Measuring the experience of playing self-paced games

Joe Timothy Cutting

Doctor of Philosophy
University of York
Computer Science
December 2018

Abstract

Self-paced games are digital games which do not require players to make a move within a particular period of time, so the game can be played at whichever speed the player desires. Game experience measures can help game designers and increase understanding of how games create engagement. This thesis aimed to develop new measures of the experience of playing self-paced games. It investigated two possible measures; measuring cognitive load using pupil dilation and measuring attention using irrelevant distractor images.

The first approach found a significant difference in pupil dilation between easy and hard variants of a task taken from the game *Two Dots.* In a subsequent study, participants played three different versions of *Two Dots* – one of which required no cognitive effort. There was no significant difference in pupil dilation due to cognitive load between the games. It seems likely that although players could use sustained cognitive effort to play the game, they chose not to, and use other strategies instead. I concluded that pupil dilation is unlikely to be an effective measure of game experience.

The second approach developed a new measure known as the *distractor recognition paradigm*. This measure surrounds the game with constantly changing irrelevant images. After playing, participants are tested on how many of these images they recognise. An initial study found a significant difference in images recognised between two very different versions of the game *Two Dots*. There was also a significant difference in distractors recognised between three more similar games. This was found to be a stronger measure of game attention than using eye tracking and also found to be effective if the distractor images were placed *inside* the game graphics. This approach succeeded in the aim of the thesis which was to find a new measure of the experience of playing self-paced games.

Contents

List of tables

List of figures

Acknowledgements

Whilst working on this PhD I have made many decisions, some big and others small. For some of these decisions I have made the right call, others have been less successful. As it happens, the wrong decisions did not really matter too much due to me making the right call on two big choices. The first correct choice was in choosing a supervisor. When I started, I thought that Paul's role was to know about research and impart his wisdom. Sometime during the four years I have spent with him I realised that that was not really what our meetings were about. Instead, he has provided support through four years of me painstakingly working things out for myself. He's listened to me moan, laughed at my jokes and politely pointed out when my ANOVA reports have typos in them. Thank you, Paul.

The other decision I got right was in marrying Danijela. Giving up my job to spend four years doing a PhD was a big jump into the unknown but she has supported me throughout this time. She has been tolerant of my grumbling about academic life, has provided much needed perspective and reminded me what a privilege it has been to study full time for a PhD. Thank you, Danijela.

I would also like to thank the other students on the IGGI PhD programme. In particular, I'd like to thank Jen Beeston for reading a draft chapter and giving useful feedback. Traditionally a PhD has been a fairly lonely business, but one of the key benefits of IGGI has been having other students around who are doing the same thing. Despite me being much older than most of them they've been welcoming and supportive. Thanks guys you've made the experience a lot more fun.

I would also like to thank YCCSA for providing cake, great offices and a cosy common room to meet and have lunch in. In four years of eating there I heard very few research ideas being shared, instead I have looked forward to the strange and random conversations that remind us that there is a world outside research. When I met the new intake of IGGI students in the common room they were discussing "the difference between a fruit and a vegetable" so it looks like this will continue.

Finally, I would like to thank Helen for making the York HCI group a welcoming and friendly place to be. I've appreciated the seminars, days trips and, of course, the cake.

Author's declaration

I declare that the work presented in this thesis is entirely my own. This work has not been submitted for any other award at this or any other university. If information has been derived from some other source, I confirm that my thesis references or acknowledges this. Some of the material in this thesis has been previously published in the following conference paper:

Cutting, J., 2017, October. Measuring Game Experience using Visual Distractors. In Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play (pp. 695-698). ACM

1. Introduction

Monument Valley is a simple but beautifully designed game played on phones and tablets. Players guide a lone figure through fantastical buildings inspired by Islamic architecture (See Figure 1). The twist is that, in this world, the laws of physics have been written by M.C.Escher so by walking one way you soon find yourself walking up walls and along the ceiling. Turn a lever and the perspective changes so that platforms which were once out of reach are now close by. The whole effect is a spellbinding puzzle game which keeps players enthralled for hours.

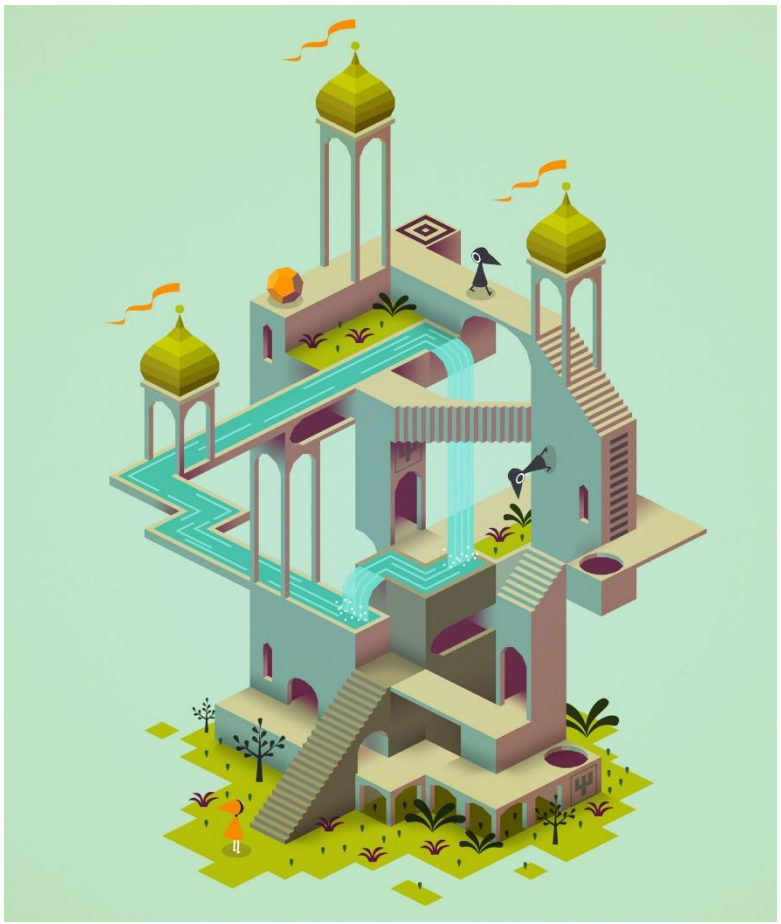

Figure 1 A screen from the game *Monument Valley* © ustwo ltd

But unlike more archetypal video games which require co-ordination and fast reflexes, players of *Monument Valley* could stop at any time without impacting their progress in the game. The game has no time limits or need for co-ordinated movement, the player is

simply making decisions at their own pace. Jennett et al. (2008) call games like this "self-paced" games because they proceed at a pace set by the player rather than imposing their own speed. In these games, players could stop at any time but the game successfully holds the player's attention for hours despite imposing no requirements for action.

Digital games now form a substantial sector of the entertainment industry with common claims that the industry generates more revenue than Hollywood (Chatfield, 2009). However, researchers and the media tend to focus on fast-paced action games such as *Fortnite* and *Dota 2*. This type of game is both hugely popular and lucrative for successful developers, but only form a fraction of the wide variety of games which are played. A substantial proportion would fall into the category of self-paced games including some of the very highest earning games such as *Candy Crush Saga* and *Puzzles and Dragons* (App Annie, 2018). Many self-paced games have widely acceptable topics without explicit sex or violence, they are quick to learn and can be played for short periods of time. Thus they fit into the genre known as *casual* games (Kultima, 2009). This includes serene artistic games such as *Monument Valley* as well as brightly coloured puzzle games like *Candy Crush Saga.* However, there are also a wide variety of self-paced games which do not fit the "casual" label. Strategy games such as *Civilization* and *X-COM* create compelling (Murnane, 2016) long-lasting experiences which keep players occupied for days using only self-paced decision making. Self-paced games thus create a wide variety of experiences for players many of which lead to deep engagement. The experience of playing games is a complex and varied phenomenon. To get a better understanding of game experience, researchers have attempted to measure it.

1.1. Measuring game experience

Ever since McCarthy and Wright (2004) proposed that technology should be considered in terms of the overall experience there have been numerous approaches to measuring the experience of playing digital games. McCarthy and Wright advocated that only holistic methods should be used to consider experience, despite this, many of these approaches have been more reductionist with the specific aim of measuring experience rather than just describing it. There are many different aspects to the experience of playing games which range from the physiological changes in players' bodies to the mental processes they use to play the game and the emotions that they feel. Most approaches to measuring this experience narrow it down to particular features which are of interest. The most common approach is to measure how engaged players are in the game because most games aim to create an engaging experience and a high level of engagement indicates a successful game. However, engagement is also a complex concept which is difficult to define. To make it more manageable, researchers have identified different aspects of engagement such as *immersion* (Brown and Cairns, 2004) and *flow* (Csikszentmihalyi, 1991) which are more tightly defined and can be measured

In a systematic review Mekler et al. (2014) found 87 studies investigating enjoyment in digital games. A wide variety of game experience questionnaires have been created

including the Game Experience Questionnaire (GEQ) (Brockmyer et al., 2009), the Immersion Experience Questionnaire (IEQ) (Jennett et al., 2008) and the Player Experience of Needs (PENS) questionnaire (Ryan et al., 2006). These measures of engagement have been used to investigate topics such as the influence of particular game designs (Thompson et al., 2012, Denisova and Cairns, 2015) or contexts (Hudson and Cairns, 2014, Nordin et al., 2014) on the game play experience. Researchers have also measured physiological properties of players' bodies to see how they are affected by the experience of playing games. They have found insights into how players are experiencing the game by looking at changes in heart rate (Ravaja, 2004), skin conductance (Ambinder, 2011) and Electroencephalogram (EEG) (Nacke et al., 2010).

There has also been a growing awareness of the need for effective measures of game experience from game designers themselves. Game designers need a reliable measure that tells them which parts of the game are engaging and which are not working so well, that way they can change the elements which are not working and keep the ones that are. Game developer Bruce Phillips (Phillips, 2006, p.22) says "I have a secret longing for the confidence in purpose that I imagine my colleagues working on productivity applications must feel. Their goals seem communicable and measurable—mine don't." As well as being able to measure how well a game meets its goals, effective measures of game experience could shed light on how games create engaging experiences and the psychological mechanisms that make this happen. These insights could feed into game design and problems such as using serious games for education and behaviour change.

When applied to real game design problems some of the issues with current game experience measures become apparent. Questionnaires can give good information about game experience but suffer from being a self-reported measure, players may only report the most memorable or intensive parts of their experience (Kahneman et al., 1993). Another limitation of questionnaires is that they ask players to describe the whole experience of playing the game so they are unable to capture changes in experience over the time of playing the game. This is a particular problem for use in game development because typically game designers wish to know which particular parts of the game create problems for the players and which lead to most enjoyment (McAllister et al., 2013, Gow et al., 2010).

Physiological approaches such as measuring heart rate, skin conductance and EEG avoid both of these issues. Measurements are taken directly from players' bodies so there is no problem of self-report. Readings are continuous for the whole length of play so particular readings can be associated with particular game events during play. However, it can be difficult to interpret this data and relate it to the actual experience of the game. If my heart rate increases during a game, am I excited by the game, frustrated by the game or is it a reaction to the caffeine in the coffee I had before playing? For an ideal physiological measure of experience, it would be possible to reliably link a change in the measure directly to a change in a psychological concept that is affected by the game.

1.1.1. Self-paced games are different

Current approaches to game experience tend not to consider self-paced games or consider their special features. Though Mekler et al. (2014)'s systematic review showed the wide interest in measuring player experience, only 5 out of 87 studies looked at self-paced games. Fast-paced games such as *Half-Life* not only dominate the research but also the conceptualisations of player experience. Most common game experience questionnaires such as the IEQ and PENS were developed on action games. This can be seen in the types of questions they ask, PENS asks about "intuitive controls" which may be important in a first-person shooter like *Half-Life* which is controlled by both hands at once and requires split second timing, but is probably less of an issue in a slow-moving minimalistic puzzle game like *Two Dots.* Similarly, the IEQ asks whether players felt they were moving through the game environment which makes perfect sense in a 3D role-playing game like *Skyrim* but does not work so well for a puzzle game like *Candy Crush Saga.* This difference between action and self-paced games is even more likely to be a problem with existing physiological measures of game experience. Both heart rate and skin conductance are linked to the psychological feeling of *arousal* (Mehrabian, 1996, Reisenzein, 1994) which may be created by the excitement and urgency of action games but is unlikely to form part of the experience of a self-paced strategy game like *Civilization.* Action games and self-paced games require different skills; action games place a premium on fast reactions and co-ordination whereas self-paced games tend to favour mental effort and strategy. It is therefore likely that they create different types of experience in players which needs to be measured using a different type of measure.

There is therefore a need from both academia and industry for a new measure of game experience which is designed for self-paced games. Ideally this measure should avoid the issues with previous measures. It should be designed and tested with self-paced games. It should not rely on participants' self-reports of their experience and have the potential for continuous measurement so that it can show variations in experience over time. The best new measure would be practical for use by both academia and industry. To be practical a measure would reliably show significant differences in experience with just a modest number of participants. It should have high discriminant validity and sensitivity (Kline, 2013). This means that unlike some current physiological measures, there should be an unambiguous link between the results of the measure and the aspect of game experience it purports to measure. Any new measure also needs to build on previous work and be "benchmarked" against previous findings to ensure that it is a reliable measure of experience. Currently the most reliable measures of game experience are validated questionnaires so a new measure should also be compared to the results of an established game questionnaire.

1.2. Research questions

Being able to measure game experience is important for both game development and more in-depth studies of how game elements combine to create a particular experience for the player. Self-paced games make up a substantial segment of digital games but current experience measures are not developed for this type of game and may be unsuited to the experiences these games create. Many existing measures also suffer from the problem that they are self-reported measures without potential for continuous measurement. Therefore, the overarching research question of this thesis is:

Can new measures be developed to measure the experience of playing self-paced games?

There are many different potential approaches to measuring experience. This thesis investigates two different measures;

1. Changes in pupil dilation
2. How well the game holds participants' attention

Unlike current questionnaire-based approaches, neither of these measures rely on self-report and both have the potential for continuous measurement. They can also both be directly linked to psychological processes so would provide a less ambiguous measure than current physiological approaches.

1.2.1. Pupil dilation

When participants are asked to perform cognitively demanding tasks their pupils increase in size (Kahneman and Beatty, 1966) which is known as pupil dilation. Many self-paced games are puzzle games or place a premium on mental effort. It seemed likely that playing these games would require cognitive load which would create changes in pupil size and that more engaging or demanding games would create particular patterns of pupil change. Thus, the specific research question investigated in this thesis is:

Can changes in pupil dilation be used to measure the experience of playing self-paced games?

1.2.2. Game attention

Digital games are well-known for holding players' attention and stopping them becoming distracted by events around them. Self-paced games are no exception to this, even though players could stop at any time and attend to surrounding events. It seems likely that players who are more engaged in the game would be less likely to pay attention to external stimuli which are irrelevant to the game

Can the experience of playing self-paced games be measured by seeing how likely players are to notice irrelevant stimuli while playing the game?

1.3. Methodology

Most game experience measures produce numerical results based on empirical tests of the game in question. The aim of this research is to develop new quantitative measures of game experience which can be used in empirical tests to measure the experience of playing a self-paced game. It follows from this that to develop a new measure I would use an experimental methodology based on performing quantitative studies. These studies were iterative and performed in groups. Each group of studies started with an initial explorative study which replicated or built on a study from the literature. Subsequent studies then iterated on this initial study, being informed by previous results, but varying the study goals or experimental setup.

Experiments are often reported using the convention that first all results are reported and then they are discussed. More recently, writers such as Pinker (2014) have suggested that this does not give a full picture of the experimental process. In reporting my experiments, I will first state the main hypothesis that is being tested. I will then report the results and analysis which relate to this hypothesis and were planned before starting the experiment. Then I will discuss the results and how they relate to the hypothesis and aims of the experiment. Sometimes these results and discussion will inspire further analysis of the data which I had not originally planned. I will report this additional analysis during the discussion to which it relates.

Most of the studies aimed to test hypotheses using an ANOVA (Analysis of variance) analysis to report if the hypothesis was supported and to estimate the magnitude of the effect size. The aim of these studies was to develop a sensitive measure of the game experience which would give rise to a large effect size when comparing different games. Smaller effects which only can only be reliably seen with large number of participants would not make useful measures. This meant that my studies could be performed with small numbers of participants. In all studies the measure of game experience was also benchmarked against the Immersion Experience Questionnaire (Jennett et al., 2008).

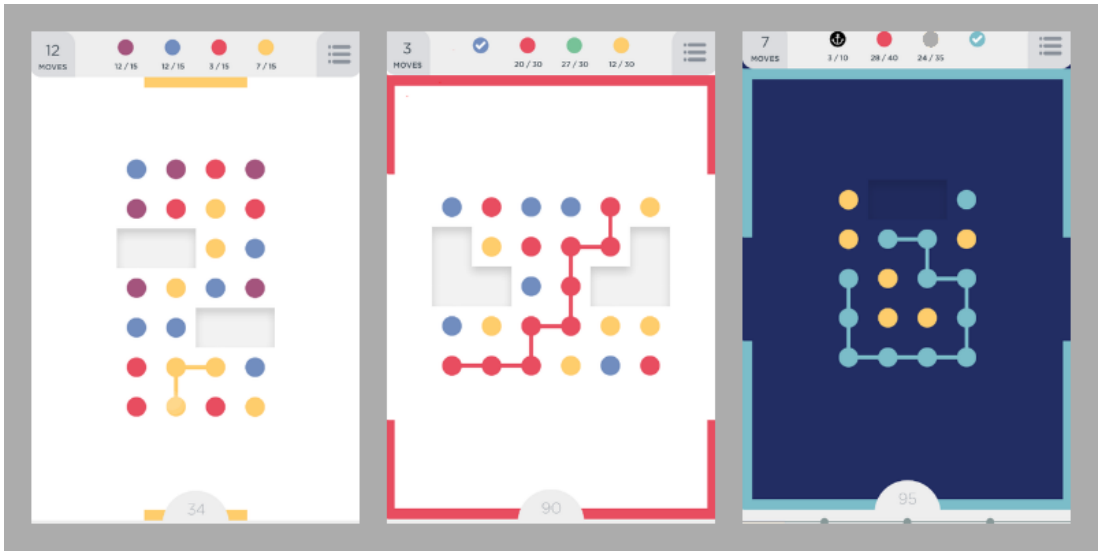

Figure 2 Three screens from the game *Two Dots*. Players have to join dots which are the same colour © Playdots Inc.

Throughout the thesis I used variants of the game *Two Dots* which is a simple self-paced game. This made it quick to prepare experiments and allowed some degree of comparison between experiments. More extensive details of the methodology are given in chapter 3.

1.4. Outline of the research

To answer the research questions described above I carried out two sequences of experiments. The first set of experiments looked at using pupil dilation to measure game experience. The second set looked at measuring how well games held players' attention. The following outlines provide a summary and rationale of the research which was carried out.

1.4.1. Outline of pupil dilation experiments

Measuring pupil dilation is a complex procedure requiring specialised equipment and analysis. To confirm that my experimental procedure worked well I began the pupil dilation experiments by replicating a study done by Kahneman and Beatty (1966). They asked participants to either memorise 4 numbers and repeat them back or to memorise 4 numbers and add one to each number before repeating it. They found that pupil dilation was higher in the task where participants had to add one to the number as well as memorise it. They ascribe the higher pupil dilation to the addition task requiring more cognitive load than just memorising. After replicating this experiment using an audio stimulus and a spoken response, I then extended it and performed a new experiment using a visual stimulus and a mouse-based response. This experiment also showed a significant difference in pupil dilation due to the cognitive load used. The next experiment considered tasks similar to those performed in the game *Two Dots*. These tasks involved visually searching for groups of dots and mental manipulation to

predict the results of the next move. The experiment compared the amount of cognitive load required between an easy task and hard task taken from the game. Once again, I found a significant difference in pupil dilation and hence cognitive load between these tasks.

The next experiments extended this measure by trying it with participants playing a real game. Participants played one of three different games, either full *Two Dots* or one of two different variants of that game. One of these variants, known as *All dots the same*, had no puzzle elements in it and would seem likely to require less cognitive load than the full version of *Two Dots.* My initial hypothesis was that the full version of *Two Dots* would require more cognitive load, and thus have higher pupil dilation, than the *All dots the same* game variant. This hypothesis was not supported by the experimental results. However, there was a difference in pupil dilation but the difference was the reverse of the expected effect – the *All dots the same* game had higher pupil dilation than the full version of the game. This is likely to be due to participants who played the *All dots the same* game were making more complex motor actions which created higher pupil dilation. To confirm this hypothesis, I repeated the three-game pupil dilation experiment. As I predicted this found that the pupil dilation for the *All dots the same* game was significantly higher than the full version of *Two Dots,* due to the more complex motor actions used in that game. These complex motor actions were probably created because participants made more complex mouse movements in the less engaging game to make it more interesting.

It is possible that the lack of difference in cognitive load between games is an experimental artefact. However, the previous experiments did find a large difference in pupil dilation due to cognitive load in the game-like task. They also repeatedly found no differences in cognitive load between variants of an actual game and much lower pupil dilation overall in the real games than the game-like task. This adds up to considerable evidence that participants did not use significant cognitive load when playing any of the game variants. This unexpected finding suggests that sustained cognitive load is not a key part of the experience of playing self-paced games like *Two Dots.* If this is the case then measuring cognitive load using pupil dilation is unlikely to be a useful measure of the experience of playing self-paced games

These experiments are described in chapters 4 and 5.

1.4.2. Outline of attention experiments

These experiments investigated a novel way of measuring game attention that I named the *distractor recognition paradigm.* In this paradigm participants played a game which is surrounded by constantly changing irrelevant distractor images such icons from the *Webdings* typeface. After the game has finished, they are tested to see how many of the distractors they recognise seeing. The idea was that the more engaged they are in the game, the fewer distractors they will recognise. Less engaging games will not hold participants' attention so well, so they will look at more of the distractors and recognise more of them afterwards.

To investigate this paradigm, I first needed to get a baseline of how many distractors participants would recognise if they were not playing a game at all. Experiments by Standing (1973) showed that participants could recognise over 90% of the images which they had seen previously. My first experiment replicated this work by showing participants 60 images, each for 5 seconds and then testing them afterwards. Like Standing I found recognition rates of over 90%.

I then performed an initial test of the *distractor recognition paradigm* by having participants play one of two very different games. The game graphics were located in the centre of the screen whilst the sides of the screen were filled with irrelevant distractor images. As predicted, participants who played the less engaging game recognised significantly more distractor images afterwards. The next step was to repeat the experiment with three more similar games and also to use eye tracking to record how often participants looked at the surrounding distractor images. This experiment also found that participants in the less engaging games recognised significantly more distractors afterwards. There were some differences in the amount of time that they looked at the distractors but overall eye tracking had a lower effect size and was a weaker measure than testing how many distractors participants recognised. In both these experiments the number of distractors recognised was inversely related to how engaging participants found the game. I therefore concluded that the distractor recognition paradigm could be an effective measure of game engagement.

For all these experiments I used icons from the *Webdings* font as distractor images. These are sufficiently different to be recognised but not particularly interesting or distinctive. The next experiment investigated the effect of using more interesting images which might be more distracting. I decided to use images of Disney characters for this purpose. I repeated the initial recognition test without any game present and found that participants recognised over 90% of images. This indicated that Disney characters had high levels of recognition which was similar to that of *Webdings* icons. I then repeated the previous three game experiment but with Disney distractors instead of *Webdings* icons. Participants did recognise more distractors but there was no significant difference in the number recognised between the different games. It appears that the Disney distractors had made the different games more similar to each other in the overall game experience that they provided. The *Full game* had become less engaging due to the Disney distractors around it. Conversely the *All dots the same* game had become more engaging due to the addition of the same Disney distractors. I concluded that using more interesting distractor images made the distractor recognition paradigm a weaker measure of attention.

During the previous experiments, participants put their head in a chin rest needed for the eye tracking equipment and were watched by the experimenter for the whole experiment. I was concerned that this was not an ecologically valid recreation of how people actually play games. So, I repeated the first three-game experiment using *Webdings* icons but removed the eye tracking, chin rest and put the experimenter in a

different room. The results showed no significant difference in the number of distractors recognised between the different games. In the previous experiments it is likely that participants who were less engaged in the game looked at the distractor images. It is possible that that without the chin rest or experimenter present they felt less constrained and looked around the room instead of looking at the distractors.

The eye tracking data showed that participants did not necessarily recognise a distractor image just because they had looked at it – their attention needed to be focused away from the game as well. I was interested to investigate what would happen if the distractor images were inside the game rather than surrounding it. The next experiment compared two games, both of which had the distractor images placed within the game. Both games were variants of *Two Dots* with the dots made much bigger and the distractor images placed inside the dots. After playing, participants recognised significantly more distractors in the less engaging game. This was despite the fact that in both games participants were looking directly at the distractors. I wanted to see whether this was a robust effect so repeated this experiment with three variants of *Two Dots.* This replication also found a same significant difference in distractors recognised. I concluded that the difference in distractors recognised is due to the level of engagement that players had in the game. Less engaged players were more likely to pay attention to the distractor images and so recognise them afterwards. More engaged players focus their attention on the colour of the dots rather than the images and recognise fewer images afterwards. This indicates that the distractor recognition paradigm can be used as a measure of game engagement even if the distractor images are within the game.

In all the previous *distractor recognition paradigm* experiments participants had not been told that they would be tested on the distractors until after the game. I performed an in-game distractor experiment with two games but this time I told participants that they would be tested on the distractors but also needed to play the game. The results showed no significant difference in the number of distractors recognised per game or the levels of immersion in each game (as measured by a questionnaire). I concluded that participants tried to focus their attention on both the distractor images and the game which they found very difficult. This made the more engaging game more difficult and the less engaging game more interesting thus reducing the difference in experience between the games and making it more difficult to measure. I concluded that the distractor recognition paradigm is only an effective measure if participants do not know that they will be tested on the distractors afterwards. These experiments are described in chapters 6 and 7.

1.5. Scope

The scope of this thesis is initially directed by the central research question *"What would be effective measures of game experience for self-paced games that do not rely on self-report and have the potential for continuous measurement?"* As such, it looks at self-paced digital

games and ways of measuring the experience of playing them. In particular, it considers two novel approaches to measuring experience – pupil dilation and attention.

The scope of the thesis is to explore the possibility of novel measures of game experience, this does not extend to producing a fully finished measure which could be used by commercial game developers without further development. All measures of experience make assumptions about the characteristics of the experience that they are measuring. By developing new ways of measuring the experience of playing self-paced games I am also exploring the psychological processes which take place whilst playing these games. Some approaches to game experience, e.g. Kultima and Stenros (2010) consider the wider context to playing games including deciding which game to play and deciding when to play it. This expanded conception of game experience is outside the scope of this thesis which considers only the experience that players feel when they are playing a single session of a game.

Furthermore, all of the game-based experiments in the thesis are performed on variants of a single game: *Two Dots.* The experiments are performed in a laboratory setting and players play the game for 5 minutes. There are, of course, an enormous variety of self-paced games which are played in different situations for much longer than 5 minutes. These are not in the scope of this work. Using the same game allowed experiments to be performed in an iterative sequence which allowed each experiment to build on the last and create a more in-depth investigation. *Two Dots* is a good candidate for a representative self-paced game. As a self-paced game is has no requirement for dexterity or rapid response. It is easy to learn and play, has very simple graphics and also shares features such as goals and levels with many other games. There are also many other similar games, most notably the hugely successful *Candy Crush Saga* and its derivatives (Dredge, 2014).

1.6. Ethics statement

All the research in this thesis was performed with due care for the participants who took part in the experiments and the impact of the research on the wider world. The research was performed according to the University of York's Code of Practice on Research Integrity. Each experiment was pre-screened by the Ethics Committee in the Department of Computer Science to ensure that it confirmed to these guidelines. Participants were all at least 18 years old and did not belong to any vulnerable groups. In particular, care was taken to consider participant welfare, the anonymity and confidentiality of participants' data and that they had given informed consent to performing the experiment. These issues are described in more detail below.

