit van de game te hebben. Dit is grotendeels gelukt. Slechts een van de experimenten toonde een significant effect van een interventie op de leukheid van de game. De validiteit hiervan is echter enigszins betwistbaar doordat de participanten geen duidelijk referentiemateriaal hadden om hun versie van de game mee te vergelijken. De resultaten van het tweede experiment lijken de bevindingen van Przybylski en collega's (2010) te onderschrijven, namelijk dat games die de speler een groter gevoel van autonomie geven, leuker gevonden worden; of, vice versa, dat games die deze autonomie wegnemen, minder leuk gevonden worden.

Tot slot, hoewel dit onderzoek niet is gedaan om ondersteuning te vinden voor Mayer's cognitieve theorie van leren met multimedia, heeft ons onderzoek waarbij we de verschillende cognitieve processen afzonderlijk benaderden wel tot merkbare effecten geleid op de totale leerwinsten na het spelen van de game. De poging het selectieproces te beïnvloeden met auditieve guidance cues leidde tot slechtere kennisstructuren na de game, de poging het organisatieproces te beïnvloeden met adaptiviteit in de complexiteitspresentatie leidde tot efficiënter leren en de poging het integratieproces te beïnvloeden door de introductie van verrassende gebeurtenissen leidde tot betere kennisstructuren. Dit toont aan dat Mayer's theorie geschikt is om als raamwerk te dienen voor game design interventies. Het geeft ook een reden om additioneel onderzoek te verrichten, zo dat bepaald kan worden of de onderscheiden cognitieve processen daadwerkelijk onafhankelijk van elkaar gezien kunnen worden.

2009

2009-01	**Rasa Jurgelenaite (RUN)**	*Symmetric Causal Independence Models*
2009-02	**Willem Robert van Hage (VU)**	*Evaluating Ontology-Alignment Techniques*
2009-03	**Hans Stol (UvT)**	*A Framework for Evidence-based Policy Making Using IT*
2009-04	**Josephine Nabukenya (RUN)**	*Improving the Quality of Organisational Policy Making using Collaboration Engineering*
2009-05	**Sietse Overbeek (RUN)**	*Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*
2009-06	**Muhammad Subianto (UU)**	*Understanding Classification*
2009-07	**Ronald Poppe (UT)**	*Discriminative Vision-Based Recovery and Recognition of Human Motion*
2009-08	**Volker Nannen (VU)**	*Evolutionary Agent-Based Policy Analysis in Dynamic Environments*
2009-09	**Benjamin Kanagwa (RUN)**	*Design, Discovery and Construction of Service-oriented Systems*
2009-10	**Jan Wielemaker (UVA)**	*Logic programming for knowledge-intensive interactive applications*
2009-11	**Alexander Boer (UVA)**	*Legal Theory, Sources of Law & the Semantic Web*
2009-12	**Peter Massuthe (TUE)**	*Operating Guidelines for Services*
2009-13	**Steven de Jong (UM)**	*Fairness in Multi-Agent Systems*
2009-14	**Maksym Korotkiy (VU)**	*From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*
2009-15	**Rinke Hoekstra (UVA)**	*Ontology Representation - Design Patterns and Ontologies that Make Sense*
2009-16	**Fritz Reul (UvT)**	*New Architectures in Computer Chess*
2009-17	**Laurens van der Maaten (UvT)**	*Feature Extraction from Visual Data*
2009-18	**Fabian Groffen (CWI)**	*Armada, An Evolving Database System*
2009-19	**Valentin Robu (CWI)**	*Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*
2009-20	**Bob van der Vecht (UU)**	*Adjustable Autonomy: Controling Influences on Decision Making*
2009-21	**Stijn Vanderlooy (UM)**	*Ranking and Reliable Classification*
2009-22	**Pavel Serdyukov (UT)**	*Search For Expertise: Going beyond direct evidence*
2009-23	**Peter Hofgesang (VU)**	*Modelling Web Usage in a Changing Environment*