1.6.1. Participant welfare

The experiments were designed so that no participants were put in situations which might cause physical harm, mental discomfort or distress. For many of the experiments, participants played computer games but these games did not contain any violence,

depictions of violence or challenging themes. Participants were instructed to try their best at the games and other experimental activities but they were always rewarded for their performance regardless of their level. Care was taken so they never felt that they had "failed" at the experiment. Participants were instructed that they could leave the experiment at any time without any loss of reward or negative consequences. Some experiments involved eye tracking equipment which did not work for every participant. This was explained to participants before the experiment, so if the eye tracker did not work for them, they were sure that it was due to the nature of the equipment rather than them, the participant, being at fault. Participants who were unable to perform experiments were still thanked and rewarded as if they had completed the experiment.

1.6.2. Anonymity and confidentiality

Data collected in the experiments was anonymised and kept confidential. Participants' personal information such as their name was removed from the data and replaced by a participant number which was then used throughout the analysis and reporting. Spreadsheets and other data files were stored in password protected systems and paper data were stored securely and protected from unauthorised access.

1.6.3. Informed consent

Participants' informed consent was required before they took part in any experiments. Before the experiment started, they were given a short information sheet written in plain language which explained the nature of the experiment they were due to perform and what they would need to do (See an example in Appendix 1). They were also given the opportunity to ask questions about the experiment. They were then given the option to withdraw from the experiment or to sign to say they consented to perform the experiment as described. All participants were fully debriefed after the experiment so that they understood exactly what had been happening during the experiment and purpose of each activity. Some experiments, such as the attention experiments described in chapter 6 required that participants did not know exactly what was going to happen in the experiment beforehand. For these experiments particular care was taken that participants understood the reason for these details being omitted from the initial brief and were satisfied with the information they had been given.

1.7. Contributions

The contributions of this thesis fall into two main areas according to which research question I was considering.

1.7.1. Pupil dilation

The main contribution from the pupil dilation experiments was that pupil dilation is unlikely to be useful as a measure of game experience for self-paced games. This is because sustained cognitive load was not found to be an important aspect of puzzle games such as *Two Dots* and may not be an important part of other games as well. The

evidence for this was that I found no significant differences in pupil dilation due to cognitive load between different versions of *Two Dots* including one that had no puzzle elements. This is further supported by an experiment that did find a significant difference in cognitive load between easy and hard tasks taken from the game of *Two Dots* and performed in isolation rather than as part of a game. This suggests that I was using an effective experimental technique for measuring cognitive load using pupil dilation. If this was the case then the reason that I did not see sustained cognitive load during actual gameplay is that players choose not to put the same amount of cognitive effort into the game as the they had in the previous experiment. Instead, they used other, less cognitive intensive strategies, for playing the game.

This other main contribution of these experiments was the development of this experimental technique to measure cognitive load using pupil dilation. This included replicating Kahneman and Beatty (1966)'s finding that pupil dilation is related to cognitive load for audio numeric processing tasks. I also extended this to visual tasks with a mouse-based response method. I document the experimental procedure needed to avoid confounds such as noise from changing light levels, training and fatigue effects and additional sources of pupil dilation such as emotional differences.

1.7.2. Game attention

The main contribution of these experiments is the development of the *distractor recognition paradigm* as a method of measuring game attention. In this paradigm irrelevant distraction images are shown in or around a game and after play has finished participants are tested on how many images they recognise. Players who are more engaged in the game recognise fewer images. This was shown to be an effective measure of the experience of playing self-paced games. This may have the potential to be a continuous measure of engagement which can detect changes in engagement over time. However, the experiments I performed did not show this, possibly because the game used did not feature strong changes in engagement.

Another interesting contribution is the finding that putting the distractor images inside the game rather than around it was also shown to be an effective measure of game attention. This has implications for the way that attention works in games. Players are looking directly at the distractor images but still do not recognise them after playing the more immersive game. This lack of recognition suggests that players' attention was focused on the screen elements needed to play the game (the colours of the dots) rather than the distractor images, so they were effectively blind to the presence of the images. This is likely to be an example of the psychological phenomena of *inattentional blindness* (Simons and Chabris, 1999).

As part of the development of the distractor recognition paradigm I replicated Standing (1973)'s finding that participants recognise around 90% of visual images that they had been previously shown. I also explored several methodological variations on this measure and there are contributions which describe how effective these variations are. Eye tracking can be used to measure how often participants look at the distractor

images. This was shown to vary depending on how engaged participants were in the game, but is a less effective measure than the test of how many images were recognised. Making the distractor images more interesting by using Disney characters was also shown to reduce the effectiveness of the paradigm. Removing the chin rest and putting the experimenter in a different room was also shown to stop the paradigm from working. The same effect could also be achieved by warning participants during the experiment introduction that they will be tested on the distractor images after the game.

2. Literature review

The overarching research question of this thesis is *What would be effective measures of game experience for self-paced games that do not rely on self-report and have the potential for continuous measurement?*
All measures are based on some property of the target that is being measured. To develop new measures of game experience it is necessary to consider the properties of the experience of playing self-paced games. Therefore, this review starts by looking at literature which considers the nature of games and concepts such as play and fun. Game designers have built up considerable experience and insight into how particular game elements create particular experiences. Accordingly, the review then looks at what designers consider the important properties which make up the experience of playing a game. It also considers other researchers who have taken a reflective approach to identifying the concepts important to game experience.
Much of the literature on game design takes a non-empirical approach. However, there is also considerable literature on empirically based models of game experience. This literature takes two main methods to measuring game experience:

1. Post-game questionnaires.
2. Physiological measures of how players' bodies respond to the game experience.

Investigating the literature on these methods showed that a large number of different game experience questionnaires have been developed. There are fewer studies into physiological measures but there are still a wide variety of measures, games and data processing techniques to consider.

The next stage of the review considered which methods would be the most likely to create a new measure of game experience for self-paced games that does not rely on self-report and had the potential for continuous measurement. It settled on two different approaches. The first was to look at whether changes in pupil dilation could be used to as a measure of game experience. As a first step to investigating this the review looked into pupil dilation and how it has been used to measure other experiences. The second approach reflects that games are good at holding players' attention and considers how this property could be measured. As an initial step to looking at this the review considered the literature on attention and how to measure distraction.

2.1. Play and fun

We use the word "play" to describe the process of interacting with a game. We also associate games with "fun". So, to understand players' engagement in self-paced games it makes sense to start with looking at these terms. Huizinga was one of the earliest modern writers on play. In *Homo Ludens* (Huizinga, 1938/2014) he looks at the play element of culture. Notably he coined the term "the magic circle" to describe the social contract that players enter when they play a game. Caillois (1961) described play as existing on a continuous range from complete unstructured playfulness (which he called *paidia*) to play with explicit rules (which he called *ludus*). He also identified four different forms of play, which are generally combined to create an overall play experience:

- Competition (which he calls *agon*). This is when a player tests their level of skill against another player or the game. *Chess* is an example of play which is mainly agon.
- Chance (which he calls *alea*). This is the element of randomness in play. A game like *Snakes and Ladders* is mainly about agea as are gambling games such as *roulette*.
- Role playing (which he calls *mimicry*). This is the element of pretending to be someone else in play. Fancy dress parties are about *mimicry* as are many forms of children's games where they pretend to be characters such as pirates or shopkeepers.
- Vertigo (which he calls *ilinx*). This is the element of physical sensation or altered perceptions. Roller coasters are about ilinx as is dancing and other physical forms of play.

Lazzaro (2004) interviewed 30 gamers of different types and identified 4 "keys to fun" that cause players to feel emotion during games.

- Hard fun (which she calls *fiero*). This is when players have a sense of challenge and frustration and then triumph and relief as they succeed at a task.
- Easy fun. This is when the game's environment or features provoke curiosity, wonder or surprise. Typically, this is not particularly challenging but provides satisfaction to the players as they discover the game's hidden information.
- Serious fun (which she also calls *excitement*). This is the sense of focus and relaxation which often comes from games with rhythm or repetitive elements such as the need to collect many items.
- People fun (which she also calls *amusement*). This the experience of playing with others which may involve communication, co-operation or competition.

Caillois and Lazarro's ideas can be applied to self-paced games. If we apply Caillois's categories to a game like *Angry Birds* (Rovio, 2009) then it clearly has *competition*, maybe some *chance* and a little *role playing* (the player takes the roles of the birds). Similarly, we can apply Lazarro's keys to the same game. There's some hard fun in the challenge,

some easy fun (curiosity) in seeing what comes next and some serious fun in the excitement of seeing what your bird will do. These are some interesting ways of looking at the features of games and the emotions they can produce. However, it is difficult to see how to develop them further to explain players' engagement in self-paced games.

Blythe et al. (2004) discuss the term "fun" and notes that fun is generally associated with distraction and transgression. They suggest that "pleasure" is a better term for describing the experience of playing games - players willingly agree to abide by the rules and demands of the game. Calleja (2011) also finds the term vague and unhelpful. He argues that it would be better to analyse games in terms of other factors which are more specific and robust. This means that rather than considering play and fun as abstract entities more insight can be gained by looking in detail at the actual games themselves and how they are designed and played.

2.2. Game design and other non-empirical approaches

Rather than trying to analyse the concepts of "play" or "fun" a more fruitful approach to understanding game experience may be to investigate the games themselves. One group of people who have considered games in depth is game designers. Game designers spend considerable time and thought in developing games which create particular experiences. As digital games have developed and become a big industry there has been a rise in game designers reflecting on the process of game design in order to help other designers create better games. This literature is extensive and may provide insights into what creates game experience. Some game designers concentrate on particular aspects of games which they describe while reflecting on their own experience. Koster (2013) describes learning as an important part of fun. Costikyan (2013) describes how uncertainty is an important part of games and looks in detail at the different techniques that games use to create that uncertainty. Cook (2012) describes how the structure of games can be described in terms of repeating "loops" or progression "arcs". Harrigan et al. (2010) look at slot machines and looks for features that would also improve casual games. Schell (2008) extends this already large number of aspects of game design by describing 100 different "lenses" which are ways that game designers can consider their work. These range from the *Emotion* of the player, through the basic *Mechanics* of the game to the ultimate *Purpose* of the designer.

These accounts contain numerous insights into what creates a game experience but these tend to lack empirical validation. They take the view that there are many different routes and aspects to engagement which can be an inspiration for creative production which can then be tested with players. An extension to this is game development methodologies such as "Rational Level Design" McMillan (2013) which is used for designing game levels. This takes the results of initial play testing and uses them to construct a model to predict how players will experience possible level designs.

As would be expected, academic writers take a more conceptual approach which is less concerned with the nuts and bolts of making games and more about trying to extract the experience of playing games into a few key concepts. Salen and Zimmerman (2004) have done a wide-ranging survey and discussion of the approaches taken by other games writers. They consider the "nature of games" in some detail. They are concerned with a wide range of games, particularly board and dice games, so may have particular relevance to self-paced games. "Completist" approaches such as Schell, Salen and Zimmerman suffer from the problem that they consider many aspects of games including game development and the nature of games. This makes it difficult to home in those that are most important to game experience. Other accounts take a "monist" approach, in that they consider the game experience to come from a single factor. So for example, Juul (2013) takes a similar approach to Salen and Zimmerman but concentrates only on the aspect of "failure" in games. This more focused approach can work well in particular situations but are often too simplistic to be applied more broadly.

Rather than consider a list of features that are important for games other researchers have attempted to create models of how games create engaging experiences. Hunicke et al. (2004) attempt to bridge the gap between game criticism and technical game development with their MDA (Mechanics, Dynamics, Aesthetics) framework. This has only three parts, all of which apply to all games so maybe more useful for understanding what games have in common than Salen and Zimmerman's approach. This can also be useful for game design but the simplicity of the framework does not describe the variety of different game experiences as well as Schell's "lenses".

Calleja (2007) has created a conceptual model for understanding game involvement and immersion which he calls the Digital Game Experience Model (DGEM). This is derived from a number of qualitative studies – mainly interviews with players of MMO games such as *World of Warcraft*. He divides the game play experience into six "frames of involvement" which players switch between. These frames are:

- *Spatial* – locating yourself within a larger area than is visible on the screen
- *Tactical* – engagement with decision making such as solving puzzles
- *Affective* – emotional engagement
- *Narrative* – engagement with story elements in the game
- *Shared* – relations with other players in the game
- *Performative* – skill with controls, such as manual dexterity

Each frame can be considered from two phases. The "Macro involvement phase" is the broader context of the game and what attracts people to play it in the first place. The "Micro involvement phase" is the moment by moment experience of the player playing the game. These different frames all contribute to what Calleja calls "incorporation" which is the process by which players' move from having to pay conscious attention to a frame to having it as internalised knowledge. Grip (2010) describes something similar by comparing the "rubber hand illusion" to the way players come to embody their game avatar. In this illusion participants come to feel that a rubber hand is part of their own body such that they flinch if the hand is threatened. Grip sees the way players come to identify with game avatars as a similar process. This is also similar to how

Koster (2013) sees play as a process of learning except that Calleja sees that as players become more skilled they are able to incorporate more frames. For example, once they have mastered the game controls, they can consider the plot. Koster and Schell would argue that a successful game continues to challenge players within the same frame, so for example, *Pacman* remains challenging because the ghosts get faster, not because the game introduces a new frame of involvement. The DGEM is mainly concerned with the different features of a game which make up the experience of playing it. These features make most sense in the context of the Massive Multiplayer Online (MMO) games from which they were derived, this model makes less sense in the context of self-paced casual games. Currently it remains a theoretical model and has no empirical validation apart from the original qualitative studies.

Other researchers have also produced models of engagement based on non-empirical approaches. The *Expanded Game Experience* model (Kultima and Stenros, 2010) considers a wider experience of games than just during play. It includes the process of choosing to play, choosing which game to play, afterplay and choosing whether to play again Similarly, the *Process model of engagement* (O'Brien and Toms, 2008) was initially derived from a non-empirical approach. It starts by describing a point of engagement (when someone starts playing the game), followed by a period of sustained engagement (when they are playing the game), ending with disengagement (when they stop playing). Subsequently there may be re-engagement, leading to another period of engagement followed by disengagement, and so on. There can also be cycles of engagement within a game, where players engage with particular sub-activities, maintain a period of engagement with them, and then disengage, returning to other aspects of the game. The process model has since been developed into an empirically based questionnaire to measure user engagement which shows that it is often difficult to make a hard distinction as to whether an approach to experience is purely non-empirical.

These particular models are interesting because they expand the idea of game experience to include wider factors than just the player playing a single game session. However, this has the net effect of making the experience more complex and difficult to measure than only looking at a single game session which starts after a player has decided to play a game. So, although an ideal measurement of experience may include wider factors than just playing the game, practical considerations of creating an effective, reliable measure may require a more focused approach.

This is a particular issue with non-empirical approaches to how game experience works. These generate a large number of factors and concepts which apply to games and the experiences around them. Some are more relevant to self-paced games than others but without empirical validation it is difficult to judge one approach over another. Non-empirical methods generally do not make testable predictions so are difficult to develop into effective measures of game experience.

2.3. Empirical models of game experience

Game designers and other writers have come up with many ideas about what triggers and sustains players' engagement in self-paced games. However, without empirical validation it is difficult to decide which ideas are the strongest and most important to the experience. McCarthy and Wright (2004) proposed that technology should be considered in terms of the overall experience. However, the experience of playing a game is hard to describe and measure (IJsselsteijn et al., 2007). Game developer Bruce Phillips (Phillips, 2006, p.22) says "I have a secret longing for the confidence in purpose that I imagine my colleagues working on productivity applications must feel. Their goals seem communicable and measurable— mine don't." McCarthy and Wright (2004) felt that the experience of technology should only be considered holistically rather than reducing it into components. Despite this, researchers have created empirically based models of game experience which take a more reductionist approach.

There are many empirically based models of game experience which take a variety of approaches however, most take a similar approach. Typically, they break down the experience of playing games into a number of subfactors, assign scores to these subfactors and then combine the scores to a single numeric value which is a measure of the strength of the experience. Most models of game experience focus on how engaging the game is and if a game scores highly in the model then it can be seen to be have been more successful in its aims. These models suggest frameworks of how to think about games and the experience they create. Looking at how these frameworks were derived shows the inspiration and empirical justification for the model, which may provide inspiration for future studies which aim to develop new measures of the experience of playing self-paced games. The different models also take different approaches to creating and validating the experience measures that they are associated with. Comparing these different approaches allows the validity of the different measures to be evaluated and may also provide inspiration for developing a new measure of game experience.

Many attempts to describe game experience are influenced by the work of Csikszentmihalyi (1991) on "Flow". Csikszentmihalyi interviewed people engaged in a range of immersive tasks such as rock climbing or carpentry. For all of these tasks he identified a state of full immersion which he called a "flow experience". Csikszentmihalyi identified nine characteristics of a flow experience namely: intense concentration, merging of action and awareness, loss of self-consciousness, a sense of control, distortion of time perception, a balance between challenge and skill, seeing the activity as intrinsically rewarding and clear goals. All of these must be present to experience flow. Initially (Csikszentmihalyi, 1991) saw that flow was a binary state, people were either in flow or not, but in later works (Csikszentmihalyi, 2013) he discusses "flow like states" which implies that it may be possible to be partially in flow.

Csikszentmihalyi and Larson (1987) discuss using experience sampling methods to assess when people were in a flow state. Typically, people are asked to complete a short questionnaire at random intervals of around 90-120 minutes. The questionnaire is designed to take less than 2 minutes to complete and starts by asking "what are you doing right now" and then other questions about mental states. This technique has also been extended by game researchers such as Kaye et al. (2018) who had players of a shooting game fill in a flow questionnaire every hour and found a relationship between positive mood and feelings of flow.

There have been many approaches to developing models and measures of game experience. These have different degrees of empirical underpinning and different approaches to deriving concepts from the empirical data they do collect. On the surface Sweetser and Wyeth (2005)'s GameFlow model seems most directly influenced by the concept of flow. They identify eight aspects of flow and from these derive 8 elements which go to make up GameFlow: *concentration, challenge, skills, control, clear goals, feedback, immersion,* and *social interaction*. Some of these such as control are a direct mapping from elements of flow; others such as social interaction are less direct. For each element they then produced 2-7 criteria which make up that element. To assess how much players will enjoy a game they go through each criterion and score it from 1-5. There are a number of issues with this approach. By relying on inspection of the game the end result is highly subjective, particularly as knowledge of the criteria may cause people to make post-hoc justifications such "I liked this so I'll give it lots of high scores". The elements they have chosen will apply differently to different games – so some games will be highly social but others not at all. By giving different numbers of criteria to each element the scores will be biased towards those with more criteria. For example, the clear goals element only has 2 criteria whereas the controls element has 7, so a game with no clear goals could still score well if the controls were good. Despite these issues GameFlow has been further developed into a set of heuristics based on reviews of real time strategy games (Sweetser et al., 2012). This assumes that issues discussed in reviews are those which actually make a difference to game enjoyment and the resulting heuristics tend to be unfocussed, numerous and concentrate on adding more complex features rather than the actual player experience. However, the biggest issue is that GameFlow does not consider different motivations that different types of player may have. GameFlow treats "game enjoyment" as a product of the features of the game rather than how a particular player relates to the game.

Calvillo-Gámez et al. (2010) investigated game experience by performing a Grounded Theory (Charmaz and Belgrave, 2007) investigation on game reviews and interviews with players and game designers. From this they came up with a theory known as the "Core Elements of Gaming Experience" (CEGE). The top level of the theory is called "Puppetry"(Calvillo-Gámez and Cairns, 2008) which states "the player's interaction with the game is formed by the player's sense of control and ownership. Control produces ownership, which in turn produces enjoyment." From the theory they created a questionnaire that was successfully used to distinguish between similar but different gaming experiences (playing Tetris with different controllers). The use of game reviews

is interesting because it creates an additional layer of interpretation between game players' experience and the theory. Game reviewers have their own agenda which is often to entertain or reinforce their audience's prior conceptions (IGN Staff, 2001, Davies, 2011). The reason for using reviews as well as player interviews is that game players are often unable to describe their experience whereas game reviewers have a particular vocabulary and terms which they feel are important (Phillips, 2006). This gives one view of the game experience, but there may be others – a film review may discuss the story and the lead actor but a film director may see the film in terms of shot grammar or *The Hero's Journey* (Campbell, 1987). Calvillo-Gamez et al only considered console and PC games rather than casual games but the idea of the game being "like a puppet which is brought to life by the player" may apply even more strongly to self-paced games.

Some other approaches take the "monist" principle that there is one key aspect game experience that measures should be based on. Qin et al. (2009) consider that "narrative is the basis or framework for computer games". Even for genres such as fighting games they see that the narrative is the story that players tell themselves about what happened. They consider the different ways that narrative can be explicitly constructed in games which range from explicit linear storytelling to the emergent narratives of "sandbox games". They do not mention which types of games they were investigating but do say that they were not interested in board games like chess and mah-jong. They were mainly interested in student game players and 70% of their respondents were students with most of the remainder being graduates. After conducting a detailed survey of existing game experience research, they created a questionnaire and adjusted it after validation on-line with a large number of participants. The final questionnaire has 27 questions divided between seven factors; *Curiosity*, *Comprehension*, *Challenge and skills*, *Empathy*, *Concentration*, *Control* and *Familiarity*. However, the questionnaire is very heavily aimed at games with a strong explicit narrative and may have limited application in other types of game.

In another monist approach Slater et al. (1994) defined *presence* as a participant's sense of "being there" in a virtual environment. As such it relates to situations where a person is represented in a 3D world. This applies to some games but also other virtual reality technologies such as *Second Life* (Linden Lab, 2003). Slater developed a questionnaire to assess presence and also to investigate the idea of "stacking" virtual environments - such as when you are in a virtual world and you start playing a computer game. This form of presence is often seen as being an important measure of game experience by researchers (Calleja, 2007, Mäyrä and Ermi, 2005) who work with 3D game genres such as first-person shooters (FPSs) and massive multiplayer online games (MMOs). Self-paced games rarely involve virtual environments. Those that do, such as *X-COM : Enemy Unknown* (Firaxis, 2012) are not interactive in the same way and may not generate presence in the same way. Cairns et al. (2014a) ran an empirical study and showed that increasing presence does not necessarily increase immersion. Lombard and Ditton (1997) extended this concept of presence and identified six different forms of presence. Three are spatial: *realism*, *transportation*, *immersion*. Three are social: *social richness*, *social*

actor within medium, *medium as social actor* (AI). They did not develop any empirical method for measuring these forms and it is arguable that by extending presence in this way they have diluted the concept and made it more difficult to discuss rather than providing a useful measure of game experience.

Sometimes it is possible to use a sound theoretical basis and a sophisticated multi-layered empirical process but still end up with problematic measure. Fang et al. (2010) developed a questionnaire to measure enjoyment during computer game play. This was inspired by Nabi and Krcmar (2004)'s model of media enjoyment which considers three levels *Affect*, *Cognition* and *Behaviour*. They initially had 66 questions to cover each of these areas. They then consulted professional game designers, performed a card sorting exercise with students studying games and an online test with 307 respondents and finally used a factor analysis to reduce the number questions to 11. A further online survey with 508 respondents validated this final questionnaire. This model has been extensively statistically validated and has the advantage that the questions refer to mental states and how the player feels rather than particular game features. Despite this, some of the questions seem a little surprising. For example, participants are asked how much they agree with this statement "The activities in this game or the actions of its character(s) are decent." This is not a question that I would personally know how to answer about any game which suggests that some aspects of the final questionnaire are problematic

Another empirical basis for game experience models is to use focus groups. Poels et al. (2007) used focus groups to talk to 16 game players about their game playing experiences. Interestingly, they spoke to a range of different gamers including both those who played regularly and less regularly. They do not give a complete list of the types of games which were discussed but they seem to be mainly console and PC games rather than mobile or casual games. They then build up a set of categories that describe the gaming experience both during play and afterwards. These were *enjoyment*, *flow*, *imaginary immersion*, *sensory immersion*, *suspense*, *competence*, *negative effect*, *control* and *social presence*. A strength of the focus group approach is that it generates a wide range of responses and categories to be considered. However, without some other form of validation it can be difficult to assess whether all the categories are needed and how robust they are. Hudson and Cairns (2014) have criticised focus groups for several issues, most notably that participants have not just played a game so it is difficult to know which particular game experience they are referring to. Poels et al did develop a Gaming Experience Questionnaire (GExpQ) which has been published by Nacke (2009). It has not been widely used empirically but Nacke did find the results correlated with some psychophysical measures. Law et al. (2018) did a large-scale validity study of the GExpQ and found no evidence for the originally postulated factor structure. They also examined the literature which cited this model and found that several versions of this model have been published in a variety of formal and less formal channels which has led to confusion and inconsistency in which versions of have been cited and used.

Brockmyer et al. (2009) also used focus groups as a basis for their questionnaire but they also used additional validation technique. Their Game Engagement Questionnaire (GEngQ) was developed as a self-reported measure of an individual's potential for becoming engaged in video game-play at differing levels. This differs from many other engagement measures in that it assesses an individual player's likelihood of becoming engaged by any game rather than how engaged they were in a particular game during a particular play session. To develop the measure, they consulted the literature and conducted focus groups before putting together a 10 question pilot questionnaire with each question rated on 5 levels. This was then expanded to 15 questions with each question rated on 3 levels. ("Yes", "No", "Sort of"). After a Rasch rating scale analysis this was expanded to 19 questions. They then validated this with 154 high school students. This was then further validated by conducting an empirical study to test how well the questionnaire predicted engagement while playing the first-person shooter game *S.T.A.L.K.E.R.: Shadow of Chernobyl* (GSC Game World, 2007). Participants were 107 male undergraduates who played for 30 minutes. After 25 minutes they heard a voice saying variants on "Excuse me, did you drop your keys?". They heard this 3 times at increasing volume. They found that scores on the GEQ predicted how likely participants were to react to the voice. Participants with a high GEQ were less likely to react. The GEQ has been substantially statistically validated and the empirical validation method is interesting and may point to using distraction as a measure of game engagement.

An alternative to focus groups is to use player interviews as an empirical basis for deriving a game experience measure. Mäyrä and Ermi (2005) proposed a game play experience model with three different kinds of immersion: sensory, challenge-based and imaginative. They developed this model after interviewing game playing children and non-playing adults. They created a questionnaire which initially had 30 questions. After testing this with 193 online participants they validated it down to 18 questions. They then used this questionnaire to classify 13 games according to the amount and types of immersion they created in players. The game *Half-Life* (Valve, 1998) came out as having the highest level of immersion and *The Sims 2* (Maxis, 2004) as the lowest. *The Sims* is the highest selling game franchise in game history (Howson, 2008). Mäyrä and Ermi do say it would be a mistake to say that *Half-Life* is a "better" game than *The Sims*. This may be because although *Half-Life* is more immersive, this immersion is only effective for a particular audience, whereas *The Sims* appeals to a wider range of players. Mäyrä and Ermi do note the "casual" (Kultima, 2009) nature of *The Sims* and it may be that casual self-paced games have lower immersion or need a different type of experience model.

The term "immersion" is often used by players, designers and reviewers when discussing game. Brown and Cairns (2004) conducted a qualitative investigation to study what players mean by this term. They conducted semi-structured interviews with 7 participants after they had been playing a game for 30 minutes and then used Grounded Theory to analyse what they said. As might be expected the term "immersion" was used to describe the degree of involvement with the game. This involvement changes with time and can be altered by various barriers, some of these are

related to the game itself such as the controls. Others are related to the player, such as the amount of time they are prepared to invest in the game. Brown and Cairns identified three levels of immersion which range from engagement through engrossment to full immersion.

Jennett et al. (2008) developed a questionnaire to measure the level of immersion in games. Some of the questions were inspired by Agarwal and Karahanna (2000)'s dimensions of *cognitive absorption*. Agarwal and Karahanna define cognitive absorption as a "state of deep involvement with software" which is exhibited by five dimensions which are *temporal dissociation*, *focused immersion*, *heightened enjoyment*, *control* and *curiosity*. Other questions were derived from the earlier immersion study by Brown and Cairns (2004). These related to emotional involvement, transportation to a different place, attention, control and autonomy. The initial questionnaire had 33 questions. They then performed two studies to explore how well the questionnaire functioned. The first study was based on the hypothesis that highly immersed players will find it more difficult to move to another task. 40 participants were split into two groups. All of them did a tangram puzzle task. One group then played *Half-Life* and the other did a box clicking task. Half way through, each group filled in the questionnaire. At the end of the play session they did the tangram task again. The immersion scores and differences in time on the tangram task were correlated across the two conditions which supported the hypothesis. However, the immersion score for the box clicking task did not correlate with the tangram task suggesting that either the questionnaire does not measure immersion well in non-game tasks or that some non-immersion factor is affecting the tangram task.

The second study involved two groups of 20 participants. One group played *Half-Life* while the other did a box clicking task. Both groups completed the questionnaire. Fixation data revealed that participants' eye movements significantly increased over time in the non-immersive condition. In contrast, participants' eye movements in the immersive condition significantly decreased over time. The questionnaire was redesigned with simpler wording and ended up with 31 questions. This was tested online with 244 participants. Analysis of the results validated the questionnaire and a factor analysis revealed five factors which were named *Cognitive Involvement*, *Real World Dissociation*, *Challenge*, *Emotional Involvement* and *Control*. A further study with a box clicking exercise showed that immersion increased with the speed of the task and the highest immersion was produced by a task in which the speed gradually increased.

The final questionnaire is known as the Immersive Experience Question (IEQ). A particular strength of this immersion model is that the five factors that make it up are not features of the game as in Calleja (2007)'s DGEM model. They are mental states and so can, in theory, be created by any game. The only exception here may be *Real World Dissociation*. All types of games can create the "losing track of time" effect but faster more graphically intense games may show a stronger effect on the questionnaire results.

Jennett et al's work validates the questionnaire statistically but also empirically using different experimental methods. These experiments did reveal several issues with this type of research. In the first two studies participants played the game *Half-Life* but the average immersion scores were lower than might have been expected. This was because a significant proportion of the participants had never played the game before and never managed to master the "WASD+mouse" controls commonly used by the First-person Shooter (FPS) game genre. Conversely, immersion scores were much higher for the "clicking boxes" task than would have been expected. Some participants who played this task said they made up their own game to click on the boxes as fast as possible.

Immersion has been further investigated in a number of empirical studies. It has been found to be affected by a number of game factors including screen size (Thompson et al., 2012), music (Sanders and Cairns, 2010), lighting levels (Nordin et al., 2014), camera position (Denisova and Cairns, 2015) and player controllers (Cairns et al., 2014b). Immersion has also been linked to game addiction (Seah and Cairns, 2008) and basic attempts made to inform game design (Huhtala et al., 2012). Nordin et al. (2013) found that time perception in games was related to immersion but concluded that the story was a complex one and more research is needed.

Cairns et al. (2013) found that immersion is increased in social games if players believe that they are playing against another person rather than a computer. Their results also indicated that it does not make any difference to immersion if the other player is located in the same room or remotely. This may vary depending on the game; it is hard to imagine that a remotely played game of *Spaceteam* (Sleeping Beast Games, 2012) would be quite the same. This is a general issue with this type of empirical study, just because they show that making a particular design change increases immersion in one particular game it does not mean that that design change will increase immersion across all games.

Most of these empirical models of tend to be derived from players' own "folk psychology" (Malle and Knobe, 1997) rather than established psychology models. Ryan et al. (2006) based their model on concepts from the *self-determination theory* (SDT) of intrinsic motivation (Deci and Ryan, 2008). In particular, they made use of the concepts from a subset of SDT known as *cognitive evaluation theory*. This maintains that players have psychological needs for *autonomy* and *competence* that they satisfy by playing games. Ryan et al created a questionnaire based on these concepts together with some other concepts such as the need for "intuitive" controls. This questionnaire is known as PENS (Player Experience of Need Satisfaction). They then tested this questionnaire on action games such as *Super Mario 64 (Nintendo, 1996), Zelda: The Ocarina of Time* (Nintendo, 1998) and *A Bug's Life* (Travellers Tales, 1998*).* They also validated it using 730 players of Massive Multiplayer Online games such as *World of Warcraft.* The PENS questionnaire takes the view that the most interesting part of player engagement is the actions that players perform due to that engagement. For example, which games they choose to play and how long they choose to play them. This is in contrast to other game engagement measures such as the Immersion Experience Questionnaire which are more interested in players' internal mental state.

Denisova et al. (2016) were intrigued by the large number of game experience questionnaires and performed a study to investigate the degree of correlation between three of the most popular; the IEQ (Jennett et al., 2008), the GEQ (Brockmyer et al., 2009) and PENS (Ryan et al., 2006) (The development and validation of these questionnaires is discussed on pages 39-41). They found a large degree of correlation between these questionnaires. In particular, the IEQ had high correlation with both the GEQ (r=0.804) and PENS (0.813). This suggests that these questionnaires may be measuring very similar underlying factors and for practical purposes are fairly interchangeable.

There have also been investigations into particular tools to measure particular aspects of the game experience. *Challenge* has been seen as an important aspect of game experience by both game designers (Adams, 2014) and academic researchers (Jennett et al., 2008). Denisova et al. (2017) have on-going work to develop a questionnaire to measure the degree of challenge in games. *Uncertainty* has also been seen as an important part of the experience of playing games (Costikyan, 2013, Abuhamdeh et al., 2015). Power et al. (2018) have developed a questionnaire to measure the degree of uncertainty in games which is based on five different factors which underpin players' feelings of uncertainty.

In summary, there are many different attempts to create models and tools for measuring and understanding game experience. The use of empirical studies to test models and questionnaires tends to reveal unexpected findings "around the edges" which give greater insight and further validation to the model. Of the models that I have considered, only PENS is derived from a psychological model and even so the derivation is fairly loose without the strong empirical evidence available from established cognitive psychology models. This makes it difficult to validate these models of game experience against other psychological findings and can also reduce their predictive power. Many of these models start with first-person shooter or massive multiplayer online games and attempt to generalise these genres across all games. They also tend to choose participants who are self-declared "gamers" and are highly skilled and motivated at video games. Mekler et al. (2014) performed a systematic review of quantitative studies on the enjoyment of digital games. Out of 87 studies almost all of them were action games; the majority were first-person shooters, racing games or sports games. Only 5 were on self-paced games. This may mean that these models do not fit well with self-paced games or the "casual" game players who play them. Exceptions are Poels et al. (2007) and Cairns et al. (2014b) who did briefly investigate casual gamers and mobile games. This was not in any great detail and there is a need to do more empirical validation of these models with self-paced games. Despite these issues, questionnaires remain the most widely used and tested measures of game experience. Any new measure of experience will need to be "benchmarked" (Kline, 2013) against existing measures and using a game experience questionnaires would be a robust and practical method of doing this. See chapter 3 for more details on this.

There are also fundamental issues with using questionnaires as a method to investigate game experience. Questionnaires are a "self-reported" measure, they rely on

participants filling out the questionnaire truthfully to give an accurate account of their experience. Although questionnaires can be validated to ensure their construct validity (Kline, 2013) there is still the problem that participants may not consider their whole experience when making their responses. For example, Kahneman et al. (1993) found that participants tended to only report the peak or the end point of their experience giving rise to the "peak-end" rule. Another key issue with all of these questionnaire-based methods is that they only measure the whole gaming experience "after the fact". It would be much more useful for game design if a new measure could indicate the changes in engagement over the time of play. That way designers could see which parts of the game play need improving and which are already successful. This would also help with investigations in to which aspects of games create which experiences. To give greater insight into which parts of a self-paced game create which experiences we would need a continuous measure which would show how the level of game engagement changes over time. One approach which attempts to solve both the problems of self-report and also offers a continuous measure of experience is to measure the physiological properties of players' bodies while they play the game. This may be a promising approach to constructing a new measure of game experience so will be considered next.

2.4. Physiological measures of game experience

Physiological measures of game experience have a number of potential advantages over self-reported measures such as questionnaires. They are less mediated by participants' mental states or their interpretation of the question. They can also be continuous measures providing measurements which relate directly to individual game events rather than the whole experience.

There are four major physiological measures which have been applied to game experience; Skin conductance, Facial Electromyography, Heart rate and Electroencephalography.

- Skin conductance is a measure of how much your skin conducts electricity. It is usually known as Electrodermal activity (EDA) but in the past has been known as galvanic skin response (GSR). Lang et al. (1993) showed that it could be used to measure emotional arousal by showing participants a series of pictures with different emotional effects. EDA responses are not instant; they take between 1-4 seconds to appear but usually have less noise and are easier to interpret than other methods such as facial muscle or heart activity (Kivikangas et al., 2011).
- Electromyography (EMG) measures the electrical activity of skeletal muscles. Facial EMG measures the activity of muscles used for facial expressions of emotion. Bolls et al. (2001) successfully used facial EMG to measure the emotional valence of participants who listened to radio adverts. They also used it to find that emotional arousal was a much better predictor of how much

participants remembered than valence. EMG is sensitive to noise both for technical reasons and also from other muscle activity such as speaking (Kivikangas et al., 2011).

- Heart rate is simply a measure of how many times your heart beats in a minute. Ravaja (2004) reports that it is linked to attention, effort, arousal, and emotion and that it has been shown to be a good measure of short-term attentional selection and long-term attentional effort. However, heart rate is also linked to many other bodily processes so interpreting the data can be difficult.
- Electroencephalography (EEG) measures brain activity by recording electrical activity along the scalp. EEG is normally divided into "bands" which correspond to the different frequencies of signal. These are named *Alpha, Beta, Gama, Delta, Theta* and *Mu* (Tatum, 2014). Studies have shown that frontal *Alpha* activity is linked to emotional responding. EEG is similar to heart rate in that it is linked to many bodily processes and activities so it may be difficult to attribute a particular signal to a definite cause.

There are some other physiological measures which have been used to look at game experience. For example, Van Den Hoogen et al. (2009) found that the amount of force that game players applied to their game controller increased with sensory immersion although the amount they tilted the controller did not. However, this measure has not been widely studied so may not generalise to other games, particularly those using different controllers.

One of the main issues with these measures is interpreting the data produced by these methods into terms which describe the player experience. A possible approach to this is to map physiological measures of game experience onto models of emotions. Russell (1980) and Mehrabian (1996) describe models of emotions in which a wide range of emotions are represented by two or three dimensions. The "core effect" of the emotion is described by two dimensions. *Valence (*or *positive)* specifies how positive or negative the emotion is and *arousal* is roughly equivalent to the strength of the emotion. The third dimension *dominance* is part of the cognitive appraisal process of an emotional event. Several physiological measures such as heart rate and skin conductance have strong links to *arousal* but other links are less clear.

Research on using physiological measures to measure game experience can be divided into those studies which look at measuring the whole experience of playing a game and those which focus on measuring the experience of particular game events on the player. These two groups are described below.

2.4.1. Overall game experience

Several researchers have used a variety of psychophysiological methods to assess the complete experience of playing a game. Nacke (2009) performed an experiment where participants played three different level modifications of *Half-Life 2* (Valve, 2004). He named these levels after the game experiences that they were designed to create; *immersive, flow* and *boredom.* The *immersive* level contained varied environments and varying challenge and relief. The *flow* level had increasing difficulty with a focus on the

mechanics of the game. The *boredom* level just repeated the same enemies with no change of difficulty. He measured skin conductance, EMG, EEG and challenge using a questionnaire. The EMG and skin conductance showed the highest reaction in the *flow* level, followed by the *boredom* level and then the *immersion* level. EEG *Delta* showed *immersion* highest followed by *flow* and *boredom*. EEG *Theta* showed *immersion* highest followed by *boredom* and then *flow*. The questionnaire showed that challenge was highest in the *flow* level followed by the *immersion* level and then *boredom*. These results indicate that different psychophysiological methods can measure different parts of the game experience but it is still not clear what these measures actually relate to.

Burns and Fairclough (2015) also looked at EEG readings as a measure of game experience. In particular they looked at "Event-related potentials" (ERPs) which are a representation of the average changes in the EEG signal in response to having perceived some stimulus. 20 participants played the driving game *WipeoutHD Fury* (Sony Studio Liverpool, 2009) at three different levels of difficulty. For each level of difficulty, they played 4 races making 12 races in all. During each race they heard 110 short audio tones. Of these tones 90 were "standard" tones and 20 were "oddball" which were played in a random order. Participants also completed an Immersion Experience Questionnaire for each level of difficulty. This showed that participants found playing the game on the easy difficultly level less immersive than the other two levels (hard and impossible). The amplitude of the ERPs during the oddball tones was also less during the harder two levels than for the easy level. This seems to show that ERPs can be correlated to immersion. The reasons for this are unclear but it may be that participants' attention was held more strongly during the two hard levels of difficulty which stopped them being so distracted by the oddball tones.

Ravaja (2009) asked participants to play remotely against a computer, a friend and a stranger. They measured emotion used EMG and arousal using skin conductance. They found that playing against a human gave more positive emotional effects and arousal. Playing against a friend gave more positive effects than a stranger. Yannakakis and Hallam (2008) measured children's heartrate, blood volume pressure and skin conductance as they played a physical game. They used this data to train an artificial neural network which then had some success in predicting how "fun" the children would report the game to be. However, as children were playing a physical game it is difficult to know whether these psychophysiological measures were part of the player's game experience or just the physical effects of the game.

Mandryk and Atkins (2007) describe a system which combines four different physiological measures (skin conductance, heart rate and two different forms of EMG). They used literature reports of these measures to build a "fuzzy logic" model which combined these measures into readings for *arousal* and *valence.* They then created another fuzzy logic model based on the circumplex model of emotion (Russell et al., 1989) which converted these readings into measures of emotions such as boredom, challenge, excitement frustration, and fun. This approach seeks to overcome the difficulty of reliably relating physiological readings to psychological states. However,

this study only involved a small number of participants (12) who played the ice-hockey based action game *NHL 2003* (EA Sports, 2002). From this study it is difficult to know how significant or accurate the measure is and to what extent they generalise to other games. It seems likely that using several physiological measures at the same time may reduce the likelihood of interference from other non-game sources changing the results. However, they will still suffer from some of the same issues such as the time lag for skin conductance and that measures such as heart rate are unlikely to be effective for self-paced games. This type of multi-measure fuzzy logic model is only possible due to previous studies on each of individual measures which used to build up the multi-measure model.

These studies illustrate that there are many different physiological measures which correspond to game experience. The difficulties are in telling which ones correspond to which parts of the experience and how to translate this into terms which tell us something broader about game experience. Multi-measure studies may make this process more reliable but they rely on previous studies on each of the component measures. Future research with a wider variety of games and participants may allow more definite links to be made but currently it is difficult to know how much significance to attach to any particular result. As with the models of game experience, most of the research has been done on action games. These are, by definition, faster paced, more intense and have more sensory immersion than the self-paced games that I am interested in. This indicates that these findings may not generalise to self-paced games. However, the slower pace of self-paced games may work well with measures which take time to react, such as skin conductance or measures which are otherwise affected by other factors such as heart rate.

2.4.2. Game events

Other researchers have made use of physiological methods to investigate events within a game. Ravaja et al. (2006) measured EMG, skin conductance and cardiac interbeat intervals while participants played the game *Super Monkey Ball*. (Amusement Vision, 2001) They coded game events and found that most positive game events (such as picking up bananas) were associated with fast "phasic" changes in the physiological measures. They also found that some events which could be seen as negative (such as falling off the board) were also associated with these phasic changes.

Weber et al. (2009) monitored participants for 50 minutes whilst they played the first-person shooter game *Tactical Ops: Assault on Terror* (Kamehan Studios, 2002). They measured heart rate and skin conductance. They divided the game into 7 different phases of play (such as *safe*, *firing* etc.) and 18 different events which can lead to one phase changing to another. (E.g. opponent disappears from the screen). They found a significant correlation between a particular game event and the heart rate of players. There was no significant correlation between skin conductance and the game event although the data seemed to point towards a small correlation which may be significant with more participants. These results could be interpreted as that both heart rate and skin conductance were measuring arousal caused by the excitement of the game. The

game events they were looking at could be quite short so the delay in the skin conductance effect may make it less accurate which would explain lack of significant correlation.

Ambinder (2011) describes how the game company Valve used skin conductance to tune an artificial intelligence director in the *Left 4 Dead* (Valve South, 2008) game series. The goal of the director is to introduce enemies dynamically to give players periods of intense gameplay followed by less intense relaxation periods. Valve measured skin conductance during gameplay and then correlated this with game play events. They then used Principle Component Analysis (PCA) to create a model of the game factors which affected players and used this to create a more enjoyable game (as measured by questionnaires). Unlike Weber's study above this study found that skin conductance was a useful measure. This is likely to be because what counted as a "game event" in *Left 4 Dead* was a wave of enemies rather than an individual enemy event. The meant that game events in *Left 4 Dead* (Valve South, 2008) lasted a lot longer than those in *Tactical Ops: Assault on Terror* so the 1-4 second delay in skin conductance response did not present a problem.

Some researchers have used a hybrid approach in which they combine physiological methods with post-game interviews or video replay. McAllister et al. (2013) measured skin conductance during a driving game and then looked for peaks and troughs in the recorded signal. They then used video replay of the game to ask players about their experiences during particular peaks or troughs in the skin conductance. This technique may be particularly useful for game development as it gets richer data from players about the times in the game when they either have problems or experience particular pleasure.

These studies show that physiological methods can be used for continuous measurement of game experience. However, even though researchers saw an association of particular game events with particular psychological measures, it is often difficult to say what this says about that particular event. The main exception to this is in measures of arousal. Arousal is closely related to tension, excitement and playing intensity and these terms can be easily mapped onto player experience. However, self-paced games are not typically associated with excitement or intensity. They do involve tension, so that it would be interesting to see whether measures of tension work the same way as for faster action games. Although physiological measures can be used for individual game events the length of the event in an action game may be too short for some measures to pick up. Self-paced games are usually much slower so this is less likely to be problem and presents an interesting opportunity for further research.

2.5. Approaches to developing a new measure of the experience of playing self-paced games

The previous discussion shows that questionnaires to measure game experience are well established and there are a wide variety to choose from. There may be room for a new questionnaire based on the specific experience of playing self-paced games but it is also possible that the differences in experience that participants report between self-paced and action games is not that large. As such a new questionnaire-based measure would not be particularly distinctive and would suffer from the same issues of existing questionnaires, namely that they are self-reported measure which does not offer the possibility of a continuous reading across the time of the game.

There seems to be more room to make a distinctive contribution using some type of physiological measure. Playing a self-paced game is physiologically a different experience from playing a fast-paced action game and it seems likely that there will be different physiological effects. One of the issues with existing physiological measures is that there is not always a clear link between the physiological measure and an associated psychological property that relates to the game experience. To create a new measure of the experience of self-paced games I considered the cognitive processes that happen while the game is played and looked for associated physiological properties which are driven by those processes. As these properties are linked to cognitive processes that happen during gameplay, measuring them should provide a measure of game experience which is both continuous and not self-reported. The two cognitive processes I decided to consider are *cognitive load* and *attention.* Both of these are features of self-paced games and there is the potential to measure them using continuous measures which are not subject to self-report.

2.6. Measuring cognitive load using pupil dilation

Typically, self-paced games tend to be either puzzle games like *Two Dots* (Playdots, 2014) or strategy games like *Civilization* (Microprose, 1991). Both of these types of games place a premium on mental effort and reasoning to complete challenges and make progress through the game. Cognitive load is a psychological concept which describes the load on cognitive control processes such as working memory (Sweller, 1988). It approximates to the amount of "mental manipulation" or "abstract thinking" that you are doing at a particular time. Baddeley and Hitch (1974) found that being under high cognitive load impaired participants' ability to memorise information. Cognitive load

has also been linked to the use of short term memory and how likely participants are be distracted from a task (Lavie, 2005, Lavie et al., 2004). The load theory of attention suggests that increases in cognitive load make participants more likely to be distracted. Digital games are notable for holding players' attention so if they do this whilst requiring high cognitive load then there may be some other mechanism involved which negates the distraction effect that would otherwise occur. Cognitive load is also required for many tasks such as filling in tax returns and mental arithmetic which are typically not seen as being engaging as playing games. Being able to measure the patterns of cognitive load use during gameplay may give insight as to why this is the case.

When participants are given a cognitively demanding task their pupils increase in size (Kahneman and Beatty, 1966), this is known as pupil dilation. Pupil size has been used for measuring processing load during reasoning, language and attention tasks (Beatty, 1982). The amount of pupil size change is related to the difficulty of the task (Jainta and Baccino, 2010) and during prolonged decision making the strongest effect is throughout decision formation, not at the end (de Gee et al., 2014). Pupil dilation has been used to measure cognitive effort in a wide range of tasks including the "n-back" task (Katidioti et al., 2014) Perceptual rivalry, Target detection, Digit-span memory and Mental multiplication (Hossain and Elkins, 2016).

Changes in attention to a brighter or darker object can change pupil size even if the pupil itself does not move (Binda and Murray, 2014). There is also evidence that higher level processing can affect pupil size – participants' pupils dilated less when they were shown a picture of the sun than a pattern of equivalent brightness. Einhäuser et al. (2008) showed participants ambiguous visual and audio stimuli such as a Necker Cube. Pupil diameter increased just before making the perceptual switch. Iqbal et al. (2004) used pupil dilation to assess the cognitive effort needed for common HCI tasks such as searching through lists and object manipulation. They found that the average pupil response for the whole task was unchanged but if they segmented the task into smaller sub-tasks then some of those sub-tasks which needed more cognitive effort showed heightened pupil response.

To be able to measure how much cognitive load a player is using at any given time in a game could give interesting insights into the player's experience of the game. In usability focused approaches to human computer interaction an interaction which requires higher cognitive load generally indicates lower usability as users need to put more cognitive effort into understanding the interface (Pomplun and Sunkara, 2003). Although the amount of cognitive effort that players put into playing a game is part of the experience of that game, there may not be a straightforward relationship between the amount of effort and how engaged they are in the game. Currently, we have very little insight into how cognitive load is linked to a players' engagement during self-paced games. It may be that more engaged players put more effort into the game and so use more cognitive load. Or the relationship may be more complex than that, so it may be that there is an optimum level of cognitive load which leads to greater engagement

or particular patterns of cognitive load use which lead to more engagement. For this reason, I will be considering that measures of cognitive load are measures of the experience of playing that game but not necessarily measures of how engaging that experience is for the player. Whatever the true picture being able to measure cognitive load during puzzle and strategy games could give insights into player's experience of those games and how it creates engagement.

It may be possible to use pupil dilation as a measure of cognitive load during game play. Pupil size changes in response to different events in a predictable way, although the size of the changes due to cognitive load can be small compared to changes due to other factors. Beatty and Lucero-Wagoner (2000) report that pupil changes due to shifts in cognitive state tend to be in fractions of a millimetre whilst changes due to differences in arousal levels are around 1mm and changes due to differences in light levels can be as high as 2-8mm. Measurement of cognitive load using pupil dilation also needs to consider issues of timing and lag between stimulus and response. Task evoked pupil responses start between 2-300ms after the stimulus (Beatty, 1982, Gagl et al., 2011) but may peak around 1200ms. Richer and Beatty (1985) also found that pupil dilation started around 1.5s *before* an associated motor action but peaked 0.5s after the action started.

Various investigations have been made to find neuro-correlates with pupil dilation. Pupil dilation has been linked to the activity of areas associated with attention and executive function (Aston-Jones and Cohen, 2005, Gilzenrat et al., 2010). Costa and Rudebeck (2016) and Joshi et al. (2015) found that pupil dilation was linked to a wide variety of brain areas which are associated with many different functions. Shine et al. (2016) found that pupil dilation was associated with integration across disparate neural regions. This suggests that pupil dilation can be influenced by a wide variety of cognitive activity which would be consistent with other studies showing change in pupil size due to due to differences in light levels, emotional states such as arousal (Loewenfeld and Lowenstein, 1993, Beatty and Lucero-Wagoner, 2000) and motor actions (Richer and Beatty, 1985). These factors may add noise to any measurements of cognitive load. Minimising the effect of these noise factors is an important part of designing experiments to measure cognitive load using pupil dilation.

2.6.1. Pupil dilation experimental design

Because pupils dilate for several different reasons it is important to use a careful experiment design to accurately measure cognitive load. Kahneman and Beatty (1966) used a simple protocol for their short-term memory task. Five participants heard strings of 3-7 digits at one second intervals. There was then a pause for two seconds and the participants then had to repeat the digits they had just heard in the same order. An image of their pupil was taken every second. During the task participants fixated on a faint grey circle printed on a white card. If participants had to recall more digits then their pupil size changed more. Kahneman and Beatty (1966) also used a similar protocol to investigate processing load. They asked participants to recall four digits but to add 1 to each of them. This resulted in larger pupil sizes than the basic recall task. It is

important to note that these initial studies were audio only, so that light levels on the pupil could be kept constant and not create spurious dilations.

Richer and Beatty (1985) asked participants to simply press a button when they judged that 5-10 seconds had passed. Participants showed a pupil response from around 1.5 seconds before their motor action which peaked about 0.5 seconds afterwards. The dilation increased with the complexity and force of the movement. Richer and Beatty concluded that the act of simply making a movement on its own creates a pupil response regardless of cognitive state. Moresi et al. (2008) asked participants to respond using several fingers or using different hands. They found that the more complex responses created larger pupil effects. These studies indicate the simple act of pressing a button or touching a screen could create pupil dilations, an effect that may cause problems when using these techniques during actual game play.

Cavanagh et al. (2014) measured pupil dilation during a screen-based decision-making task. Participants sat 65 cm from the screen and were prevented from moving using a stationary head mount. 4 participants were rejected from the experiment leaving a total of 20. They initially asked participants to fixate on a cross for 1 second before showing the stimulus. All stimuli were balanced to be of equal total luminance throughout the experiment so the amount of light reaching participants' eyes is constant throughout the experiment. This is known as luminance balancing and is done to avoid changes in light levels affecting the pupil size and confounding the results. The experiment was a between-subjects design in which the pupil responses from different sets of stimuli were compared. Participants in all conditions made the same motor movements. This allowed pupil dilation data between conditions to be compared as the noise from luminance and motor effects would be the same for each condition. This study is interesting because it shows that pupil dilation can be used to measure cognitive load during a screen-based task in which participants respond with a motor action.

2.6.2. Data analysis

Modern eye tracking equipment measures the size of each eye between 30-200 times a second. Participants may have different sized eyes and different reactions to the same stimulus. The signal will also contain noise. Some of this is created by the pupil measuring technique and other by difficulty in ensuring that each participant performs only the cognitive processing that they are being ask to undertake. To control for these issues researchers have used a number of data analysis techniques which are described below.

Kahneman and Beatty (1966)' experiment sampled the pupil size only once a second which gives a small number of discrete size readings for each experiment. For each condition they took the peak value of the pupil dilation and compared them using a t-test. They also took averages of the pupil dilation across all the subjects and compared them visually on a graph. Jainta and Baccino (2010) normalised all pupil sizes by subtracting the size of the pupil during an initial no-stimulus calibration state. They measured pupil size 25 times a second and calculated separate t-tests for each sampling

point to compare the pupil dilation for each task. By normalising pupil sizes, they account for individual differences in pupil size and resting state. Cavanagh et al. (2014) also normalised pupil sizes from a pre-stimulus baseline. They used a repeated-measures analysis of variance (rANOVA) to compare pupil sizes during different areas of interest.