2009-24	**Annerieke Heuvelink (VUA)**	*Cognitive Models for Training Simulations*
2009-25	**Alex van Ballegooij (CWI)**	*RAM: Array Database Management through Relational Mapping*
2009-26	**Fernando Koch (UU)**	*An Agent-Based Model for the Development of Intelligent Mobile Services*
2009-27	**Christian Glahn (OU)**	*Contextual Support of social Engagement and Reflection on the Web*
2009-28	**Sander Evers (UT)**	*Sensor Data Management with Probabilistic Models*
2009-29	**Stanislav Pokraev (UT)**	*Model-Driven Semantic Integration of Service-Oriented Applications*
2009-30	**Marcin Zukowski (CWI)**	*Balancing vectorized query execution with bandwidth-optimized storage*
2009-31	**Sofiya Katrenko (UVA)**	*A Closer Look at Learning Relations from Text*
2009-32	**Rik Farenhorst (VU) and Remco de Boer (VU)**	*Architectural Knowledge Management: Supporting Architects and Auditors*
2009-33	**Khiet Truong (UT)**	*How Does Real Affect Affect Affect Recognition In Speech?*
2009-34	**Inge van de Weerd (UU)**	*Advancing in Software Product Management: An Incremental Method Engineering Approach*
2009-35	**Wouter Koelewijn (UL)**	*Privacy en Politiegegevens*
2009-36	**Marco Kalz (OUN)**	*Placement Support for Learners in Learning Networks*
2009-37	**Hendrik Drachsler (OUN)**	*Navigation Support for Learners in Informal Learning Networks*
2009-38	**Riina Vuorikari (OU)**	*Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*
2009-39	**Christian Stahl (TUE)**	*Service Substitution -- A Behavioral Approach Based on Petri Nets*
2009-40	**Stephan Raaijmakers (UvT)**	*Multinomial Language Learning: Investigations into the Geometry of Language*
2009-41	**Igor Berezhnyy (UvT)**	*Digital Analysis of Paintings*
2009-42	**Toine Bogers (UvT)**	*Recommender Systems for Social Bookmarking*
2009-43	**Virginia Nunes Leal Franqueira (UT)**	*Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*
2009-44	**Roberto Santana Tapia (UT)**	*Assessing Business-IT Alignment in Networked Organizations*
2009-45	**Jilles Vreeken (UU)**	*Making Pattern Mining Useful*
2009-46	**Loredana Afanasiev (UvA)**	*Querying XML: Benchmarks and Recursion*

2010

2010-01	**Matthijs van Leeuwen (UU)**	*Patterns that Matter*
2010-02	**Ingo Wassink (UT)**	*Work flows in Life Science*
2010-03	**Joost Geurts (CWI)**	*A Document Engineering Model and Processing Framework for Multimedia documents*
2010-04	**Olga Kulyk (UT)**	*Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*
2010-05	**Claudia Hauff (UT)**	*Predicting the Effectiveness of Queries and Retrieval Systems*
2010-06	**Sander Bakkes (UvT)**	*Rapid Adaptation of Video Game AI*
2010-07	**Wim Fikkert (UT)**	*Gesture interaction at a Distance*
2010-08	**Krzysztof Siewicz (UL)**	*Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*
2010-09	**Hugo Kielman (UL)**	*A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*
2010-10	**Rebecca Ong (UL)**	*Mobile Communication and Protection of Children*
2010-11	**Adriaan Ter Mors (TUD)**	*The world according to MARP: Multi-Agent Route Planning*
2010-12	**Susan van den Braak (UU)**	*Sensemaking software for crime analysis*
2010-13	**Gianluigi Folino (RUN)**	*High Performance Data Mining using Bio-inspired techniques*
2010-14	**Sander van Splunter (VU)**	*Automated Web Service Reconfiguration*
2010-15	**Lianne Bodenstaff (UT)**	*Managing Dependency Relations in Inter-Organizational Models*
2010-16	**Sicco Verwer (TUD)**	*Efficient Identification of Timed Automata, theory and practice*
2010-17	**Spyros Kotoulas (VU)**	*Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*
2010-18	**Charlotte Gerritsen (VU)**	*Caught in the Act: Investigating Crime by Agent-Based Simulation*
2010-19	**Henriette Cramer (UvA)**	*People's Responses to Autonomous and Adaptive Systems*
2010-20	**Ivo Swartjes (UT)**	*Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*
2010-21	**Harold van Heerde (UT)**	*Privacy-aware data management by means of data degradation*
2010-22	**Michiel Hildebrand (CWI)**	*End-user Support for Access to Heterogeneous Linked Data*
2010-23	**Bas Steunebrink (UU)**	*The Logical Structure of Emotions*

2010-24	**Dmytro Tykhonov**	*Designing Generic and Efficient Negotiation Strategies*
2010-25	**Zulfiqar Ali Memon (VU)**	*Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*
2010-26	**Ying Zhang (CWI)**	*XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*
2010-27	**Marten Voulon (UL)**	*Automatisch contracteren*
2010-28	**Arne Koopman (UU)**	*Characteristic Relational Patterns*
2010-29	**Stratos Idreos(CWI)**	*Database Cracking: Towards Auto-tuning Database Kernels*
2010-30	**Marieke van Erp (UvT)**	*Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval*
2010-31	**Victor de Boer (UVA)**	*Ontology Enrichment from Heterogeneous Sources on the Web*
2010-32	**Marcel Hiel (UvT)**	*An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*
2010-33	**Robin Aly (UT)**	*Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*
2010-34	**Teduh Dirgahayu (UT)**	*Interaction Design in Service Compositions*
2010-35	**Dolf Trieschnigg (UT)**	*Proof of Concept: Concept-based Biomedical Information Retrieval*
2010-36	**Jose Janssen (OU)**	*Paving the Way for Lifelong Learning*
2010-37	**Niels Lohmann (TUE)**	*Correctness of services and their composition*
2010-38	**Dirk Fahland (TUE)**	*From Scenarios to components*
2010-39	**Ghazanfar Farooq Siddiqui (VU)**	*Integrative modeling of emotions in virtual agents*
2010-40	**Mark van Assem (VU)**	*Converting and Integrating Vocabularies for the Semantic Web*
2010-41	**Guillaume Chaslot (UM)**	*Monte-Carlo Tree Search*
2010-42	**Sybren de Kinderen (VU)**	*Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach*
2010-43	**Peter van Kranenburg (UU)**	*A Computational Approach to Content-Based Retrieval of Folk Song Melodies*
2010-44	**Pieter Bellekens (TUE)**	*An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*
2010-45	**Vasilios Andrikopoulos (UvT)**	*A theory and model for the evolution of software services*
2010-46	**Vincent Pijpers (VU)**	*e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