Marshall was concerned about using pupil dilation in environments with varying light levels so came up with the *Index of Cognitive Activity* (ICA) (Marshall, 2007, Marshall, 2002) This uses wavelet theory to create an index based on small changes in the pupil size rather than absolute values. Previous studies mainly worked with stimuli which were presented at a particular time and caused participants to begin processing that stimulus immediately. By looking for changes in pupil dilation Marshall's approach suggests that it may be possible to detect the onset of processing without knowing when the stimulus occurred. The current formulation of the ICA is complex and not clearly published but it may be possible to use a simpler method inspired by this approach. de Gee et al. (2014) filtered the input data to remove both very high and very low frequencies. They used a *third order Butterworth* filter with a passband of 0.05–4 Hz. As previously mentioned, pupils dilate with varying speeds and orders of magnitude depending on the stimulus. By using filters, it may be possible to identify the frequencies produced by different stimuli and remove those produced by game elements such as changing graphics and motor actions. This would leave only the cognitive load produced by the game experience itself.

These studies suggest that pupil dilation could be an effective way of measuring the cognitive load used whilst playing games. These changes in cognitive load could be an effective measure of the experience of playing the game. There are a variety of factors in the experience of playing games which may add additional noise to the pupil dilation data. These include light levels, motor actions and differences between individual players. Previous research has managed to minimalize these issues and use pupil dilation as a measure of cognitive load in other situations and it is likely that I could use similar techniques to apply this method to games.

2.7. Measuring how well games hold our attention

As has been noted previously, a key aspect of the experience of playing games is how engaging the game is to play. Most games aim to be engaging, so a measure of engagement is, to a certain extent, a measure of how successful the game is in meeting its goals. To be engaged in playing a game means that the game is fully holding your attention and preventing you from being distracted by events around you or thoughts about unrelated activities. This means that attention in games in seen differently from more traditional usability focused studies (e.g. Holland and Morse (2001), Weinberg et al. (2013)) which often aim to create interfaces which are usable without demanding users' full attention, as that attention may be needed for other tasks such as driving a

car. This feature that engaging games hold your full attention is acknowledged by other game experience measures. For example, the IEQ (Jennett et al., 2008) asks "To what extent did you notice events taking place around you?" and the less that players notice events around them, the more engaging the game is. This is one area in which self-paced games are likely to be similar to fast-paced action games, as both types of games hold players' attention. If we could find a way of measuring how well participants' attention is held then this could serve as a measure of engagement which is a key part of the experience of playing self-paced games. To consider how this ability to hold attention could be measured I first considered psychological approaches to attention, then looked at alternative approaches to measuring attention both in non-game applications and also in games.

Psychologists have taken many approaches to looking at visual attention. It is generally agreed that attention is not one process but a feature of several different processes, each of which can be "paying attention" or not to a stimulus (Rensink, 2015). One of these processes is "visual search" which is what happens when we see an image for the first time and are scanning to find a particular item which may or may not be there. This has been extensively studied and the way that particular stimuli can "pop out" and grab attention is well understood (Wolfe, 2014, Duncan and Humphreys, 1989). To investigate how well games hold attention we need to look at processes which hold attention to a particular stimulus or allow participants to become distracted. The load theory of attention (de Fockert et al., 2001, Lavie et al., 2004, Lavie, 2005) states that participants are more likely to become distracted under high cognitive load but less likely to be distracted if under high perceptual load. Games typically ask players to process complex visual information which puts them under perceptual load. Load theory would explain that the reason that games hold players' attention is that the high perceptual load makes the players less likely to be distracted by surrounding events. Lleras et al. (2017) replicated Lavie's load theory experiments but also found that if they changed the instructions slightly then participants under high cognitive load were *less* likely to be distracted which is the opposite effect, so it is unclear how robust these findings are. Also, like the visual search studies described above, load theory studies have only been done on tasks which take a fraction of a second rather than the prolonged focused attention which happens during game play.

Other researchers have investigated attention over a longer period of time by looking at the phenomenon of "mind wandering" in which people lose track of the task they are supposed to be doing. Smallwood et al. (2008) found that participants whose minds wandered during a task were less able to construct a situational model of what had been happening. Levinson et al. (2012) found that participants with high working memory were more likely to mind wander, suggesting that excess working memory is required for mind wandering. Unsworth and Robison (2018) found that mind wandering can happen regardless of differences in arousal levels and that there are different types of mind wandering which happen at different arousal levels. None of these experiments looked at mind wandering during games and it is possible that mind wandering does happen whilst participants are playing a game. However, this research

suggests that it is a complex phenomenon which is not yet well understood so using it as a measure of game experience may be difficult.

Another psychological attentional process which takes place over longer time periods is known as "inattentional blindness". This refers to situations in which participants do not pay attention to a stimulus even though it is right in front of their eyes, in these situations they can be effectively "blind" to the stimulus. The most famous demonstration of inattentional blindness was done by Simons and Chabris (1999). Participants were told to watch players pass a basketball between themselves and to count the number of passes. Unknown to the participants another experimenter dressed as a gorilla passed through the basketball players and was not seen by a proportion of participants because they were focused on counting the number of passes.
Drew et al. (2013) repeated the gorilla experiment but with radiologists who were performing a familiar lung nodule detection task. The final trial included an image of a gorilla which was 48 times larger than the average nodule. 83% of the radiologists failed to spot the gorilla despite eye tracking showing they were looking directly at it. In Simon and Chabris's experiment 46% of participants failed to see the gorilla which is considerably less than the 83% found by Drew at al. This suggests that participants who are experienced in a particular field or situation may have "learnt" not to notice unexpected stimuli which creates a higher level of inattentional blindness than that experienced by participants experiencing a novel situation such in the Simons and Chabris's experiment. It is likely that inattentional blindness occurs during digital games as players find it difficult to focus their attention on more than one aspect of the game at the same time. Players become experts in these games by learning how to succeed at the game, this may create a high level of inattentional blindness.

Other researchers have looked in more detail into which factors increase levels of inattentional blindness. Most et al. (2001) found that levels of blindness depend on both how similar the unexpected object is to other objects in the display and also the attentional set of the observer. Cartwright-Finch and Lavie (2007) found that inattentional blindness was increased if participants were under high perceptual load during their main task. Once again, these situations occur extensively in games. Typically, games feature many similar objects, players adjust their attentional set depending on the genre of game they are playing and players are under very high perceptual load as they visually scan all the game elements and consider how to manipulate them. The presence of all of these factors in digital games suggest that it is highly likely that inattentional blindness happens during gameplay and the more immersed players are in the game the stronger the effect.

Rees et al. (1999) used fMRI brain scanning to investigate whether participants in an inattentional blindness experiment were really "not seeing" the unexpected stimulus or whether they were just not remembering that they had seen it. They found that when attention was diverted elsewhere the brain did not differentiate between real words and random letters and concluded that participants really did "not see" the unattended stimulus. Zendle et al. (2018) looked to see if people who played digital games were

primed to recognise the subject of the game more quickly after they had finished playing the game. To do this they asked participants to play one of two games which were themed around either around the topic of mice or the topic of trucks. They were then tested how fast they could classify images relating to the game theme. They found that participants were *negatively* primed towards the theme of the game. So, for example, those participants who played the game about mice reacted slower in the classification task to pictures of mice. Both games had the same underlying gameplay and the theme was irrelevant to actually playing the game. So, it looks like participants actively ignore the irrelevant aspects of the game to concentrate on the game elements that they need to succeed at the game. This ignoring continues after the game has finished which slows down their reactions to the classification task. Zendle et al found that this effect happens after only 20 seconds of gameplay. This suggests that if players are immersed in a game then their attention is specifically removed from aspects of the game which are not needed to succeed.

2.7.1. Measuring attention

There has been some work to measure attention in non-game contexts. Kinoshita (1995) showed participants text and numbers and found that they only remembered a stimulus if they had been paying attention to it. Smallwood et al. (2008) asked participants to read a Sherlock Holmes story and interrupted them periodically to ask if they were "on task". They found that participants who reported less time on task were less likely to be able identify the villain of the story.

Some studies have already looked at measuring how well games hold players' attention. Nordin et al. (2013) manipulated attention in games but it did not change levels of reported immersion so probably is not a good foundation for creating a new measure of engagement. As described earlier in this chapter (2.4 Physiological measures of game experience) Burns and Fairclough (2015) found that more immersed players had smaller ERPs in response to "oddball" tones. Although they do not explicitly make the connection to attention this may be because more immersed players had their attention held more strongly by the game so responded less strongly to the distracting tones. Brockmyer et al. (2009) validated the Game Engagement Questionnaire by interrupting players with a spoken voice asking them if they had dropped their keys. Players who were more engaged in the game were less likely to respond to the voice. Jennett et al. (2009) played ten different audio sounds to players as they played an action game. Players were then asked which sounds they remembered after the game. Those players who were more immersed in the game (as tested by an IEQ) remembered fewer sounds.

None of these studies involved self-paced games but the successful ones all measure game attention by offering a distraction to players while they play the game and then measuring how distracted they were from the game. The more immersed players are in the game, the less likely they are to be distracted. All of these studies used audio distractors. However, it may also be possible to use visual distractors rather than audio distractors. Standing (1973) found that participants were extremely good at recognising

pictures that they had been shown previously. He showed participants 1000 different images and found recognition rates of over 90% in a subsequent test. Brady et al. (2008) repeated this study and found a similarly high level of image recognition. This suggests that participants would have a high rate of recognition for visual images and that they would make good distractors for a measure of attention. Jennett and Brockmyer found that measuring attention using distractors was an effective measure of engagement in action games and it looks likely that a similar measure may also work for self-paced games.

2.8. Chapter conclusion

I have considered the question of what triggers and sustains players' engagement in self-paced games from a number of approaches. Considering the terms "play" and "fun" suggests some basic classifications of the types of activities that players perform when playing games but does not help explain why some games are more engaging than others. Games designers and other non-empirical researchers have come up with a long list of factors which may contribute to engagement in self-spaced games. However, without further empirical confirmation it is difficult to judge which of these are valid. Various researchers have developed different empirical models of game experience. Most of these were developed using other game genres such as first-person shooters but they could still form a basis for empirical investigation into players' engagement into self-paced games. All of these empirical models rely on "after the fact" questionnaires which cannot capture the experience of individual game events and suffer from being self-reported measures.

Physiological measures such as skin conductance, heart rate and EEG can give a measure of the overall game experience but also capture players' reaction to individual game events. They also have the advantage that they are not self-reported. However, many of these measures do not have an unambiguous link to psychological states and interpreting them can be difficult and is often specific to each particular type of game. The most well understood measures have been used to measure emotional "arousal" during action games but this is unlikely to be an important part of the experience of playing self-paced games.

Pupil dilation has been widely studied and shown to be a robust measure of cognitive load in a number of situations. It is subject to noise and confounding factors, many of which are present during self-paced gameplay. However, researchers have pointed to techniques which can be used to ameliorate these issues. The 2-300ms delay between stimulus and response may be problem for action games but suits the slower pace of self-paced games so that pupil dilation looks like a promising technique for understanding player's experience of playing self-paced games. This gives rise to first research question that I will investigate in this thesis.

Can changes in pupil dilation be used to measure the experience of playing self-paced games?

The experiments which investigated this research question are described in chapters 4 and 5.

Self-paced games are interesting because of the way that they hold players' attention without requiring immediate action. There have been successful attempts to measure game attention in action games by attempting to distract players while they play the game and then measuring how distracted they were from the game. The more immersed players are in the game the less likely they are to be distracted. This technique may be suitable as a new measure of the experience of playing self-paced games, it is not a self-reported measure and may be able to provide a continuous measure of experience over the time of the game. This gives rise to the second research question:

Can the experience of playing self-paced games be measured by seeing how likely players are to notice irrelevant stimuli while playing the game?

The experiments which investigated this research question are described in chapters 6 and 7.

3. Experimental setup

This thesis describes two different approaches to developing a new measure of the experience of playing self-paced games. One approach investigates using pupil dilation to measure differences in cognitive load during the game and the other looks at measuring how well the game holds players' attention. Both approaches consist of an iterative sequence of experiments which develop and test the measure which is being investigated. To do this both approaches have a number of elements in common. They both use the same self-paced game for the experiments. They both use the same eye tracking equipment and they use the same game experience questionnaire to provide a benchmark measure of game experience to compare with. Rather than provide duplicate rationale and details of these elements in the chapters which describe the experiments, it makes more sense to describe these aspects of the experimental setup in one place. Therefore, this chapter describes elements of the experimental setup which are common to both approaches. They are the game that participants played and its variants, the eye tracking equipment and the game experience questionnaire used to benchmark the measures which were being developed.

3.1. Two Dots

To investigate the experience of playing self-paced games I needed an example of a self-paced game to use in my experiments. Although there are a wide variety of self-paced games from strategy games like *Civilization* to casual games such as *Candy Crush Saga* there are certain attributes which make some more suitable than others for running experiments. I wanted to be able to recruit participants who had never played the game before so the game needed to be one that people could learn to play quickly. It is much easier to recruit participants for short experiments than ones that take a long time and taking less time to perform experiments would allow me to do more experiments with more participants so the game needed to be playable in a short time. To maximise the number of potential participants it also needed to have an accessible subject matter without graphic violence. This list of requirements is very similar to Kultima (2009)'s conception of *casual design values.* Kultima describes those values as *Acceptability*, *Accessibility*, *Simplicity* and *Flexibility.* So, it seemed natural that I would use a casual game for my experiments. Another important criterion was that the game would be fun and engaging. I considered designing my own casual game. However, I could not guarantee that it would be as engaging as existing commercial games so I decided to make a copy (known as a clone) of a successful casual self-paced game. By making a

copy of the game I would have complete control of the presentation and environment of the game and would be able to modify it to suit my experiments. The game I chose was *Two Dots* (http://weplaydots.com/twodots/). *Two Dots* is a very successful (Crook, 2014) self-paced game with a simple graphic style and the minimum of additional rewards and bonuses. It is a mobile game available for Apple and Android devices. The simplicity meant that it was easier to clone and also easy for participants to learn how to play.

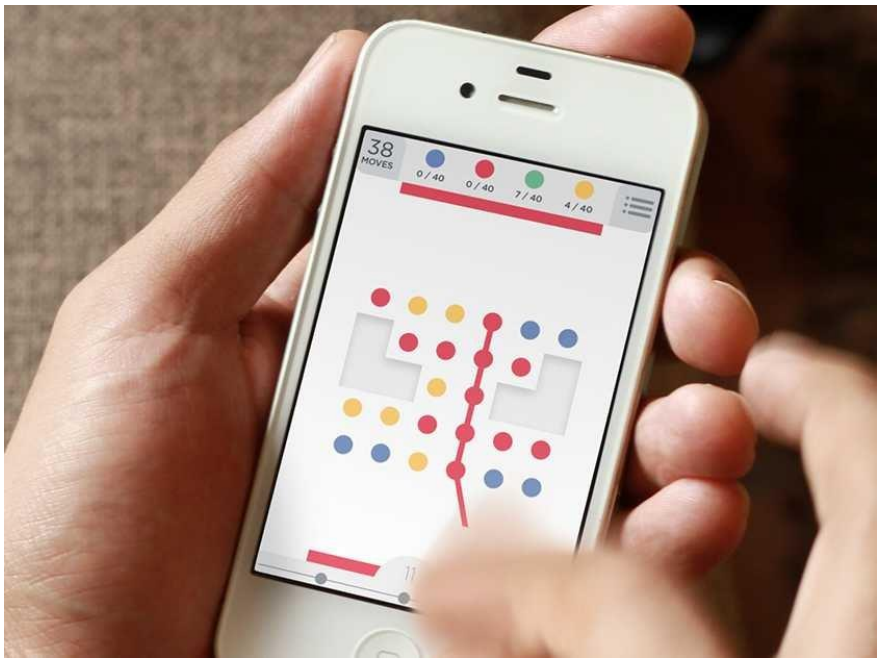

Figure 3 Two Dots being played on a phone. Players join the dots in the grid to meet the targets at the top of the screen within the move limit

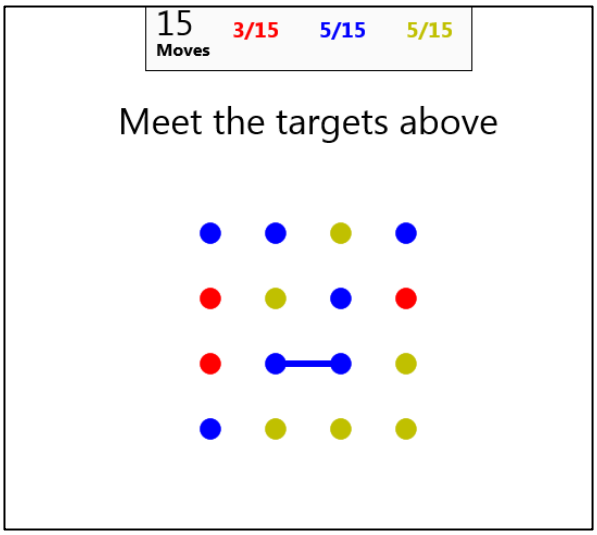

Figure 4 My clone of Two Dots which is played using a mouse on a computer

3.1.1. How to play *Two Dots*

The core gameplay of *Two Dots* is very simple. As a mobile game the original game was played on a touchscreen but my clone was played using a mouse. The game has a grid of dots of different colours. The player has to find two or more dots of the same colour which are next to each to other. The player then drags a line which joins the dots. When they release the line the dots that they have joined disappear. Then the gaps are filled by the remaining dots dropping down. Any gaps still remaining are filled by new dots dropping from the top of the screen. Each level has a set of targets at the top of the screen which indicate how many dots of each colour need to be joined. For example, the screen in Figure 3 shows that the player needs to join 40 blue dots as well as 40 red, green and yellow dots. To succeed at the level the player needs to join this number of dots within a certain number of moves. This move limit is shown in the top left of the screen. If they join enough dots within the move limit, they successfully complete the level and move onto the next one. If they fail to remove that number of dots, then they fail the level and have to play it again from the start.

3.1.2. Variants of the game

To perform experiments on measuring game experience I created several different versions of the game of *Two Dots.* These are described below:

Monochrome version

I created monochrome versions of the games to avoid strong changes in light levels which would change the dilation of participants' pupils. Several of the experiments measured the size of participants' pupils while they were playing the game. Pupil dilation is affected by changes in the amount of light reaching the eye. If an experimental stimulus changes in brightness this could potentially introduce noise into the pupil response data. Some pupil dilation experiments (Just and Carpenter, 1993, Cavanagh et al., 2014) use "luminance balanced" stimuli. In these experiments all stimuli are of a constant luminance throughout the experiment so the amount of light reaching participants' eyes is constant throughout the experiment. This is to avoid changes in light levels affecting the pupil size and confounding the results. In *Two Dots* the number of dots displayed on the screen can change and the game also displays popup boxes with information which means that it is not possible to completely luminance balance the screen display from one move to the next. However, I did want to avoid any unnecessary large changes in luminance during the stimulus, so I changed the varying colours of the original *Two Dots* game to be different symbol patterns which are shown below.

Figure 5 The different symbols used in the *Two Dots* game stimulus

The symbols are designed to be very different shapes to make it easy to perform visual search across them (Wolfe, 2014). All of them are luminance balanced with each other so they have an equal average level of brightness. This is done by making sure that each symbol has the same proportion of dark and light pixels. This means that any given grid of symbols will have the same luminance as another grid of the same size. However, luminance may change if the grid changes size or the game has to display a popup box with information.

The commercial version of *Two Dots* does actually have an option for colour blind players which replaces the different coloured dots with symbols and pilot testing showed that players found the game experience symbol version of the game to be the same as the coloured version

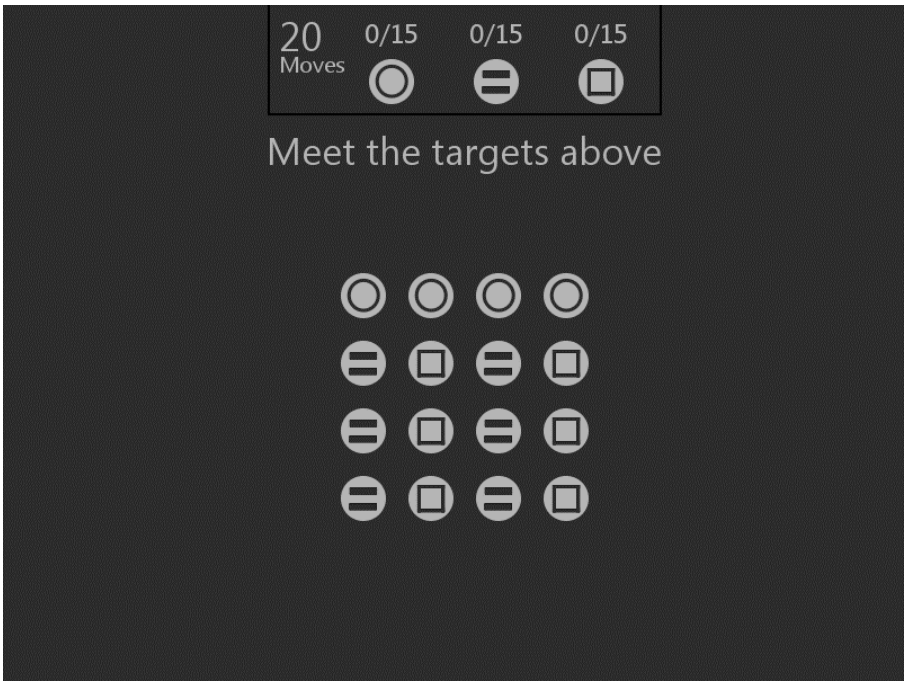

Figure 6 The monochrome version of the *Full game* of *Two Dots*

Gameplay variants

I also created variants of *Two Dots* which were designed to give players different play experiences. Many of my experiments were concerned with measuring the experience of playing self-paced games. To do this I needed to manipulate the experience that participants had, but I needed to be sure that their difference in experience was due to differences in how they found the game rather than other factors. To do this I created three variants on the game of *Two Dots* which were similar in that players still

performed the same activity of joining dots but the gameplay was changed so players had a different experience to that of playing the standard version of the game. Each of the game variants was based on the standard version of *Two Dots* but had some game play element removed. Because of this, I refer to the standard version of *Two Dots* as the *Full game* to contrast it with other versions which have elements removed. There are two main components to *Two Dots*, the goals and the dots. I created two different versions of the game by simplifying each of these components in turn. I also created an additional version of the game by simplifying both of these elements in the same game.

No goals game

This version simplifies the goals component of the *Full* game. It is the same as the standard version of *Two Dots* except that there are no targets for the number of dots to connect and no limit on the number of moves. Because there are no targets to meet in joining dots this means that there are also no levels in this game. So, players are given the instruction "Now play how you like" and just join dots which disappear and add to the dot totals until the time runs out.

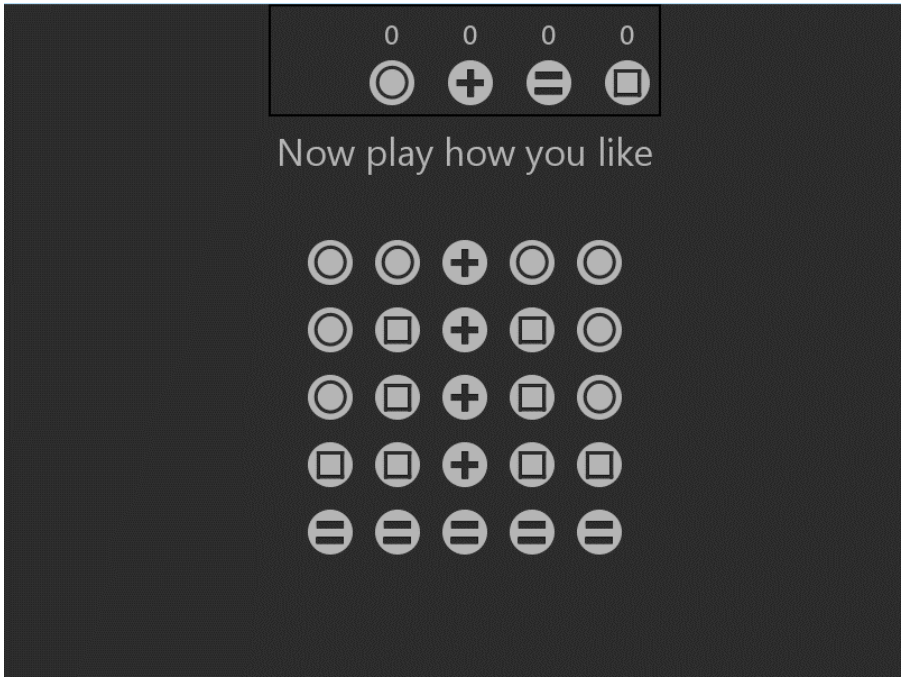

Figure 7 The *No goals game*. Players can still join the dots in the centre of the screen. The game keeps track of how many of each type have been joined but there are no targets and no move limit.

All dots the same game

This version simplifies the dots component of the *Full game.* It is the same as the standard version except that all the dots are the same colour or symbol. Because they are all the same, players have no difficulty in finding two or more dots to join together. However, this version of the game does have targets for the number of dots which need to be joined and once these are met the player does go onto the next level.

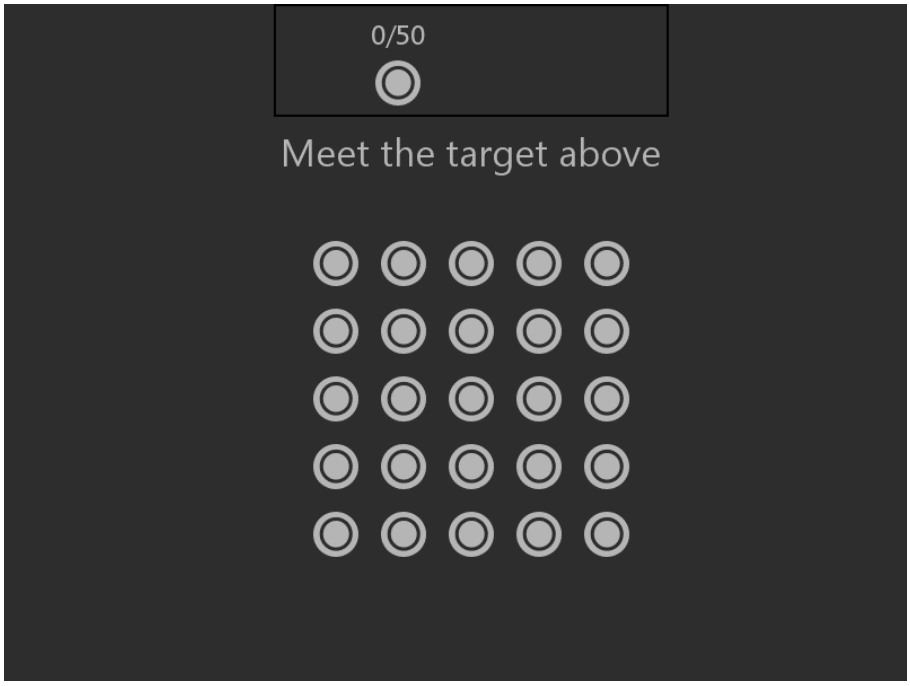

Figure 8 The *All dots the same* game has all the dots the same symbol. Players can still join the dots in the centre of the screen. Players have to reach the target at the top of the screen to get to the next level.

Bad game

For this version of the game I simplified both the dots and the goals of the game. This produces a version of the game in which all the dots are the same colour and the targets are also removed so there are no goals to the game. All players can do is join the dots of the same colour which then drop down and are replaced by other dots of the same colour. As there are no targets or goals the game just continues until the time runs out. This is likely to be the least engaging game variant so I called it the *Bad game.*

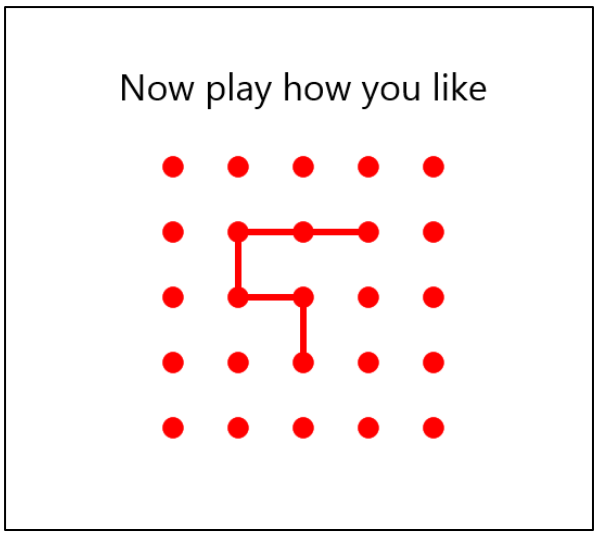

Figure 9. A screen from the *Bad game*. Players can join dots to remove them, but there is no challenge or target.

3.1.3. In-game distractors

The experiments in chapter 7 rely on inserting irrelevant distractor images into the game. The distractor images had to be part of the game so that players were looking directly at them whilst playing the game. The images used are icons from the *Webdings* typeface (the reason for this is described in more detail in section 6.1) and they change every 5 seconds. To insert the distractor images into the game of *Two Dots* I increased the size of the dots and put the distractor images inside the dots. The distractor images change every 5 seconds. Players have to join dots of the same colour and the actual distractor images are irrelevant to the gameplay. Examples of the game with in-game distractors are shown in Figure 10 and Figure 11.

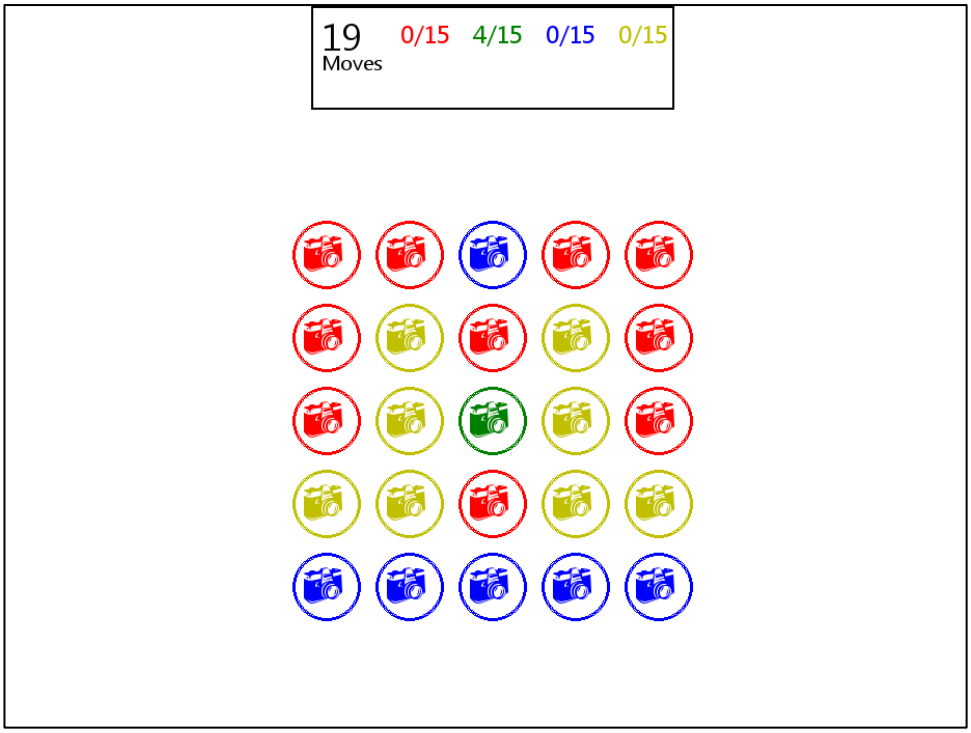

Figure 10 A screen from the *Full game* with in-game distractors. Players have to join dots of the same colour. The images inside the dots change every 5 seconds.

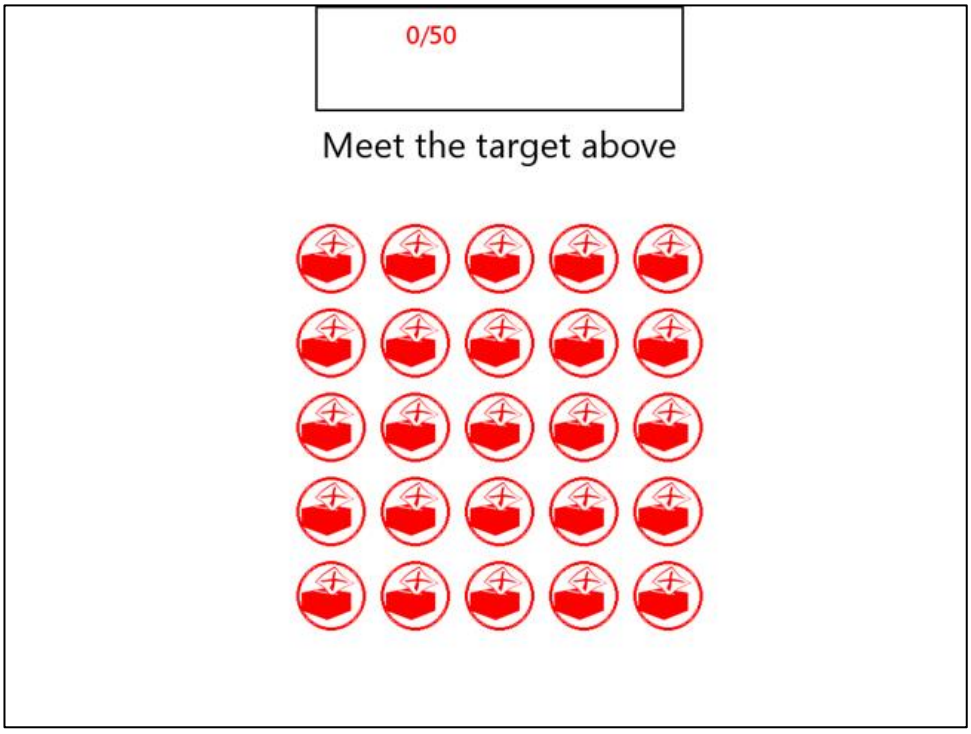

Figure 11. A screen from the *All dots the same* game with in-game distractors.

3.2. Eye tracking equipment

Many of the experiments described make use of eye tracking equipment. The same eye tracker and configuration was used for all the experiments so it makes sense to describe it here. The eye tracking equipment used was an Eyelink 1000 Plus made by SR Research (https://www.sr-research.com/). This uses a desk mounted camera which tracks eye movements on a desktop screen. Participants place their chin on a chin rest which keeps their eyes a constant distance from the screen. This is shown in Figure 12.

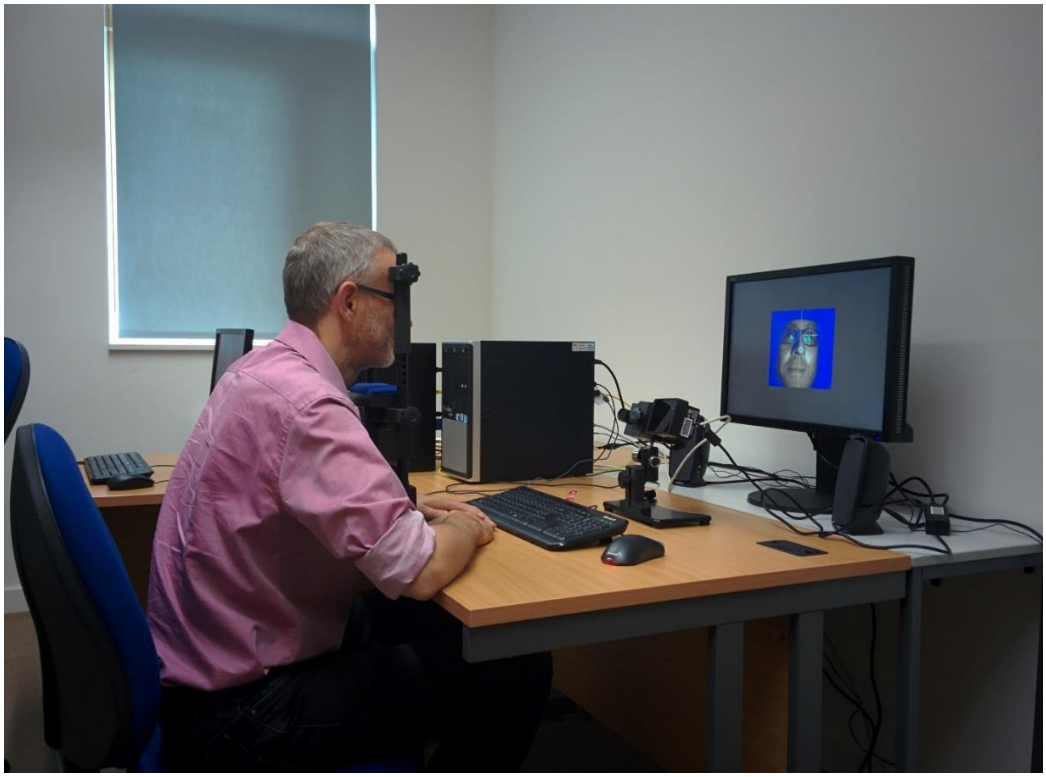

Figure 12 Eyelink eye tracker being used by a participant

The eye tracker works by projecting infrared light on the participant's face. This is then reflected from their face and eyes and picked up by the camera. Image processing in the eye tracker then processes the camera signal and converts this into data on eye saccades, fixations and pupil size. The tracker has a sample rate of up to 2000Hz which may be needed for tracking fast moving saccades. For my experiments I was only interested in pupil dilation and fixations so I operated the tracker at 250Hz.

Before eye tracking could begin all participants went through a registration procedure which is described below. Participants sat in the chair and put their face in the chin rest looking towards the screen. The chin rest was kept in a constant position for all participants to ensure that they all had the same view of the screen and stimulus. However, participants could adjust the height of the chair so that they were comfortable. After the participant had adjusted the chair, I then adjusted the image processing threshold so that the infra-red camera had a clear image of their pupils. Eye

pupils do not reflect infra-red light well so they show up strongly against the rest of the face which does reflect the infra-red light. Heavy eye makeup absorbs infra-red and glasses can distort the light so participants with either of these can make this adjustment more difficult but I managed to get a clear image for most participants. The few cases where this was not possible are mentioned in the participants section of the relevant experiment. The next step was to calibrate the tracking system using a 9-point calibration procedure. In this procedure participants look at 9 different points on the screen in turn and press the space bar when they have focused on each one. This is then verified by repeating the calibration procedure and checking that the tracking for both calibrations are within 1 degree of each other.

3.3. Game experience questionnaire

The aim of this research is to come up with new measures of the experience of playing self-paced games. To do this I needed some other pre-existing experience measure to compare my new measure to and serve as a benchmark to assess its effectiveness.

For some of the experiments in chapters 5, 6 and 7 participants play different games and have a different game experience depending on the game. To confirm that their experience was indeed different I measured their experience using a standard measure of game experience. I also wanted to compare my new measure with this existing measure to see how they relate to each other. As discussed in the literature review (see chapter 2) there are many different existing measures of game experience. Most of these measures can be divided into those which are based on questionnaires and those which are based on measuring physiological properties of the player's body. Physiological measures have some advantages but it is often difficult to unambiguously relate the property being measured to a particular aspect game experience. Also, many of the measures may be specific to the experience of playing action games and thus not suitable for the experience of playing self-paced games. After considering these issues I decided that a physiological measure was not suitable as a benchmark measure of game experience.

As the literature reviews notes, many different game experience questionnaires have been developed. I did not find a measure which had been developed using self-paced games so I looked for a measure which was based on broad game experience concepts rather than those which only apply to a particular genre of games. This type of measure was more likely to be wide ranging enough to also measure the experience of playing self-paced games even though it had not been developed with them in mind. I was also looking for a measure which had been well validated and widely used in the literature. If a measure is widely used then this suggests that it may be robust and produces useful results in a variety of situations. Based on these criteria I considered the GEQ (Brockmyer et al., 2009), PENS (Ryan et al., 2006) and the IEQ (Jennett et al., 2008). The development, validation and characteristics of these measures are discussed on pages 39-41. Work by Denisova et al. (2016) suggests that all three of these questionnaires produce similar results so it is likely that any one of them could have been used.

I narrowed down the choice by considering the type of measure I was looking to develop. This thesis considers two different possible approaches to measuring game experience; the first is based on measuring cognitive load and the second is based on measuring how well games hold player's attention. Cognitive load has been linked to attention (Lavie et al., 2004, Lavie, 2005) with the finding that increased use of cognitive load was associated with participants being more likely to be distracted. This implies that both of the approaches I consider are related to the way that games hold players' attention. Jennett (2010) found evidence that immersion is a form of selective attention. which led me to choose the IEQ over the other measurements.

Before using the IEQ in the main experiments I wanted to confirm that the IEQ was an effective measure of the experience of playing self-paced games. To do this I performed a pilot study with 16 participants, 8 played the *Full game* of *Two Dots* and another 8 played the bad game (described earlier in this chapter). As the bad game is specifically designed to be less engaging, I expected that the levels of immersion would be lower in that game than the *Full game.* After playing the game each participant filled in an IEQ about their game experience There was significant difference between the two games ($p<0.001$) with a very large effect size ($\eta_p^2=0.390$). This confirmed that the IEQ was an effective measure of game experience for self-paced games and so I decided to use it throughout the thesis.

4. Measuring cognitive load using pupil dilation

In the literature review chapter, I considered existing methods of measuring the experience of playing digital games. There are two main approaches; questionnaires and physiological based measures. There are many different game experience questionnaires which have had some degree of success in measuring the experience of playing action-oriented games. However, they all suffer from the same two issues. They are self-reported measures which means they are subject to error due to players not reflecting accurately on their whole experience. Also, because players answer the questionnaire after the game has finished, questionnaires are unable to measure changes in the game experience over time. Physiological measures aim to solve the problems with questionnaire-based methods. They are based on monitoring properties of the player's body such as heart rate and skin conductance. When playing a game these properties may change depending on the experience the player is having. Existing physiological measures are mainly designed to measure the changes in arousal which occur during fast-paced action games. Self-paced games are much slower and are less likely to lead to the same changes in arousal and so may need to use different physiological measures. I therefore decided to develop a new physiological measure of the experience of playing self-paced games.

To further summarise discussion from the literature review, when developing this measure, I considered the particular experience of playing self-paced games. I then looked for a physiological measure that reflected that experience. Self-paced games tend to be either puzzle games such as *Two Dots* (Playdots, 2014) or strategy games like *X-COM* (Firaxis, 2012). In both these game genres mental effort and reasoning are key to completing the game's challenges and making progress. It seems likely that this this mental effort is similar to the psychological concept known as "cognitive load" (Sweller, 1988). Cognitive load approximates to the amount of "mental manipulation" or "abstract thinking" that you are doing at a particular time.

Cognitive load has been measured by detailed measurements of changes to participants' eyes. When participants are under high cognitive load the pupils of their eyes become larger. This is known as "pupil dilation" and has been used by psychologists (Kahneman and Beatty, 1966, Jainta and Baccino, 2010) to measure cognitive load. Their experiments showed that pupil dilation is a reliable measure of

cognitive load which creates a strong effect size between participants who are under different levels of cognitive effort. This suggests that pupil dilation has the potential to be an effective online measure of how much cognitive load players are using when they play self-paced games. As mental effort is a key part of self-paced games, being able to measure cognitive load seems like it will be a useful measure of the experience of playing the game which also links to an established psychological concept. This chapter describes two initial experiments designed to build on the work of existing pupil dilation studies and apply them to measuring cognitive load in self-paced games.

The first experiment in the chapter replicates a study by Kahneman and Beatty (1966) which measures cognitive load using pupil dilation. This experiment used an audio-based stimulus and participants spoke their response. The next experiment extends this to use a visual stimulus and mouse-based response which is a form that is more useful for measuring pupil dilation in self-paced games. The design of these experiments used several techniques to prevent other factors adding noise to the pupil dilation readings. These techniques are described in section 4.1 below. The lessons learnt in experimental design are then used in the next chapter to investigate using pupil dilation to measuring the experience of playing self-paced games.

4.1. Experimental design and pilot studies

To design the experiments in this chapter I first considered possible confounding factors which might interfere with the measurement of cognitive load using pupil dilation. I also performed pilot studies which helped inform the experimental design.

4.1.1. Confounding factors

There are a number of factors which can confound pupil dilation measurements (Beatty and Lucero-Wagoner, 2000) including changes in light levels as well as participant arousal and stress. The most difficult to control are changes in light levels. If participants are viewing stimuli on screen, then most changes in graphics will change the amount of light reaching their eyes. Light levels on their pupils can also be changed by participants moving their head or changing what they are looking at. Beatty (1982) also found that making a motor action caused a pupil response, although it was not completely clear from his experiment whether it was making the action which caused the response or just deciding to make the action. Digital games typically feature changing graphics, numerous motor actions to control the game and often changes in arousal and stress. Because of this, experiments which attempt to measure cognitive load using pupil dilation must be carefully designed to minimise the additional noise from these additional factors which may affect the pupil dilation measurement.

4.1.2. Experimental design

Because of these confounding factors I decided to use an experimental paradigm based on comparing the results of two similar conditions. In each condition participants are exposed to the same stimulus and have to give a similar type of response. The only difference between the conditions is the task that participants are asked to do. I then compare the results of the two conditions and any significant difference in pupil dilation should be due to the task. Because the additional noise from other factors should be the same for each condition any additional pupil dilation from other factors will be cancelled out when the two conditions are compared.

The pupil dilation experiments in this chapter are based on a set of pupil dilation experiments performed by Kahneman and Beatty (1966). These experiments showed a strong effect using a small number of participants. This made them a good starting point for my experiments as it suggested that similar experiments could find significant results without large sample sizes. The first experiment in this chapter attempts to replicate Kahneman and Beatty's original experiment. They used an audio stimulus and a range of tasks of increasing difficulty. My replication keeps the audio task but only uses one task of moderate difficulty. Almost all digital games involve looking at a screen and most self-paced games are controlled with a mouse or a touchscreen. The second experiment extends the experimental setup to be more similar to self-paced games. So, participants perform the same task as the first experiment but see the stimulus on a screen and click with a mouse to give their response.

The purpose of these experiments is to develop and explore techniques for measuring pupil dilation due to cognitive load during self-paced games. For each experiment I performed a significance test to see if the main hypothesis was supported, but I also performed other analyses to investigate other aspects of the experiment which could inform future experimental design. These included training and fatigue effects, blink rates and whether participants had focused their gaze on the central stimulus.

4.1.3. SMI Eye tracker pilot studies

Initially I performed two experiments using a glasses-based eye tracker made by the SMI company. This eye tracker consists of pair of special glasses which contain cameras and infrared light sources. They enable participants' gaze to be tracked anywhere they look in the surrounding area, without needing to constrain the participants' field of view. I encountered some problems using this eye tracker so decided to treat these two experiments as pilot studies for the main experiments which are described in this chapter. The SMI eye tracker experiments and associated problems are described below.

The first experiment I performed using this eye tracker attempted to replicate an experiment performed by Kahneman and Beatty (1966). Participants perform two number tasks, one of which is harder than the other. They hear the numbers spoken to them and give their response by speaking. Kahneman and Beatty (1966) found

significantly higher pupil dilation in the harder task than the easy task. I attempted to replicate their experiment using a within participants study with three conditions and 10 participants. Participants performed 10 trials of the easy task followed by 10 trials of the hard task and then another 10 trials of the easy task. This experiment compared the mean pupil dilation across the whole length of each trial. I found a significant difference in pupil dilation between the hard task and the second easy task but not the first easy task and the hard task. I concluded that there was a significant training effect which made the first easy task much more difficult for participants and so increased the pupil dilation that was measured. I later repeated this experiment with a superior eye tracker and experimental method which is described in Experiment Pupil 1: Audio stimulus 4.2 below.

Due to this training effect the second experiment using the SMI eye tracker began with 10 training trials and then interleaved the easy and hard trials. This experiment used a similar number task but showed participants a visual stimulus and they gave their answer by clicking a mouse. In this experiment I divided each trial into 1 second long windows and compared the mean pupil dilation for each window rather than taking the mean pupil dilation across the whole length of the trial. This experiment did not show a significant difference in pupil dilation between conditions but did show a moderate effect between conditions. (η_p^2=0.180). I later repeated this experiment with a superior eye tracker and experimental method which is described in Experiment Pupil 1: Audio stimulus 4.3 below.

There were a number of problems with these experiments. I discovered that this particular design of eye tracker does not allow for accurate readings of pupil dilation. As the eye tracker is contained in a pair of special glasses it has to fit snugly over the participants' face and align the cameras to their eyes precisely. I found that for some participants the shape of their face and nose did not work well with the eye tracker and the results were less precise. There were also a number of other problems with the experimental design and setup such as unreliable synchronisation between the eye tracker and the stimulus computer. I therefore decided to stop using this eye tracker and move to a more suitable Eyelink eye tracker system which is used for all of the other eye tracker experiments in this thesis. Due to these various problems the results from these experiments are unreliable and will not described be any further in this thesis. They should be considered pilot studies which did inform the design of subsequent experiments. In particular the results support the use of training periods, interleaved stimulus and analysing pupil dilation in 1 second windows across the time of the trial. The effect size found in the second experiment was also used in a power calculation to estimate the number of participants needed for similar experiments.

4.2. Experiment Pupil 1: Audio stimulus

This audio stimulus experiment aimed to use pupil dilation to measure the difference in cognitive load between two different tasks. The experimental design was influenced by the pilot studies described above and included two particular features to reduce the risk of confounds to the data. These features were an interleaved stimulus and considering the data in "windows" of 1 second. They are described in more detail below.

Interleaved stimulus

Pilot pupil dilation studies showed that there may be a strong training effect on pupil dilation as participants get used to doing the task. There may also be an effect on pupil dilation which is caused by taking a break and coming back to the task. To remove these confounding factors this experiment interleaves the trials from the two conditions without a break between them. This means that participants perform each condition with almost the same of experience in each one. The order in which participants start the conditions is counterbalanced so half of them start with the easy condition followed by the hard condition and the other half start with the hard condition followed by the easy condition.

Consider 1 second window

Pilot studies showed that pupil dilation varied considerably across the time of each trial. From a participant's point of view the easy and hard conditions are exactly the same up until they have to give an answer. So, for most of the trial you would also expect the pupil dilation to be similar. To get a more detailed picture of what is going on I decided to consider the pupil dilation in "windows" of 1 second duration. I defined the pupil dilation for each window as the mean pupil dilation of all the readings within that time period. As each trial lasted 10 seconds this meant that each trial produced at least 10 different pupil dilation readings for analysis. I piloted this experiment with two participants which indicated that the period of maximum difference between conditions was a 1 second window starting 8.5 seconds from the start of the trial. (Possible reasons for this being the period of maximum difference are considered in the discussion in section 4.2.5). Most of the time windows I considered were consecutive and measured in whole numbers of seconds from the start of the trial. However, the pilot seemed to show the window of maximum effect size started at 8.5 seconds so I used this for the main hypothesis in order to maximise the chance of seeing a significant difference between conditions.

4.2.2. Hypothesis

The hypothesis of the experiment is:
For a 1 second window starting at 8.5 seconds from the start of the task, the size of participants' pupils will increase when they are performing the hard task compared to the easy task.

Design

This was a within-subjects design with two conditions. The independent variable was the difference in cognitive load needed for each condition. The dependent variable is the amount of pupil dilation in each condition. Participants were asked to perform a different task for each condition, an easy task (low load) and a hard task (high load). For the easy task participants heard 4 numbers from 0 to 4. They had to remember these for 2 seconds and then repeat them in the same order as they had heard them. The hard task was similar except that participants had to add one to the number and then repeat the result. The stimulus for each task was identical apart from the instructions to tell the participant which task to do. This was to minimise the chance of introducing noise into the pupil dilation data through other differences between conditions such as the brightness of the stimulus. I also collected data on the baseline pupil dilation without any stimulus and number of times that participants blink during the study. This additional data was used to get a better idea of how participants' eyes are behaving during the experiment.

Participants

I performed a power calculation to decide how many participants to use. Starting from the effect size seen in the pilot experiment (η_p^2=0.180) I calculated that 12 students gave a good chance of a significant result. I took a conservative approach and aimed for 14 participants. 20 students and staff from the University of York took part in the study. However, 4 were rejected because they did not follow the instructions. In the hard condition, they performed the addition as soon as they heard the number rather than when giving the answer at the end. I discovered this by asking them about their approach in the post experiment debrief. Another 2 were rejected because they got more than 60% of the answers wrong in the hard task. This left 14 (5 female) participants who were considered for the trial. Ages ranged from 20 to 30 (mean = 25.5). 9 of the participants were not native speakers of English.

Materials

Participants were asked to carry out two different tasks which are developed from Kahneman and Beatty's initial study. Both tasks involved a similar audio stimulus:

1. Pause for 2 seconds. This is used to get an initial baseline reading of participants' pupil sizes.
2. Participants hear instructions which tell them which task they are doing
3. Pause for 3 seconds
4. Participants hear 4 spoken random numbers from 0 to 4 with one second between them.
5. Pause for 2 seconds.
6. Participants hear 4 clicks, each one second apart. After each click, they need to give their answer.
7. Pause for 7 seconds. This gave participants' pupils a chance to relax after performing the task.

For each task the set of numbers that they heard was randomly generated by the website *random.org.* The numbers were generated with the condition that all the numbers in each set of 4 was different. Each participant heard the same sets of numbers for each condition. The spoken numbers were pre-recorded and spoken by the same female voice for each number. During the task participants were asked to focus on a light grey cross (RGB 51,51,51) in the middle of a dark grey (RGB 85,85,85) 24″ screen. The screen size was 51.5cm wide by 32.5cm high and the central cross was 4cm cm high. Participants were positioned in a chin rest so that their eyes were 95cm from the screen so that the screen filled 31.5° their participants' field of view. The cross filled 4.8 degrees of their field of view.

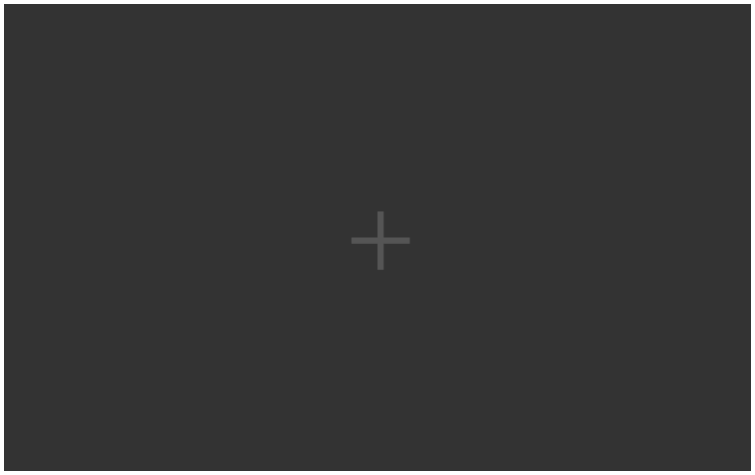

Figure 13 The constant screen displayed during the audio stimulus experiment.

The stimulus was created using the Eyelink Experiment builder software. The size of participants' pupils was measured using an SR Research Eyelink 1000+ eyetracker which recorded at 1000 frames a second. The study took place in a quiet room with no natural light and constant lighting levels.

4.2.3. Procedure

Each participant took part in both conditions. To control for possible training or fatigue effects the order of the conditions was counter-balanced so that half the participants performed the easy task first followed by the hard task. The other half of participants did it the other way around. Participants performed a consent procedure and were then calibrated with the eye tracker. They then had the first task explained to them and they performed 5 trials of this task. They then had the second task explained to them and performed 5 trials of that task. These ten trials made up the training phase of the study. For the main testing phase of the study participants alternated between doing 5 trials of the first task followed by 5 trials of the second task until they had done 30 main trials.

Before each trial, they heard audio instructions which told them which task they needed to perform.

4.2.4. Results

All of these results consider only trials number 11-40 and ignore the first ten trials. This is because those first ten trials were considered as training trials where participants are still learning the tasks. Where pupil dilation is reported, it is a normalised pupil dilation calculated by dividing the diameter of the participant's pupil by a baseline pupil diameter which is measured at the start of each trial. Unless otherwise stated the number of participants (N) is 14.