2010-47	**Chen Li (UT)**	*Mining Process Model Variants: Challenges, Techniques, Examples*
2010-48	**Withdrawn**	
2010-49	**Jahn-Takeshi Saito (UM)**	*Solving difficult game positions*
2010-50	**Bouke Huurnink (UVA)**	*Search in Audiovisual Broadcast Archives*
2010-51	**Alia Khairia Amin (CWI)**	*Understanding and supporting information seeking tasks in multiple sources*
2010-52	**Peter-Paul van Maanen (VU)**	*Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*
2010-53	**Edgar Meij (UVA)**	*Combining Concepts and Language Models for Information Access*

2011

2011-01	**Botond Cseke (RUN)**	*Variational Algorithms for Bayesian Inference in Latent Gaussian Models*
2011-02	**Nick Tinnemeier(UU)**	*Work flows in Life Science*
2011-03	**Jan Martijn van der Werf (TUE)**	*Compositional Design and Verification of Component-Based Information Systems*
2011-04	**Hado van Hasselt (UU)**	*Insights in Reinforcement Learning*
2011-05	**Base van der Raadt (VU)**	*Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.*
2011-06	**Yiwen Wang (TUE)**	*Semantically-Enhanced Recommendations in Cultural Heritage*
2011-07	**Yujia Cao (UT)**	*Multimodal Information Presentation for High Load Human Computer Interaction*
2011-08	**Nieske Vergunst (UU)**	*BDI-based Generation of Robust Task-Oriented Dialogues*
2011-09	**Tim de Jong (OU)**	*Contextualised Mobile Media for Learning*
2011-10	**Bart Bogaert (UvT)**	*Cloud Content Contention*
2011-11	**Dhaval Vyas (UT)**	*Designing for Awareness: An Experience-focused HCI Perspective*
2011-12	**Carmen Bratosin (TUE)**	*Grid Architecture for Distributed Process Mining*
2011-13	**Xiaoyu Mao (UvT)**	*Airport under Control. Multiagent Scheduling for Airport Ground Handling*
2011-14	**Milan Lovric (EUR)**	*Behavioral Finance and Agent-Based Artificial Markets*
2011-15	**Marijn Koolen (UvA)**	*The Meaning of Structure: the Value of Link Evidence for Information Retrieval*
2011-16	**Maarten Schadd (UM)**	*Selective Search in Games of Different Complexity*

2011-17	**Jiyin He (UVA)**	*Exploring Topic Structure: Coherence, Diversity and Relatedness*
2011-18	**Mark Ponsen (UM)**	*Strategic Decision-Making in complex games*
2011-19	**Ellen Rusman (OU)**	*The Mind ' s Eye on Personal Profiles*
2011-20	**Qing Gu (VU)**	*Guiding service-oriented software engineering - A view-based approach*
2011-21	**Linda Terlouw (TUD)**	*Modularization and Specification of Service-Oriented Systems*
2011-22	**Junte Zhang (UVA)**	*System Evaluation of Archival Description and Access*
2011-23	**Wouter Weerkamp (UVA)**	*Finding People and their Utterances in Social Media*
2011-24	**Herwin van Welbergen (UT)**	*Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*
2011-25	**Syed Waqar ul Qounain Jaffry (VU)**	*Analysis and Validation of Models for Trust Dynamics*
2011-26	**Matthijs Aart Pontier (VU)**	*Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*
2011-27	**Aniel Bhulai (VU)**	*Dynamic website optimization through autonomous management of design patterns*
2011-28	**Rianne Kaptein (UVA)**	*Effective Focused Retrieval by Exploiting Query Context and Document Structure*
2011-29	**Faisal Kamiran (TUE)**	*Discrimination-aware Classification*
2011-30	**Egon van den Broek (UT)**	*Affective Signal Processing (ASP): Unraveling the mystery of emotions*
2011-31	**Ludo Waltman (EUR)**	*Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*
2011-32	**Nees-Jan van Eck (EUR)**	*Methodological Advances in Bibliometric Mapping of Science*
2011-33	**Tom van der Weide (UU)**	*Arguing to Motivate Decisions*
2011-34	**Paolo Turrini (UU)**	*Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*
2011-35	**Maaike Harbers (UU)**	*Explaining Agent Behavior in Virtual Training*
2011-36	**Erik van der Spek (UU)**	*Experiments in serious game design: a cognitive approach*