Hypothesis

The hypothesis proposed that there would be an increase in pupil dilation between the easy and hard conditions when measured in a one second window starting 8.5 seconds into the trial. The hypothesis was supported as there was a significant difference in the mean pupil dilation for each participant during this window; $F(1,13)= 121.619$, $p <0.001$ There was an extremely large effect size , $\eta_p^2 = 0.903$ between the two conditions.

Condition	Mean	SD
Easy	1.032	0.040
Hard	1.140	0.053

Table 1 Normalised pupil dilation for a one second window staring at 8.5 seconds

Pupil dilation over time

	Hard condition	Easy condition			
Time	Mean (SD)	Mean (SD)	Effect size η_p^2	p value	F(1,13)
0	1.013 (0.013)	1.009 (0.013)	0.056	0.395	0.774
1	1.042 (0.022)	1.018 (0.026)	0.471	0.005	11.560
2	1.064 (0.031)	1.022 (0.036)	0.591	0.001	18.819
3	1.075 (0.044)	1.031 (0.039)	0.511	0.003	13.588
4	1.100 (0.050)	1.032 (0.046)	0.671	<0.001	26.511
5	1.112 (0.056)	1.026 (0.044)	0.748	<0.001	38.688
6	1.142 (0.067)	1.060 (0.049)	0.750	<0.001	38.948
7	1.163 (0.062)	1.068 (0.056)	0.814	<0.001	56.945
8	1.153 (0.062)	1.039 (0.044)	0.875	<0.001	90.827
8.5	**1.140 (0.055)**	**1.033 (0.042)**	**0.903**	**<0.001**	**121.619**
9	1.123 (0.053)	1.030 (0.040)	0.846	<0.001	71.521
10	1.078 (0.055)	0.998 (0.036)	0.682	<0.001	27.819

Table 2 Normalised pupil dilation over the time of each trial

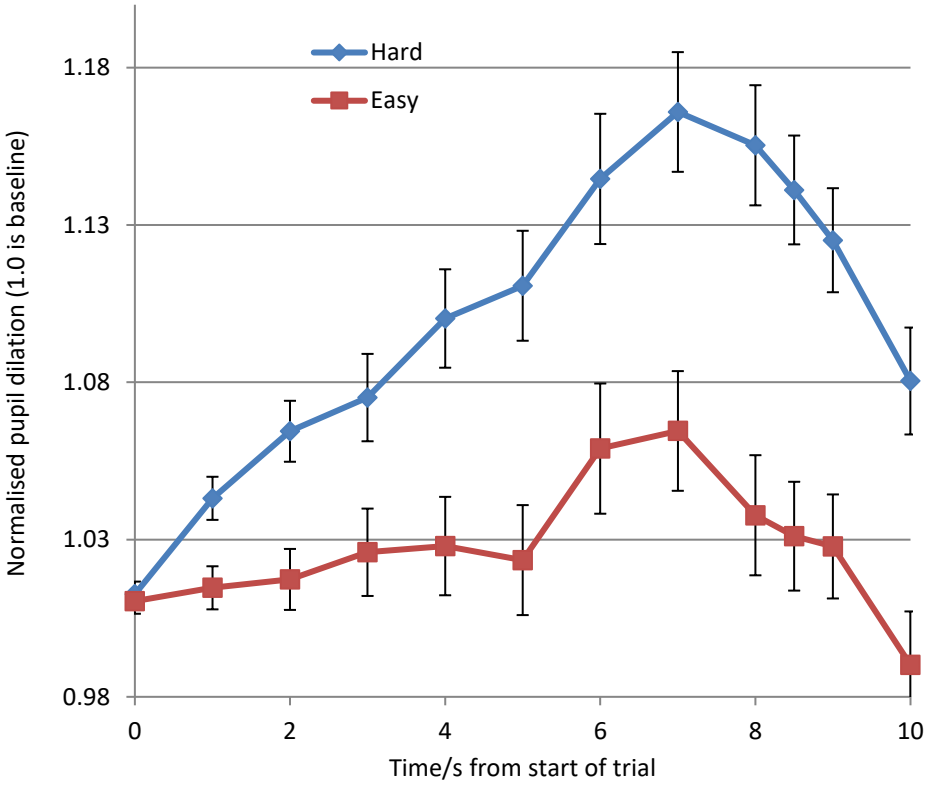

Figure 14 Normalised pupil dilation across the time of the trial (Error bars show standard error)

Blink rate

Condition	Mean	SD
Easy	1.26	0.44
Hard	1.38	0.49

Table 3 Blinks per second for each participant between the two conditions

I calculated the total number of times each participant blinked per second during all the trials for each condition. There was not a significant difference (p=0.086) between the rates of blinking for the easy and hard condition, although there is a large effect size between conditions; F(1,13)= 3.439, p=0.086, η_p^2=0.209.

Baseline

The baseline pupil dilation for each trial is the mean pupil dilation for the initial two second pause time when participants are doing nothing except looking at a fixation cross. Baseline pupil dilation is given in "arbitrary units" specific to the Eyelink eye tracker. There is a significant difference in the baseline pupil dilation between the easy and hard condition. There is an extremely large effect size between conditions. F(1,13) = 11.370, p=0.005, η_p^2=0.467.

Condition	Mean	SD
Easy	5429.56	645.15
Hard	5552.85	628.74

Table 4 Pupil size during the baseline period between two conditions

Training and Fatigue effects

To see if there are any training or fatigue effects, I calculated Pearson's correlations between the mean pupil dilation at 8.5 seconds and the trial number. There were no significant correlations between pupil dilation and trial number in either condition.

Condition	Pearons's r	Significance p	t value	df
Easy	-0.156	0.411	-0.835	28
Hard	0.0825	0.665	0.438	28

Table 5 Correlations between pupil dilation and trial number

Are participants focused on the central stimulus?

If participants are not focused on the central stimulus then this could potentially distort the pupil dilation measurement because their eyes may be at angle to the camera which would make them appear smaller. I investigated to see if participants were looking at the centre of the screen for the whole duration of the experiment. To investigate this, I created an area of interest (AOI) 640 pixels x 640 pixels across in the centre of the screen. This corresponds to 10.4° of participants' field of view. I then looked at what percentage of time participants looked within this AOI.

Participant ID	% time on central Area of Interest
50	96.24
51	80.08
52	87.12
53	99.92
54	98.98
56	85.04
59	91.48
60	99.96
62	99.65
63	**65.35**
66	99.33
67	83.98
68	**52.54**
69	99.27

Table 6 Percentage time each participant spent looking in the central area of interest

This showed that almost all the participants were looking inside the central area for at least 80% of the time. However, there were two participants who were outliers and

spent much less time looking inside the central area. Participants 63 and 68 only looked in the central area for 65.35% and 52.54% of the time respectively. I was concerned that this might be a confound on the results so I recalculated the previous results without these two participants. The results are very similar for all measures apart from the blink rate which slows a slight difference in effect size. These results are shown below.

Hypothesis testing after removing two participants

The hypothesis for this reanalysis was unchanged; that there would be an increase in pupil dilation between the easy and hard conditions when measured in a one second window starting 8.5 seconds into the trial. This hypothesis was once again supported as there was a significant difference in the mean pupil dilation for each participant during this window. There was also an extremely large effect size between the conditions; $F(1,11)= 96.966$, $p<0.05$, $\eta_p^2=0.898$.

Condition	Mean	SD
Easy condition	1.03	0.05
Hard condition	1.14	0.06

Table 7 Normalised pupil dilation for a one second window starting at 8.5

Pupil dilation over time after removing two participants

	Hard		Easy				
Time	Mean	SD	Mean	SD	Effect Size η_p^2	p value	F(1,11)
0	1.013	0.014	1.010	0.014	0.047	0.478	0.539
1	1.043	0.024	1.016	0.025	0.623	0.001	18.187
2	1.064	0.033	1.020	0.033	0.705	<0.001	26.303
3	1.073	0.048	1.027	0.035	0.600	0.002	16.502
4	1.101	0.054	1.026	0.044	0.752	<0.001	33.368
5	1.113	0.061	1.023	0.043	0.774	<0.001	37.753
6	1.144	0.072	1.059	0.050	0.788	<0.001	41.004
7	1.164	0.066	1.063	0.057	0.876	<0.001	77.504
8	1.155	0.066	1.038	0.047	0.884	<0.001	83.562
8.5	**1.141**	**0.060**	**1.032**	**0.045**	**0.898**	**<0.001**	**96.966**
9	1.124	0.057	1.026	0.042	0.858	<0.001	66.395
10	1.079	0.058	0.992	0.032	0.763	<0.001	35.499

Table 8 Normalised pupil dilation over the time of each trial

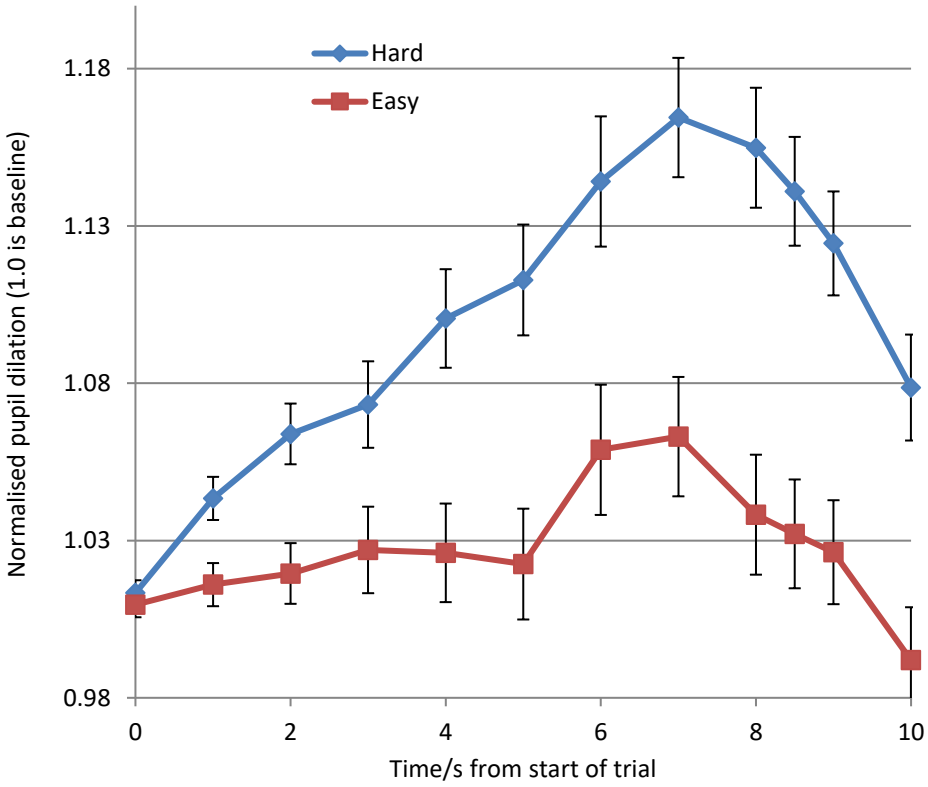

Figure 15 Normalised pupil dilation over the time of each trial (Error bars show standard error)

Blink rate

Condition	Mean	SD
Easy condition	1.24	0.43
Hard condition	1.31	0.45

Table 9 Blinks per second between the two conditions

There is no significant difference between the number of blinks per second for the easy and hard condition. However, there is a moderate effect size between conditions. $F(1,11)= 1.305$, $p=0.278$, $\eta_p^2=0.106$

4.2.5. Discussion

The hypothesis of the experiment proposed that the hard condition would have higher pupil dilation than the easy condition. This was measured as the mean pupil dilation for a one second window starting at 8.5 seconds into the trial. This hypothesis was supported with an extremely large effect size ($\eta_p^2=0.903$).

I considered the pupil dilation across the whole time of the trial. To do this it is useful to review what participants are doing at the different time periods of the trial (see Table 10 below).

Time /s	Easy condition action	Hard condition action
-5 to -3	Baseline measurement	Baseline measurement
-3 to 0	Instructions	Instructions
0	Hear first number	Hear first number
1	Hear second	Hear second
2	Hear third number	Hear third number
3	Hear fourth number	Hear fourth number
4 to 6	Pause	Pause
6	Say first number	Say first number plus one
7	Say second number	Say second number plus one
8	Say third number	Say third number plus one
9	Say fourth number	Say fourth number plus one
10 to 17	Rest period	Rest period

Table 10 Participant actions at different times of the trial

The window which starts 8.5 seconds into the trial is the time of maximum difference between the conditions. This corresponds to the time in the trial when participants have said the first two results and are about to remember the third number and add one to it. Although the two trials have no difference at the 0 second point there is a significant difference between the conditions at all other time periods.

Looking at the pupil dilation over the whole trial shows the similarities and differences between the two tasks. In both tasks, the pupil dilation starts from just above the baseline. However, the hard task starts to increase more steeply even though the task is the same for the first 6 seconds. This could be down to increased anxiety or it could be some participants not following the instructions and performing the addition as they hear the number rather than when they have to speak it. Although some participants admitted that they had done this and were removed from the analysis, it may be that some participants performed the addition early and did not admit it. Both tasks peak at around 7 seconds and then decrease although the easy task decreases faster than the hard one. These results are similar to those obtained by Kahneman and Beatty (1966) as shown in Figure 16 below.

They found a similar peak in pupil dilation at around the 8-9 second mark after the start of the stimulus and describe participants using most effort to remember all the numbers and then as they say each one, they can "unload" it from memory and use less effort. The hard task also involves performing an addition task for each number which is spoken. Unlike the memory aspect of the task this addition task takes a constant amount of effort for each number which causes the pupil dilation to decrease at a slower rate.

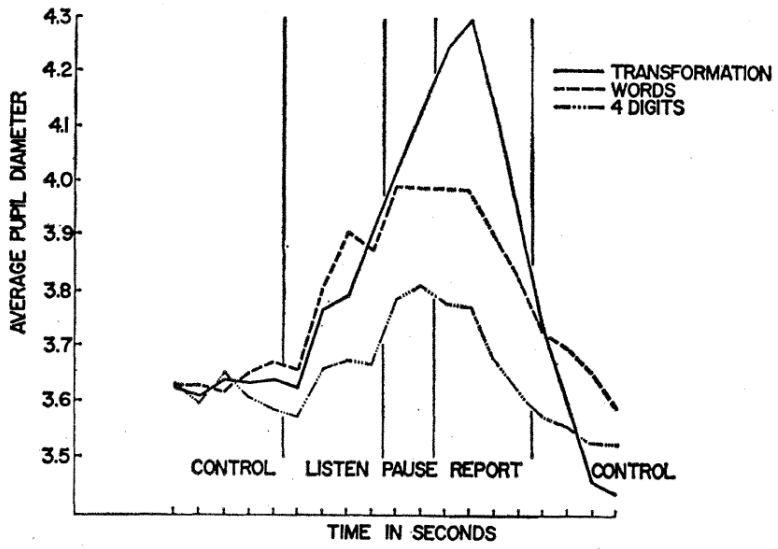

Figure 16 Kahneman and Beatty (1966) pupil dilation against time. The equivalent experiment is the "TRANSFORMATION" condition marked by the steep unbroken line.

Davis (1994) found that harder tasks caused participants to blink less. I was interested to see if this could be an alternative measure of cognitive load. There is a large effect size (η_p^2=0.209) of more blinks during the hard condition but it is not significant (p=0.086). The effect is also in a different direction to that found by Davis; in this experiment participants in the harder task blinked more than the easy task. This may indicate that blinks could be used a measure of cognitive load. However, the effect size is considerably lower than pupil dilation so it is unlikely to prove a more useful measure

I also analysed pupil data from the two second "baseline measurement" period before each trial. During this period participants look at a fixation cross and do not perform any other task. This is used to normalise the pupil dilation measurements. I compared the raw baseline pupil dilation and found there was a significant difference between conditions and a very large effect size (η_p^2=0.467). The mean baseline measurement for the hard condition is about 2% higher than the easy condition. This is probably due to the additional anxiety that participants feel as they know they are about to do the harder task. It is also possible that this is "residual" pupil dilation left over by the previous task, but as participants have a 7 second rest period after each trial this seems unlikely. The task comparison pupil dilation is normalised against this baseline for each trial so it is factored out of the comparisons between conditions. This suggests that future studies should make use of a series of baseline measurements throughout the study to compensate for longer lasting effects such as anxiety which may bias the pupil dilation measurement.

Initial pilot experiments had shown that there was a significant training effect as participants got used to the tasks. Calculating correlations between the trial number and pupil dilation showed that there is a small training effect (r=-0.156) for the easy condition but almost none for the hard condition. The difference between this result and

the pilot experiments may be due to the inclusion of 10 training trials at the beginning of the experiment. After these training trials, the pupil dilation for each task does not change much over the length of the study. This suggests that participants do not find that either task gets significantly easier as the study progresses. This may also be a result of interleaving the trials for the difficult and easy trials. Participants experience both conditions after a similar amount of training and swap from one condition to the other every five trials so do not become accustomed to one condition over the other.

I was concerned that if participants are not focusing on the centre of the screen then the pupil dilation measure may be distorted or this may introduce other confounds into the data. There were two participants who looked at the centre of the screen for less than 80% of the time. I removed these two participants from the data and repeated the same analysis. It made very little difference to the main hypothesis test which was still significant with a huge effect size. Looking at this reduced data set over the full time of the trial showed very little difference from the data set produced by the full set of participants. The only real difference in results from removing these two participants was when comparing the number of times that they blinked. With all participants, the effect size (η_p^2) between conditions was 0.209, without these two participants it dropped to 0.106. It seems likely that participants who did not focus on the centre of the screen were slightly more likely to blink during the hard condition but were otherwise similar to the other participants. I considered removing participants who do not focus on the central stimulus from future studies. However, there is no evidence that this causes a difference in the results and as I had already removed 6 participants for other reasons, removing any more might lead the study to be underpowered. Therefore, future studies should probably keep including participants even if they do not focus consistently on the centre of the screen.

Limitations

The main limitation of this study is that the task had an audio stimulus and a speech response. This ensured that the pupil dilation was not affected by factors such as changing light levels or motor movements but makes it difficult to generalise more common screen-based tasks which use a mouse to give response. Another limitation is that this task put participants under time pressure to give the right answer at the correct time. This made the task more difficult and would have increased pupil dilation. Although this is similar to some tasks involved in fast paced action games it is very unlike the typical situation in a self-paced game where players have as much time as they want. A final limitation is that participants found the task very difficult. Two participants were removed due to getting more than 60% of the answers wrong. Another four were removed because they performed the addition as soon as they heard the numbers, they usually did this because otherwise the task was too difficult.
All wrong answers were removed from the analysis which reduced the total data set and may increase the overall variance. It also suggests that this task is at the limits of many participants' abilities which makes it less similar to other more common or enjoyable tasks that I may wish to measure.

4.3. Experiment Pupil 2: Visual stimulus

The previous experiment confirmed that I could use pupil dilation as a robust method for measuring cognitive load during an audio stimulus task. The results were significant with an extremely large effect size which suggests that they form a good foundation for future experiments.

This next experiment aimed to extend this technique to a situation which was more similar to playing a self-paced game. The last experiment used an audio stimulus and a spoken response. This has the advantage that it avoids potential additional pupil dilation which could be caused by looking at a visual stimulus or making a motor action to give a response. However, a typical self-paced game uses a screen to show players the state of the game and they respond using a touchscreen or mouse. This next experiment used the same number task but changed the mode of delivery so that participants saw their task stimulus on a screen and made their response by using a mouse.

To avoid potential confounds due to additional pupil dilation caused by the visual stimulus or motor actions both conditions used exactly the same visual stimulus and required exactly the same types of motor action. This meant that any additional pupil dilation created by additional factors was the same for each condition and so was factored out when the conditions were compared. The only way that conditions differed was in the instructions which were given to participants. These instructions specified which task they were to perform. As in the previous experiment one task required high cognitive load while the other required low cognitive load.

Hypothesis

The previous audio experiment confirmed the hypothesis that the highest effect size between conditions would be during a 1 second window which started at 8.5 seconds from the start of the task. This experiment uses the same task as the audio experiment so it seems likely that the highest effect size would be in same time window.
Therefore, the hypothesis of the experiment is:
For a 1 second window starting at 8.5 seconds from the start of the task, the size of participants' pupils will increase when they are performing the hard task compared to the easy task.

Design

This was a within-participants design with two conditions. The independent variable was the difference in cognitive load needed for each condition. The dependent variable was the amount of pupil dilation in each condition. Participants were asked to perform a different task for each condition. One task is easy (low load) and the other is hard (high load). The stimulus for the tasks is described under *Materials* (below). The only difference

between the stimulus for each condition is the instructions which participants are given at the start of the trial.
I also collected data on the baseline pupil dilation without any stimulus, the length of each trial, the number of times that participants blink during the experiment and whether they were looking within a central area of interest. This data was used to get a clearer picture of what was happening during the experiment which could be useful when designing new experiments and reflecting on the results of this one.

Participants

The previous experiment had 14 participants and showed a significant result with a strong effect size. This study used a very similar task so I decided to use the same number of participants. 14 students and staff from the University of York took part in the study. 8 were women and ages ranged from 18 to 55 (mean=29.0). 11 of the participants were not native speakers of English.

Materials

Participants were asked to carry out two different tasks which are developed from the initial study by Kahneman and Beatty (1966) . Both tasks involved a similar visual stimulus displayed on a screen. Each trial consists of the following stages:

1. Pause for 2 seconds. This is used to get an initial baseline reading of participants' pupil sizes.
2. Participants hear instructions which tell them which task they are doing
3. Pause for 3 seconds
4. Participants see 4 random numbers from 0 to 4 appear on the screen, one at a time with one second between them.
5. Pause for 2 seconds.
6. Participants see the numbers from 0 to 4 on the screen. They click on 4 of them one after the other to give their answer.
7. Pause for 7 seconds. This gave participants' pupils a chance to relax after performing the task.

For each task, the set of numbers that they saw was randomly generated by the website *random.org.* The numbers were generated with the condition that all the numbers in each set of 4 was different. Each participant saw the same sets of numbers for each condition. Some studies such as Cavanagh et al. (2014) which measure pupil dilation have used luminance matched stimuli to prevent changing light levels from adding noise to the data. I decided that this was not possible for this experiment due to the way that the stimuli changes over time. However, the stimuli are designed to minimise changes in luminance even if they could not be removed completely. This means that the stimuli displayed on the screen are low contrast and small so that they fill a small fraction of participants' field of view. During the task participants were asked to focus on a light grey cross (RGB 51,51,51) in the middle of a dark grey (RGB 85,85,85) 24″ screen. The cross was 4 cm high. Numbers were displayed in the same shade of grey and the same size as the cross.

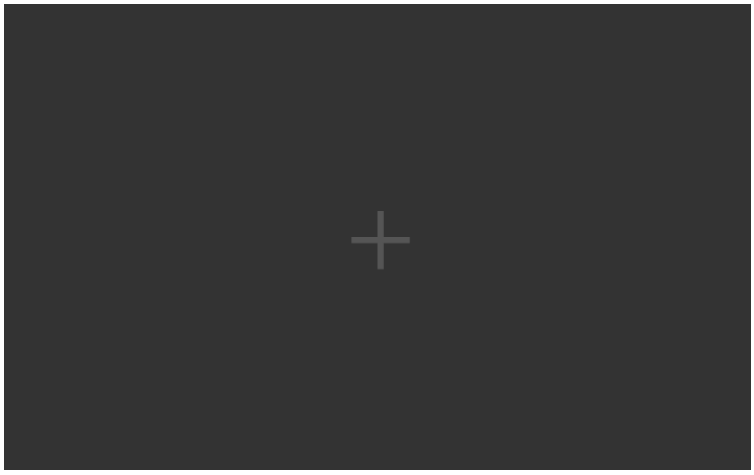

Figure 17. Initial cross shown during ten second pause

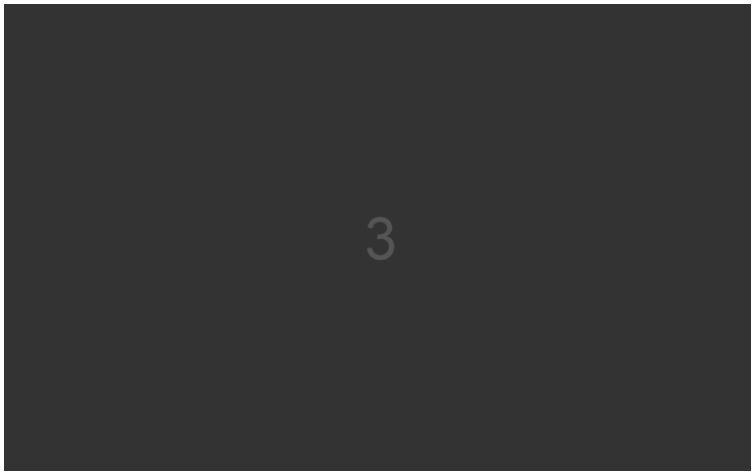

Figure 18. One of the four numbers displayed during the memorise section of the stimulus

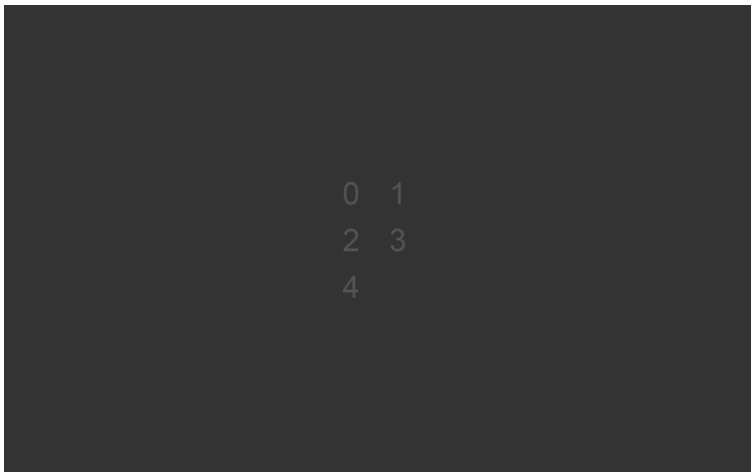

Figure 19. The numbers that participants click on to give their response

The stimulus was created using the Eyelink Experiment builder software. The same Eyelink eyetracker was used as the previous audio experiment. Participants were positioned in a chin rest in the same way as the previous audio experiment.

4.3.2. Procedure

Each participant took part in both conditions. To control for possible training or fatigue effects the order of the conditions was counter-balanced so that half the participants performed the easy task first followed by the hard task. The other half of participants did it the other way around. Participants did a consent procedure and were then calibrated with the eye tracker. They then had the first task explained to them and they performed 5 trials of this task. They then had the second task explained to them and performed 5 trials of that task. These ten trials made up the training phase of the study. For the main testing phase of the study participants alternated between doing 5 trials of the first task followed by 5 trials of the second task until they had done 30 main trials. Before each trial, they heard audio instructions which told them which task they needed to perform.

4.3.3. Results

All of these results consider only trials number 11-40 and ignore the first ten trials. This is because those first ten trials were considered as training trials where participants are still learning the tasks. Where pupil dilation is reported, it is a normalised pupil dilation calculated by dividing the diameter of the participant's pupil by a baseline pupil diameter which is measured at the start of each trial. Unless otherwise stated the number of participants (N) is 14.

Hypothesis

The hypothesis proposed that there would be an increase in pupil dilation between the easy and hard conditions when measured in a one second window starting 8.5 seconds into the trial. The hypothesis was supported as there was a significant difference in the mean pupil dilation for each participant during this window; $F(1,13)= 23.441, p<0.001$, $\eta_p^2=0.643$. There was an extremely large effect size (η_p^2) of 0.643 between the two conditions.

Condition	Mean	SD
Easy	1.08	0.02
Hard	1.12	0.04

Table 11. Summary of normalised pupil dilation for a one second window staring at 8.5 second.

Pupil dilation over time

I considered the pupil dilation over the whole size of the trial (See Table 12 and Figure 20).

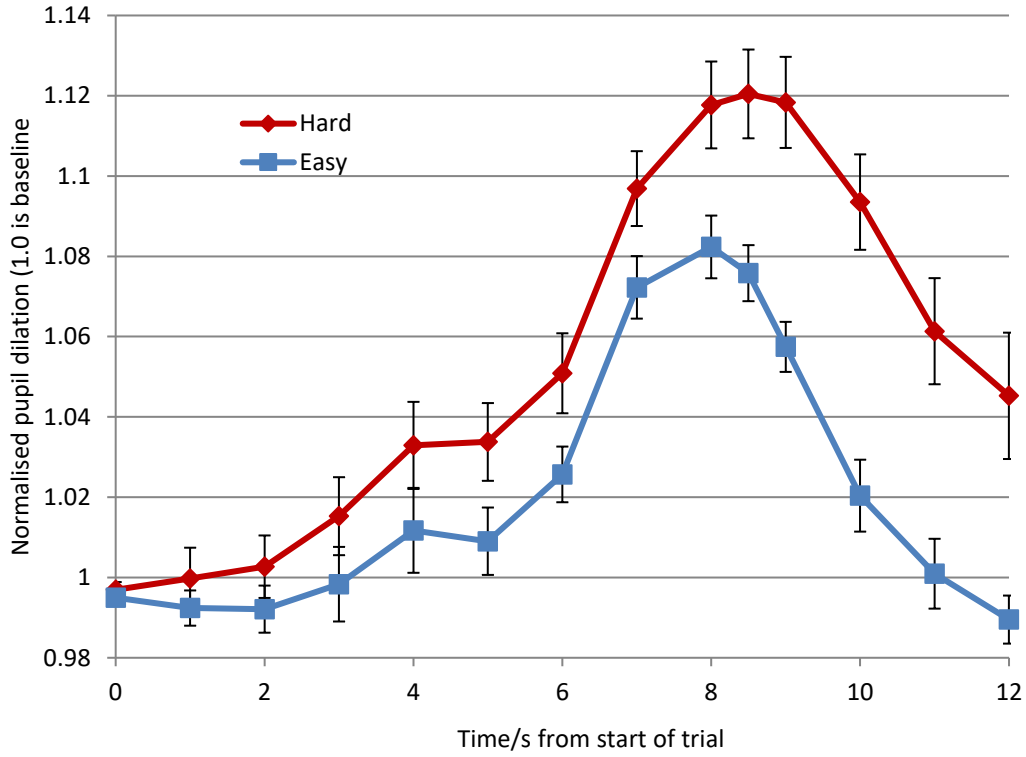

Figure 20 Normalised pupil dilation across the time of the trial (error bars show standard error)

	Hard condition	Easy condition			
Time /s	Mean (SD)	Mean (SD)	Effect size η_p^2	p value	F(1,13)
0	0.997 (0.007)	0.995 (0.007)	0.045	0.449	0.609
1	1.000 (0.029)	0.992 (0.018)	0.188	0.107	3.004
2	1.003 (0.030)	0.992 (0.024)	0.460	0.005	11.061
3	1.015 (0.036)	0.998 (0.035)	0.584	0.001	18.222
4	1.033 (0.038)	1.012 (0.037)	0.593	0.001	18.941
5	1.034 (0.034)	1.009 (0.029)	0.536	0.002	15.012
6	1.051 (0.036)	1.026 (0.024)	0.482	0.004	12.087
7	1.097 (0.035)	1.072 (0.029)	0.417	0.009	9.304
8	1.118 (0.040)	1.082 (0.029)	0.528	0.002	14.564
8.5	**1.120 (0.042)**	**1.076 (0.025)**	**0.643**	**<0.001**	**23.441**
9	**1.118 (0.044)**	**1.057 (0.022)**	**0.729**	**<0.001**	**34.886**
10	1.094 (0.045)	1.020 (0.032)	0.662	**<0.001**	25.437
11	1.061 (0.049)	1.001 (0.030)	0.572	0.001	17.342
12	1.045 (0.056)	0.990 (0.022)	0.468	0.005	11.429

Table 12 Normalised pupil dilation over the time of each trial

Blink rate

Condition	Mean	SD
Easy	0.404	0.202
Hard	0.422	0.212

Table 13 Summary of blinks per second for each participant between the two conditions

I calculated the mean number of blinks per second for each participant during all the trials for each condition. There was no significant difference between the rates of blinking for the easy and hard condition, however there was a moderate effect size. F(1,13)= 2.857, p=0.115, η_p^2=0.180

Baseline

The baseline pupil dilation for each trial is the mean pupil dilation for the initial two second pause time when participants are doing nothing except looking at a fixation cross. Baseline pupil dilation is given in "arbitrary units" specific to the Eyelink eye tracker. There was not a significant difference in the baseline pupil dilation between the easy and hard conditions. There was a small effect size between conditions.; F(1,13)= 0.919, p=0.355, η_p^2=0.066.

Condition	Mean	SD
Easy	4870.96	398.64
Hard	4907.01	395.44

Table 14 Pupil size during the baseline period between two conditions

Trial Duration

In this experiment participants could take as much time as they liked to give their responses, although they were encouraged to answer as quickly as possible. This is unlike the audio experiments in which participants had to give their answers within a particular time period.
There was a significant difference in the amount of time that participants took to respond to the easy and hard conditions. There was also an extremely strong effect size between conditions; F(1,13)= 27.407, p<0.001 , η_p^2=0.678

Condition	Mean/ms	SD/ms
Easy	2964	682
Hard	4503	1483

Table 15 Number of milliseconds between seeing the prompt to respond and giving the last response click.

Training and Fatigue effects

To see if there are any training or fatigue effects, I calculated Pearson's correlations between the mean pupil dilation at 8.5 seconds and the trial number. There were no significant correlations between pupil dilation and trial number in either condition. This indicates the amount of pupil dilation due to the task does not change from the beginning to the end of the experiment.

Condition	Pearson's r	Significance t value	Significance p	DF
Easy	0.00331	0.0175	0.986	28
Hard	-0.158	-0.846	0.405	28

Table 16 Correlations between pupil dilation and trial number

I also calculated Pearson's correlations between the mean duration of each trial and the trial number. There were significant correlations for both the easy and hard condition. This indicates that participants are faster towards the end of the study than they are at the beginning.

Condition	Correlation r	Significance t value	Significance p
Easy	-0.403	-2.333	0.027
Hard	-0.518	-3.201	0.0034

Table 17 Correlations between trial duration and trial number

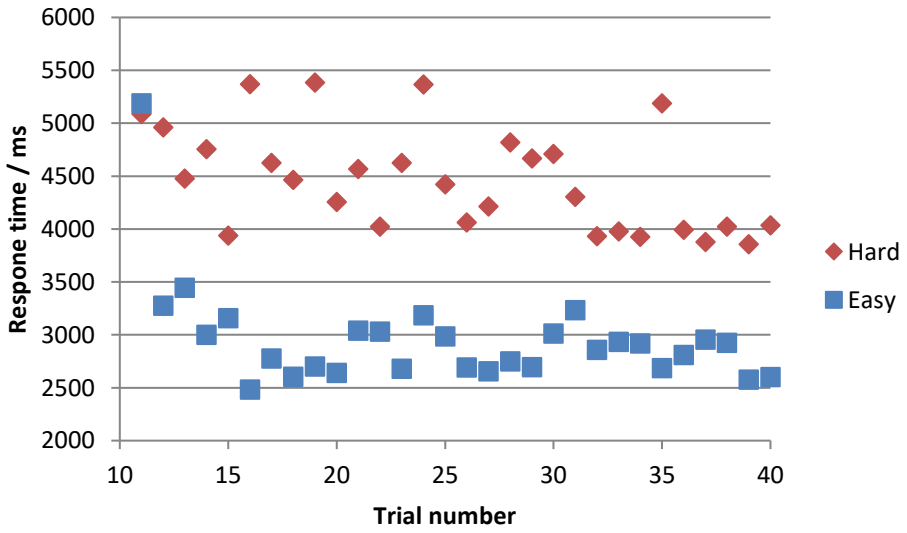

Figure 21 Scatterplot of mean trial duration and trial number

Are participants focused on the central stimulus?

I created an area of interest (AOI) 640 pixels x 640 pixels across in the centre of the screen. This corresponds to 10.4° of participants' field of view and is the same area of interest that I created for the audio experiment. I then looked at what percentage of time participants looked within this AOI. As can be seen all the participants were looking in this area for at least 97% of the study.

Participant ID	% time on central Area of Interest
70	100
71	100
72	99.96
73	99
74	98.42
75	99.98
76	99.91
77	99.33
78	100
79	97.22
80	99.76
81	100
82	100
83	97.98

Table 18 Percentage time each participant spent looking in the central area of interest

Removing one doubtful participant

One participant (Id 83) initially had difficulty with the task but then suddenly got a lot better. When I discussed this with him after the study, he said that he started using a method "like in online games". I was not sure what this entailed so left his data in the study for the initial analysis. This section describes the same analysis but without this "doubtful" participant. The hypothesis that there would be a significant difference in pupil dilation between conditions at the 1 second window starting at 8.5 seconds was still supported, as there was a significant difference in the mean pupil dilation for each participant during this window; $F(1,12)= 49.767, p<0.05, \eta_p^2=0.806$. The effect size ($\eta_p^2$) between the two conditions increased from 0.643 with this participant to 0.806 without him. The maximum effect size between the conditions still happened at the 9 second period but increased from 0.729 to 0.822.

Condition	Mean	SD
Easy	1.074	0.02
Hard	1.125	0.04

Table 19 Normalised pupil dilation for a one second window staring at 8.5

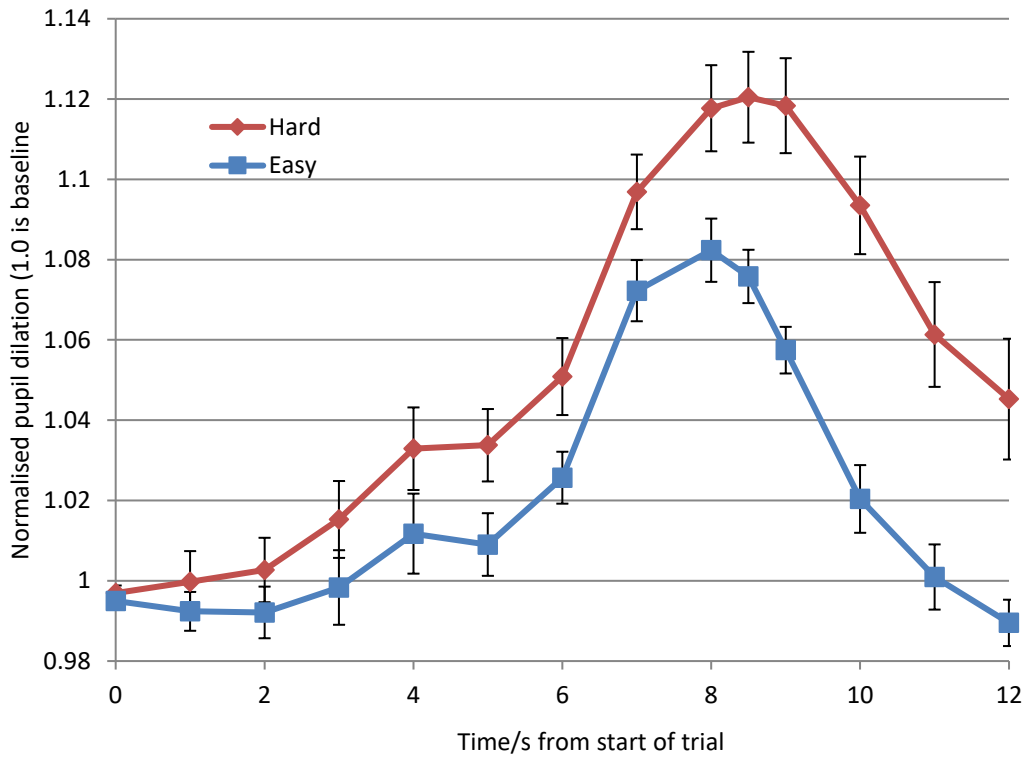

Figure 22 Normalised pupil dilation over the time of each trial (error bars show standard error)

	Hard Condition		Easy Condition		Effect size η_p^2	p value	F(1,12)
Time/s	Mean	SD	Mean	SD			
0	0.996	0.007	0.994	0.006	0.051	0.437	0.647
1	0.997	0.028	0.990	0.016	0.170	0.143	2.451
2	0.999	0.028	0.989	0.021	0.438	0.010	9.343
3	1.012	0.035	0.995	0.033	0.573	0.002	16.095
4	1.031	0.039	1.010	0.038	0.568	0.002	15.770
5	1.035	0.035	1.008	0.030	0.568	0.002	15.761
6	1.053	0.036	1.025	0.025	0.590	0.001	17.289
7	1.100	0.034	1.070	0.028	0.690	<0.001	26.719
8	1.122	0.039	1.079	0.028	0.774	<0.001	41.065
8.5	**1.125**	**0.040**	**1.074**	**0.025**	**0.806**	<0.001	**49.767**
9	**1.124**	**0.041**	**1.057**	**0.022**	**0.822**	<0.001	**55.508**
10	1.099	0.043	1.019	0.032	0.733	<0.001	32.974
11	1.066	0.048	1.000	0.031	0.634	0.001	20.783
12	1.049	0.057	0.988	0.022	0.523	0.003	13.150

Table 20 Normalised pupil dilation over the time of each trial

I recalculated the differences between the blink rate, baseline measure and trial duration without this participant and I also recalculated correlations to investigate training and fatigue effects. All of these results were similar to the previous analysis so are not

reported here. In summary, the overall effect size between conditions was significantly higher without this participant but there were only small changes to the other measures.

Analysis relative to click time

As there is a significant difference between the time taken to complete the easy and hard conditions. I looked for some way to compare the two conditions which took account of this. I compared the two conditions relative to the time when participants clicked their first answer. These are shown in Table 21 and Figure 23.

	Hard condition	Easy condition			
Time/s relative to first click	Mean (SD)	Mean (SD)	Effect size η_p^2	p value	F(1,13)
-8	1.001 (0.005)	1.000 (0.006)	0.004	0.821	0.053
-7	1.002 (0.018)	0.999 (0.007)	0.059	0.383	0.815
-6	1.007 (0.027)	0.998 (0.019)	0.340	0.022	6.704
-5	**1.016 (0.031)**	**1.000 (0.027)**	**0.804**	**<0.001**	**53.254**
-4	1.031 (0.034)	1.009 (0.036)	0.697	<0.001	29.846
-3	1.039 (0.035)	1.017 (0.032)	0.631	<0.001	22.200
-2	1.044 (0.033)	1.014 (0.027)	0.667	<0.001	25.994
-1	1.066 (0.039)	1.037 (0.027)	0.624	<0.001	21.563
0	1.102 (0.038)	1.076 (0.030)	0.560	0.001	16.574
1	1.117 (0.042)	1.082 (0.031)	0.609	0.001	20.221
2	**1.112 (0.046)**	**1.056 (0.022)**	**0.707**	**<0.001**	**31.374**
3	1.077 (0.045)	1.022 (0.025)	0.633	<0.001	22.393

Table 21 Normalised pupil dilation over time relative to the time of participants' first click

There are two peaks in the maximum difference between the two conditions. The first is 5 second before the first click with an effect size of 0.804 and the second is at 2 seconds after the click with an effect size of 0.707.

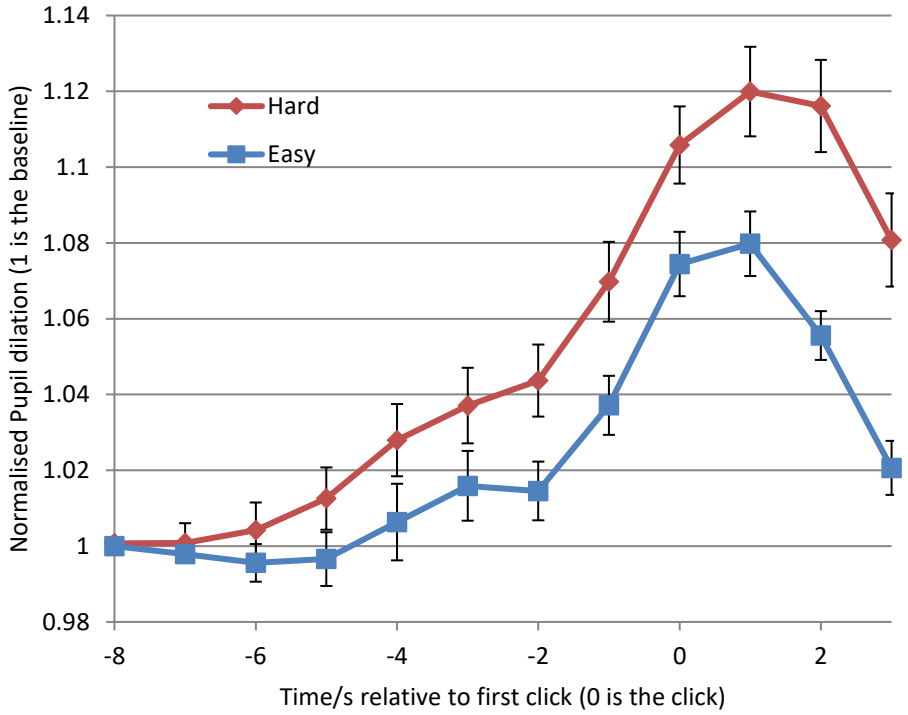

Figure 23 Normalised pupil dilation over the time of each trial relative to the first click (Error bars show standard error)

4.4. Discussion

The hypothesis of the experiment proposed that the hard condition would have higher pupil dilation than the easy condition. This was measured as the mean pupil dilation for a one second window starting at 8.5 seconds into the trial. This hypothesis was supported with an extremely large effect size (η_p^2=0.643).

The time of maximum difference between conditions is very close to my prediction and similar to the previous audio experiment. I predicted that 8.5 seconds into the trial would be the time of maximum difference between conditions. At that time, the effect size (η_p^2) is 0.643. However, the effect size is actually highest at the 9 second time period when it is 0.729. This is 0.5s later than in the previous audio stimulus experiment. In this visual stimulus experiment participants do not have to answer within a particular time and also have to click on the right answer which takes more time than just saying the answer. Both of these factors may mean that participants take an additional 0.5s to process and give their answer. Taking this into account, the period of maximum pupil dilation difference in the audio stimulus experiment is very similar to the equivalent time period for the visual stimulus. This corresponds to the time in the trial when participants have clicked on the first two results and are about to remember the third number and add one to it. Although the two conditions have very little difference at the 0 and 1 second time points there is a significant difference between the conditions at all other time periods.

Participants in this study did not have to give their answers within a particular time period, unlike the previous Audio stimulus study, although they were encouraged to answer as quickly as possible. This meant that individual trials within the study varied in the time taken to complete them by each participant. There is a significant difference between the time taken to complete the task in each condition. There is also a strong effect size (η_p^2=0.678). The difference between the mean time to complete a trial in the hard and easy conditions is 1539ms. It is possible that this difference in time explains some of the difference in pupil dilation when comparing conditions at particular time periods across the trial. However, looking at the graph of pupil dilation over time shows that the peak of the hard condition is only around 0.5 of a second in front of the peak of the easy condition. This suggests that comparing both conditions during the same time 1 second window will still give a reasonable indication of the difference between them. I also analysed the data relative to the time of the first click which I will discuss later.

I looked in detail at how the pupil dilation varied over the time of the trial. Table 22 (below) shows what participants are doing at the different time periods of the trial.

Time /s	**Easy condition action**	**Hard condition action**
-5 to -3	Baseline measurement	Baseline measurement
-3 to 0	Hear instructions	Instructions
0	See first number	See first number
1	See second number	See second
2	See third number	See third number
3	See fourth number	See fourth number
4 to 6	2 second pause	2 second pause
6 to 6+time needed to click on the numbers (t)	Click on four numbers in order	Click on each of the four numbers+1 in order
6+t to 13+t	7 second rest period	7 second rest period

Table 22 Participant actions at different times of the trial

In both tasks, the pupil dilation starts from just above the baseline. However, the pupil dilation during the hard task starts to increase more steeply even though the task is the same for the first 6 seconds. This could be down to increased anxiety or it could be some participants not following the instructions and performing the addition as they hear the number rather than when they have to speak it. The easy task peaks at around 8 seconds and hard task peaks a little later at 8.5 seconds. Then both tasks decrease although the easy task decreases faster than the hard one. Similarly to the previous experiment this fits in with Kahneman and Beatty (1966)'s description of participants using most effort to remember all the numbers and then as they click on each one they can "unload" it from memory and use less effort. The hard task also involves performing an addition task for each number. Unlike the memory aspect of the task this addition task takes a constant amount of effort for each number which causes the pupil

dilation to decrease at a slower rate.

As with the previous experiment I looked to see whether blink rate could be used as an alternative measure of cognitive load to pupil dilation. I looked at whether participants blinked more in the hard task than the easy task. There was no significant difference between conditions although there was a moderate effect size. This result is similar to the audio stimulus experiment which also found no significant difference in blink rates. Davis (1994) found that that harder tasks caused participants to blink less. The tasks that Davis investigated were those performed by pilots flying fighter-bombers and it is likely that this difference from Davis's result was due to these experiments using a different task and mode of presentation. Although this task is cognitively difficult for participants, there is not much variation in the screen display or complex information to take in which may explain the lack of difference in blink rates.

I also analysed data from the two second "baseline measurement" period which happens before each trial. During this period participants look at a fixation cross and do not perform any other task. This is used to normalise the pupil dilation measurements. I compared the raw baseline pupil dilation between conditions. I found there was no significant difference (p=0.355) and a small effect size (η_p^2=0.066). This is in contrast to the audio experiment which had a significant difference between conditions for the baseline measurement. This may be because the audio experiment needed participants to respond within a particular time period. This made the task harder and more stressful so when participants anticipated the stress of the hard task this increased their pupil dilation which changed the baseline measurement. The task comparison pupil dilation is normalised against this baseline for each trial so it is factored out of the comparisons between conditions.

Previous experiments had shown that there was a significant training effect as participants got used to the tasks. Calculating correlations between the trial number and pupil dilation showed that there is a small and not significant training effect for the hard condition and almost no effect for the easy condition. This is similar to the audio stimulus study which showed no significant change in pupil dilation over the time of the study. I also investigated to see if there was a training effect on the length of time which it took participants to complete each trial. This was very different from the pupil dilation training result. There was a significant moderate negative correlation between the length of each trial and the trial number for both the easy (r=-0.403) and the hard condition (r=-0.518). This indicated that participants got quicker at the task throughout the study even though they had already had 10 training trials before it. Some of this speed increase may be due to learning how to click on their answer quicker rather than the actual performance of the mental task. Looking at a scatterplot of the trial number against time taken shows that initially the time taken varies considerably between trials but it settles down to a shorter, more consistent time for the last few trials.

I was concerned that if participants are not focusing on the centre of the screen then the pupil dilation measure may be distorted or there may be other confounds. I used the

same central Area of Interest (AOI) that I had for the audio experiment and calculated what percentage time of the experiment participants looked within it. All the participants looked within this AOI for at least 97% of the study time. This suggests that unlike the audio experiment all the participants were focused on the stimulus for the whole duration of the experiment. This is probably due to this being a visual-stimulus based study so that participants had to look at the stimulus to complete the task.

One participant initially had difficulty with the task but then suddenly got a lot better. When I discussed this with him after the study, he said that he started using a method "like in online games" which had helped him succeed at the task. I was not sure what this entailed so left his data in the initial analysis. I then tried rerunning the analysis without his data. Although the overall shape of both graphs remained the same this increased the effect size (η_p^2) at 8.5 seconds from 0.643 to 0.806. Although the point of highest effect size remained at 9 seconds it also increased from 0.729 to 0.822. This suggests that this participant may not have been following the instructions as given and was performing the plus one calculation when he first heard the numbers rather than when he had to give the answer. This indicates that in this type of pupil dilation experiment participants can appear to follow the instructions but are actually performing different mental tasks at different times to those that they have been instructed. This can affect the pupil dilation data and reduce the effect size between conditions. This suggests that future experiments should be careful to reject participants who seem to be behaving differently or have unusual patterns in their experiment data as this may indicate that they have not been performing the task as requested.

As previously mentioned, one of the issues with analysing data from this study was the variable length of each trial which may create issues with comparing data from the same time period. To compensate for this, I analysed the data relative to the time when participants made the first click to give their answer. The results give a graph of pupil dilation against time which is similar to original analysis. Both conditions start together and rise to a single peak, with the hard condition rising higher, before dropping down at the end of the trial. However, there are two points of maximum effect size. The first is at 5 seconds before the first click with an effect size of 0.804. The other is at 2 seconds after the click when the effect size is 0.707. The peak of difference at -5 seconds may be caused by participants who did not follow the instructions performing some addition early. Although significant this difference is not caused by the hard condition producing a particularly large pupil dilation, it is just that the easy condition has not risen at all and the variance is very low. This creates the very large effect size between conditions. The other peak which happens 2 seconds after the first click roughly corresponds to the time 8-9 seconds into the trial when participants are recalling the third number. This second peak is thus comparable to the single peak seen when the data is analysed relative to the start of the trial. Subsequent experiments are also likely to use tasks which take a variable amount of time. To make sure that participants are performing similar mental processes at each time period the data should be analysed relative to the time when participants first respond to give their answer.

Limitations

The main limitation of this study is that participants can take as much time as they like to provide their responses. Analysis showed that the time they take varies significantly between conditions. There is also a significant practice effect over the time of the study which means that they perform the later trials faster than the earlier ones. This makes it more difficult to compare pupil dilation at equivalent points in the tasks. Participants click four times during a trial to provide their response, so although it is possible to anchor data analysis around the time of one click, the time of the other clicks can still vary and make comparison more difficult. However, the purpose of this study was to develop a method of measuring pupil dilation during self-paced games. Self-paced games also allow participants to provide responses in a variable amount of time and may feature multiple responses and practice effects. That the main hypothesis of the experiment was still supported despite these potential confounds, shows that this technique may be robust enough to create a measure during real gameplay.

My hypothesis for this study was that the period of peak difference between conditions would be the same as the audio stimulus study and occur at 8.5 seconds into a trial. The actual peak was slightly later at 9 seconds, which is probably due to removing the time limit and asking participants to click rather than say their response. There was one participant included in the study whose performance suddenly improved. This may be because did not follow the instructions and started performing the addition task earlier in the trial. Removing this participant from the data showed a significant increase in effect size. Analysing the pupil dilation during the hard task, showed that there is a small but significant peak around the time when participants hear the numbers. This was not present in the easy condition but was still present even with the "doubtful" participant removed. This suggests that other participants are also performing the addition earlier than instructed during at least some of the trials. This suggests that a key limitation of addition tasks such as this one is that participants may not perform mental tasks during the instructed time period and instead take shortcuts to succeed at the task. These shortcuts can then only be detected by looking at the pupil dilation data.

4.5. Chapter conclusions

These two experiments both showed that pupil dilation can be used as a measure of cognitive load during number-based memory and manipulation tasks. The first experiment demonstrated this with an audio stimulus and spoken response. The second experiment demonstrated this with a visual stimulus and a mouse click response which is similar to that used during self-paced games. Both experiments showed a significant difference between easy and hard tasks with an extremely large effect size between conditions. Both experiments used a small number of participants (14) and if future pupil dilation experiments have similar effect sizes then they may also show significant results with similar small sample sizes. Not requiring a large sample size is a useful property for a measure of game experience and these experiments suggest that this may be the case for pupil dilation.

The experiments included a number of measures to prevent additional noise factors from confounding the data. Both experiments took a baseline measure of pupil dilation before participants started each trial. The audio stimulus experiment showed a significant difference in baseline between the easy and hard tasks. This is probably due to differences in arousal between conditions; participants were more stressed by the hard condition even before they started the task. Both experiments normalised all pupil dilations readings against the baseline measurement which ensured that these differences in arousal did not affect the analysis. Future pupil dilation experiments on games may create similar changes in arousal so should also normalise pupil dilation against a baseline measurement for each trial. The experiments also included a training period and interleaved the easy and hard conditions. This was to prevent training effects from adding confounds to the data and appears to have been successful as no significant training effects were found in either experiment.

There were differences between experiments in the amount of time that participants had to give their responses. In the audio stimulus experiment participants had to give their responses within a particular time as indicated by a click sound. In the visual stimulus experiment participants were simply required to give their responses as quickly as possible rather than within a set time period. This meant that the time to give those responses varied from trial to trial. There was a significant difference in time taken between the easy and hard condition. There was also a significant training effect which meant that participants were able to complete the task more quickly in the later trials. My initial data analysis compared time windows in each condition which were measured from the start of the trial. However, the different conditions take different amount of time to complete. To ensure that participants are at a similar stage in the trial for each condition I also analysed the data by measuring the time window relative to the moment that participants made their first click to produce their answer. Both analyses produced similar significant results and effect sizes. Future studies which allow participants to respond in variable amount of time should perform the analysis relative to the moment of the first participant response to avoid confounds due to differences in response times.

I also investigated some participant behaviours during the experiment to see if they affected the experimental results. In the audio stimulus experiment several participants looked away from the centre of screen for significant periods of the experiment. There was no evidence that this affected their pupil dilation data and such participants should be included in future experiments. In the visual stimulus experiment one participant suddenly improved their performance. They probably did this by performing the addition early in the experiment rather than later as they had been instructed. Further analysis showed that by not following instructions they reduced the difference between conditions and overall effect size of the experiment. Future experiments should be aware that participants may not follow the instructions given and perform different mental tasks from those instructed. There may be few outward signs of this apart from

the pupil data so care should be taken to remove any participants who behave differently during the experiment.

In summary, these two experiments showed that pupil dilation can be an effective way of measuring cognitive load in tasks which require memory and manipulation of numbers. Performing these experiments allowed me to explore the experimental technique and analysis necessary to maximise effect size and reduce confounds from other sources of pupil dilation. In the next chapter, I apply these findings to the task of using pupil dilation to measure the cognitive load used while playing a self-paced game and to using that measure to understand the experience of playing that game.

5. Measuring cognitive load during gameplay

In the previous chapter I showed that pupil dilation can be used to measure the amount of cognitive load used when performing a task. Initially I replicated with modern equipment and analysis an audio based experiment first performed by Kahneman and Beatty (1966). I then showed that this technique is robust enough to be extended to tasks with a visual stimulus which participants response to with a mouse. This promises that pupil dilation could be used as a measure of cognitive load during more complex interaction such as during a self-paced game. The goal of this chapter is to extend this measurement technique first to "game-like" tasks and then to measuring cognitive load whilst participants are playing a real game.

5.1. Overview of experiments

The experiments in this chapter look at measuring changes in cognitive load whilst participants play a game. They then look at linking those changes in cognitive load to changes in the player's experience of the game. The analysis looks at both changes in experience within a single game and also consistent changes in experience which are found between games of different design. There are three experiments in this chapter which are all based on a clone of the popular mobile game *Two Dots* which is described in chapter 3. Even though it is a simple game there are still several potential issues with measuring pupil dilation during a game of *Two Dots*. These issues are described below.

5.1.1. Issues with measuring changes in pupil dilation during the game of *Two Dots*

There are three main issues with measuring changes in pupil dilation during the game of *Two Dots.* They are *Timing, Luminance* and *Response* and are described in more detail below.

Timing

Pupil dilation due to cognitive effort starts between 2-300ms after the stimulus (Beatty, 1982, Gagl et al., 2011) and can take some time to revert back to the base measure

afterwards. In *Two Dots* players may make one move which takes a lot of effort and then immediately follow this with another move which takes much less effort. Because the second move follows immediately this makes it difficult to tell which aspects of the pupil dilation correspond to which move.

Luminance

Pupil dilation is affected by changes in the amount of light reaching the eye. If an experimental stimulus changes in brightness this could potentially introduce noise into the pupil response data. To reduce changes in luminance coming from the stimulus I converted the game of *Two Dots* to monochrome (this is described further in chapter 3). This reduces the changes in luminance coming from the stimulus but does not remove them completely. If a stimulus has constant luminance throughout the experiment it is known as "luminance balanced". Because it was not possible to completely luminance balance the whole stimulus over time the experiments in this chapter compared the pupil dilation between tasks rather than looking at absolute values. Both tasks in the first experiment used an identical stimulus so even though the luminance of the stimulus did change throughout the experiment, participants in each task see the same luminance at each stage. Therefore, the differences in pupil dilation between tasks was due to the mental processes performed rather than any differences in luminance.

Response

In the desktop version of *Two Dots* players use a mouse to drag a line between dots to join them, lifting their finger from the mouse button completes a move and removes the dots. So, a move consists of starting from one dot, joining it to at least one other and then deciding to finish the move. Typically, in a pupil dilation experiment I need to link pupil dilation across time to a particular moment in time which can be called the "response event". In *Two Dots* the response happens across a particular span of time and it is not easy to see when players put in the mental effort to decide on the move. It could have been at the start of the move, during it or only right at the end.

5.1.2. Experimental plan

For the first experiment which looked at pupil dilation in games I decided to simplify the game of *Two Dots* to avoid the issues discussed above which relate to the timing and complexity of the player response. In this experiment participants complete a series of game like "puzzles". These puzzles are the same tasks which they would perform in the actual game. However, unlike the game they have no choice in which puzzles they will attempt and they give their answer using a single click rather than joining dots. If the differences in cognitive load used in the puzzles are reflected by differences in pupil dilation then it is likely that similar experiment could be used to measure the amount of cognitive load used when playing a full game of *Two Dots.* I then performed an exploratory analysis of the changes in pupil dilation across time during a full game of *Two Dots* to see if these reflected any changes in the game experience across time. The second experiment looked at three different versions of the game *Two Dots* (described in chapter 3) which should require different amount of cognitive load to play. This experiment compared the pupil dilation between these different game versions to see if

it reflected these differences in cognitive load. I then repeated this in the third experiment with a revised hypothesis that none of these game versions imposed sustained cognitive load.

5.2. Experiment Pupil 3: *Two Dots* puzzle

The main aim of this experiment is to see whether pupil dilation can be used to measure the difference in cognitive effort used between solving an easy game puzzle and solving a hard game puzzle. This experiment has an analogous design to the pupil dilation experiments in the previous chapter. As in those experiments, participants perform both an easy task and a hard task. Also like the previous experiments the tasks are presented in groups of 5 trials with each group containing tasks of the same type. Those groups are then interleaved which reduced the likelihood that participants are better at one task due to training effects. Both easy and hard tasks are derived from the game *Two Dots.* As part of the training phase of the experiment participants also play a full version of *Two Dots* while their pupil dilation is recorded. This pupil dilation data was then used for an exploratory analysis to investigate whether there are patterns of pupil dilation across longer periods of game play which indicate particular game play experiences.

Hypothesis
Previous pupil dilation studies (reported in chapter 4) showed that pupil dilation varies across the time period of a task and that when comparing pupil dilation between tasks it is useful to consider the trial period in "windows" of 1 second duration. The pupil dilation for each window is calculated by taking the mean value of all pupil measurements within that time period.
A pilot study (see below) found that the hard task takes on average around 5 seconds longer to complete than the easy one. If I measured the time window from the start of the trial then comparing the same time window in the easy and hard tasks would be comparing pupil dilation at different stages in the task. For example, at same the time when participants in the easy task were giving their answer, participants in the hard task would still be thinking and not give their answer for another 5 seconds. To avoid this issue, I measured the start of each time window from the moment when participants clicked to give their answer. This meant that all time windows were relative to participant action so were more likely to compare similar mental processes. The same analysis technique was used for a pupil dilation experiment in the previous chapter and found to be effective.

A pilot study (see below) determined that the difference between conditions was likely to be greatest during the 1 second period starting 3 seconds before participants have clicked to give their answer. For the purposes of this hypothesis pupil dilation will be measured as the mean value of this 1 second period.
The hypothesis of the experiment is:

The size of participants' pupils will increase when they are performing a hard game puzzle compared to an easy game puzzle. In particular this will be significant during the 1 second time window which starts 3 seconds before participants click to give their answer.

5.3. Method

Design

This was a within-subjects design with two conditions. The independent variable was the difference in the task which participants performed for each condition. The dependent variable was the size of participants' pupils during each of the different condition.

Tasks

Participants were asked to perform a different *Two Dots* related task for each condition. One task is easy (low load) and the other is harder (high load). In the easy condition participants are performing a simple pattern matching task, whereas in the hard task they perform the same pattern matching task but also need mentally manipulate the symbols and perform a new pattern matching task on the result For the easy task participants saw a grid of 5 x 5 symbols. They had to find a group of 3 identical symbols which were next to each other in such a way that they could be joined in the game *Two Dots.* At the bottom of the grid is a single example of each type of symbol. Once participants had found the group of 3 symbols they clicked on the same symbol at the bottom of the grid. For the hard task participants saw the same 5 by 5 grid of symbols. They also had to find the same group of 3 identical symbols. Once they found this group, they had to mentally imagine removing these symbols and letting the other symbols drop down as they would in *Two Dots.* After they drop down, they form a new group of 3 identical symbols. Participants then have to click on the symbol which forms this new group. This is explained in more detail in the *Materials* section below. The only difference between the stimuli for each condition is the instructions which participants are given at the start of the trial.

At the start of each trial I measured the mean baseline pupil dilation without any stimulus. This baseline measurement was used to normalise the pupil dilation measurements collected during the rest of the trial. So all pupil dilation analysis is done on "normalised pupil dilation" which is measured pupil dilation divided by the baseline measure for that trial. I also collected data the length of each trial, saccade amplitude and the number of blinks, fixations and saccades. This data was used to get a clearer picture of what was happening during the experiment which would be useful when designing new experiments and reflecting on the results of this one.

Materials

Participants were asked to carry out two different tasks which are developed from previous pupil dilation experiments and the game *Two Dots*. Both tasks involved the same visual stimulus and had the same sequence of stimulus which are listed below.

1. Pause for 2 seconds. This is used to get an initial baseline reading of participants' pupil sizes.
2. Participants hear instructions which tell them which task they are doing
3. Pause for 3 seconds
4. Participants see a grid of 5 by 5 symbols on the screen. There is also 1 symbol of each of the four different types at the bottom of the screen.
5. Participants use the mouse to click on the symbol at the bottom which answers the task they are doing.
6. Pause for 7 seconds. This gave participants' pupils a chance to relax after performing the task.

The tasks were based on tasks commonly performed when playing *Two Dots.* For the easy task participants viewed a grid of symbols like the one below.

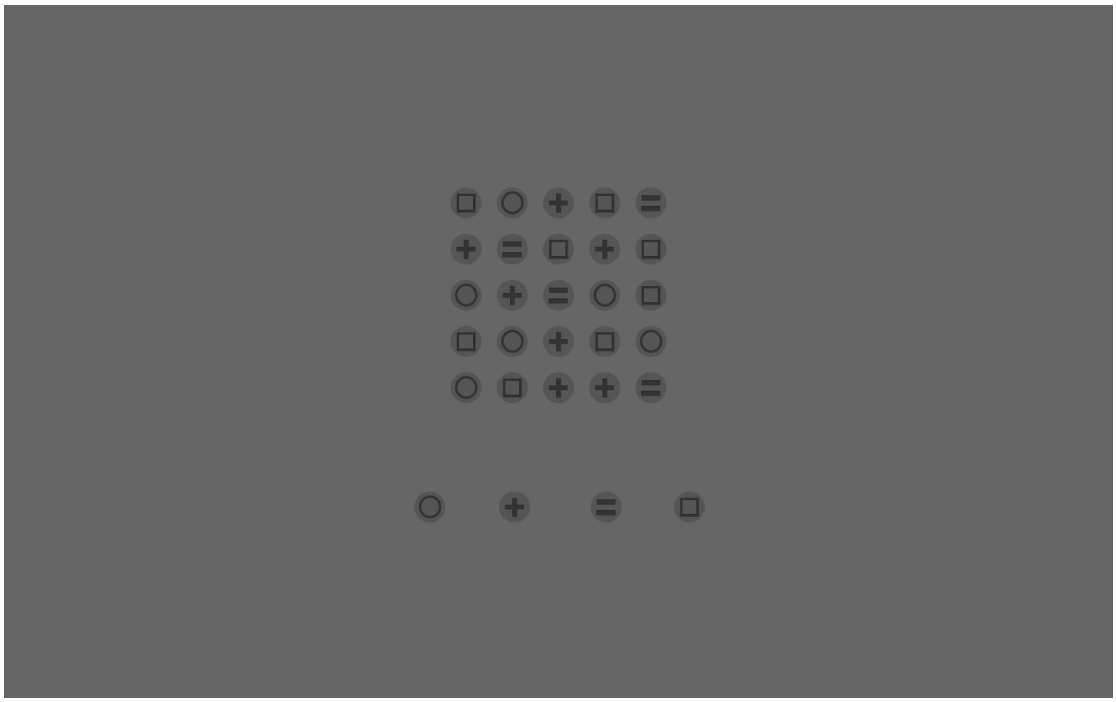

Figure 24 Example stimulus screen for the easy task

They then had to find three identical symbols which were next to each other in such a way that they could be joined when playing *Two Dots*.

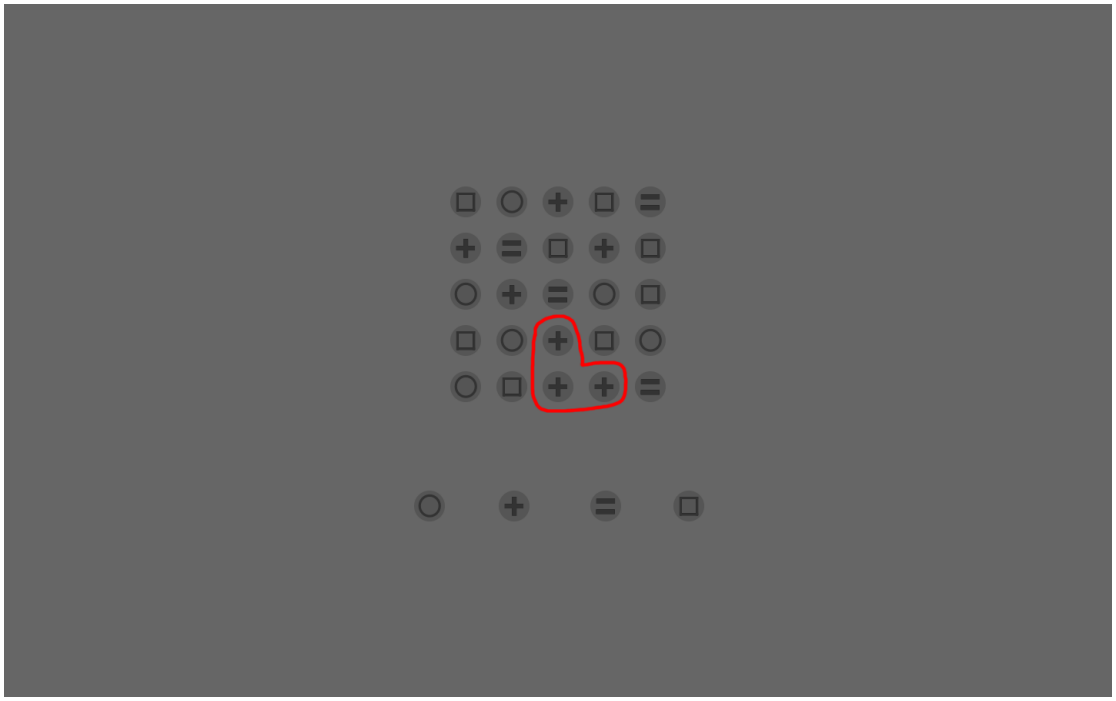

Figure 25 The same stimulus showing which three symbols could be joined when playing Two Dots

In this example there are three crosses together as indicated by the red line. Once participants had found this group, they would indicate their choice by clicking on the cross symbol in the row at the bottom of the screen. For the hard task participants viewed a similar grid of symbols and found a group of three identical symbols in the same way as the easy task.

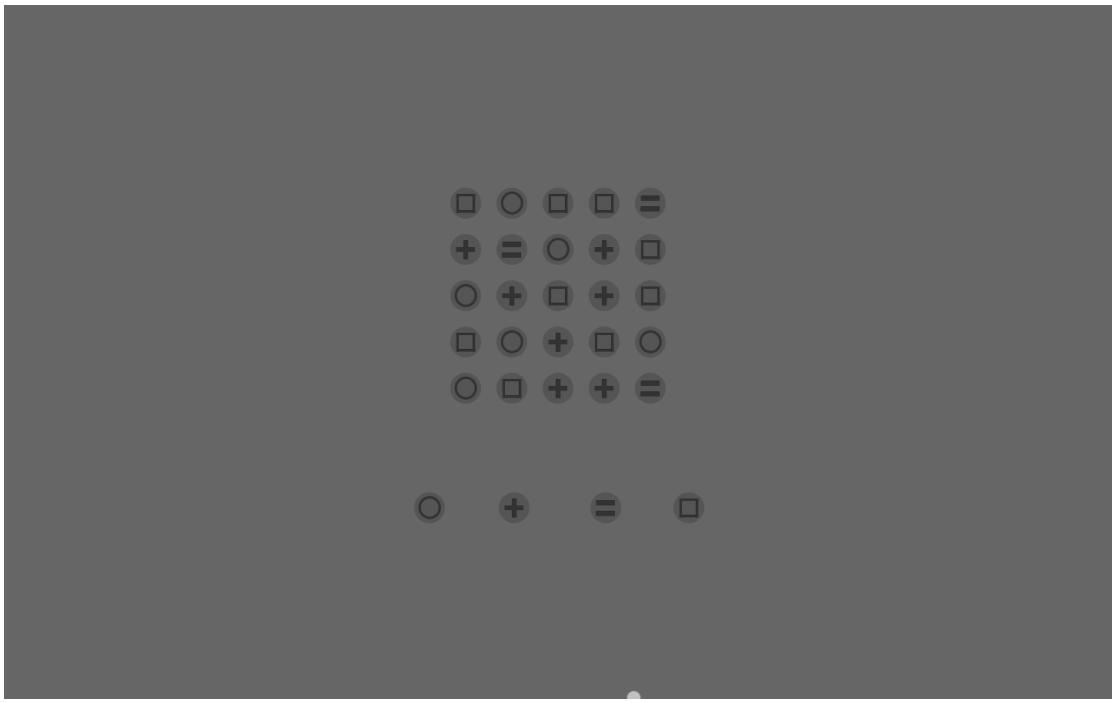

Figure 26 The stimulus for the hard task is identical in form to the stimulus for the easy task

Once they had found this group, they had to mentally “remove” these three symbols and imagine how the other symbols would fall down.

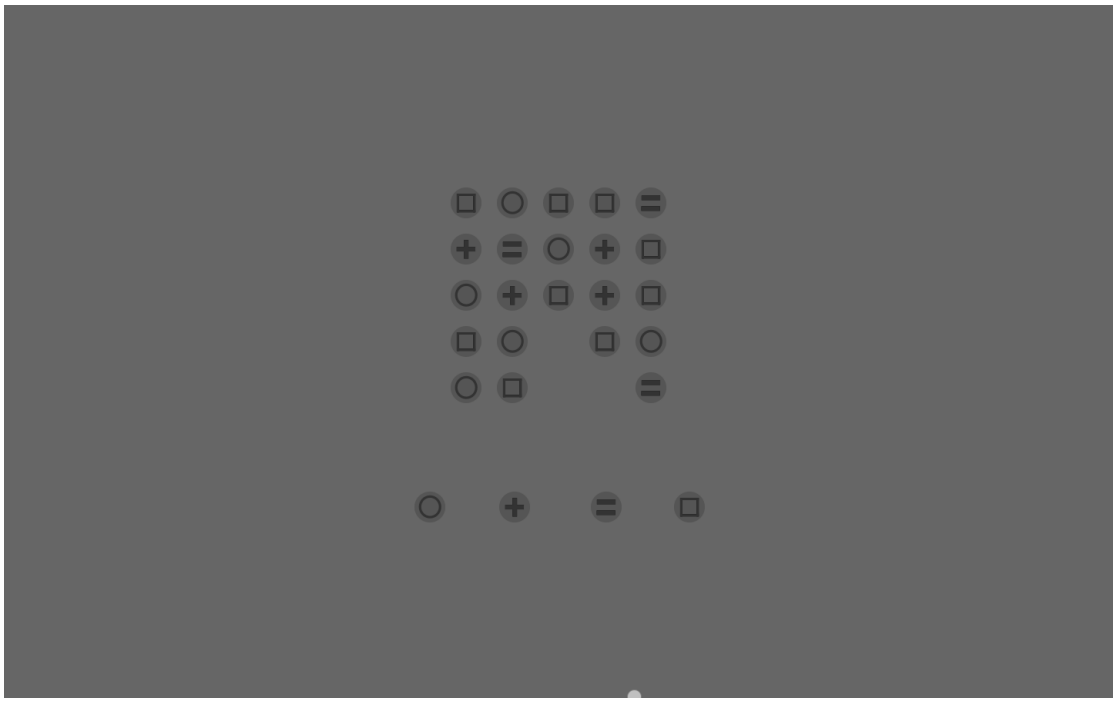

Figure 27 Participants have to imagine removing the group of three symbols

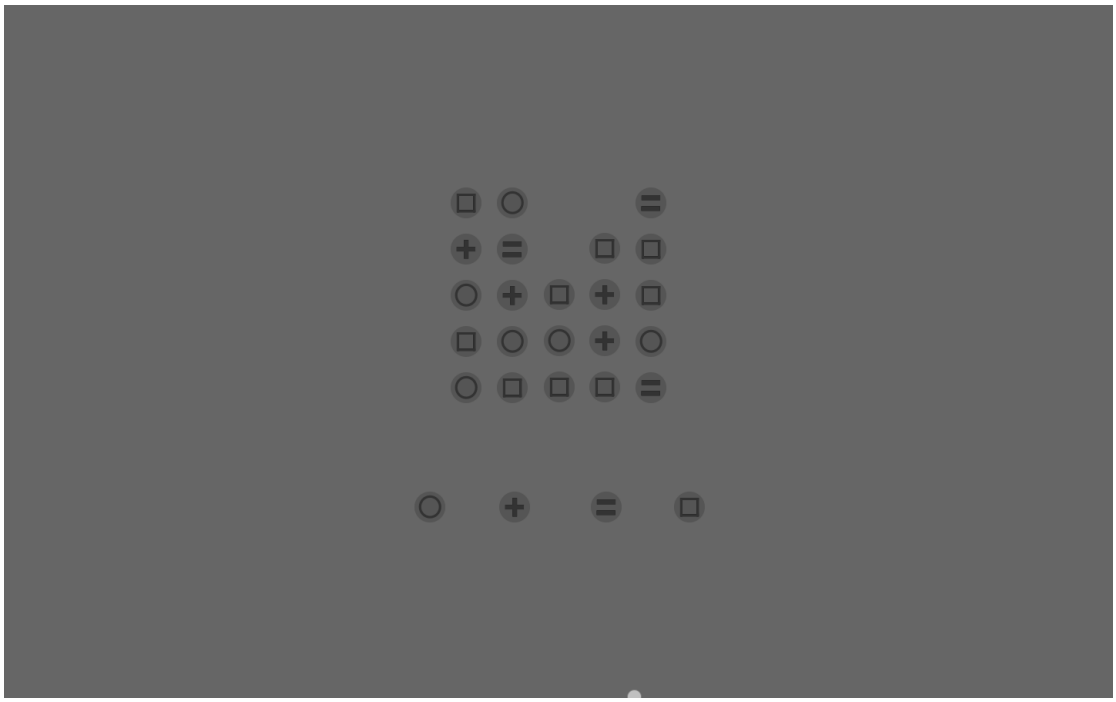

Figure 28 They then imagine the other symbols dropping down

Once they had imagined these symbols dropping down, they then had to find a new group of three identical symbols and click on the appropriate symbol at the bottom of

the screen. In the example below there is a group of three squares so participants would click on the square symbol in the row at the bottom of the screen.

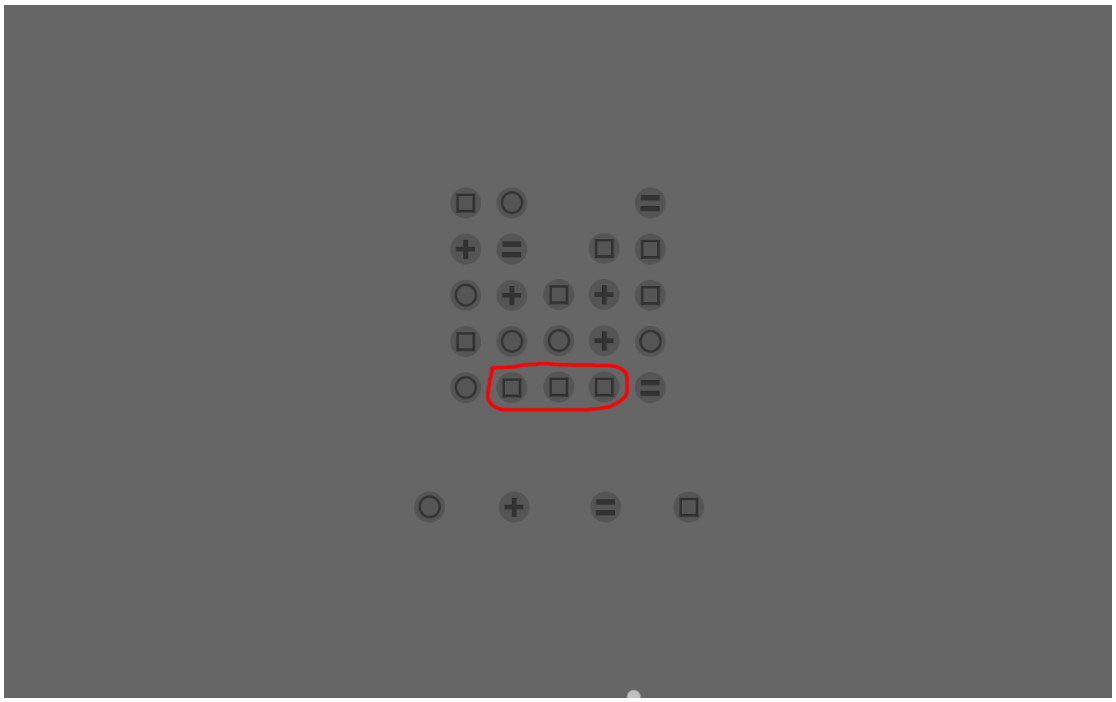

Figure 29 Finally, they find a new group of three in the symbols which have dropped down. They then click on this symbol

It is important to remember that the stimulus for the hard task is static and identical in form to the stimulus for the easy task. The symbols are not actually removed or drop down – participants have to imagine this happening.

Participants

My initial pupil dilation experiments in chapter 4 used a power calculation to estimate that 14 participants would be enough for a significant result. This was indeed the case as these experiments had 14 participants and found significant results with strong effect sizes. I decided to use a similar number of participants for this study. 16 students and university staff took part in the study. However, one participant was wearing very strong glasses (>8 dioptres) which meant that the eye tracker could not follow his pupil movements. Another participant had difficulty understanding the task and may not have been performing it correctly. These two participants were removed from the analysis leaving 14 (4 female). Ages ranged from 20 to 55 (mean=30.4). Seven of the participants were not native speakers of English.

Procedure

Each participant took part in both conditions. To control for possible training or fatigue effects the order of the conditions was counter balanced so that half the participants performed the easy task first followed by the hard task. The other half of participants did it the other way round.

Each participant performed a consent procedure and was calibrated with the eye tracker. The eye tracker then started recording. The participant played the game of *Two Dots* for five minutes. This included a short tutorial so that they understood how to play the game. The participant then came away from the eye tracker and was shown a paper version of the easy task described above. Once the participant understood the task and gave the correct answer, they were shown a paper version of the hard task and I made sure that they understood the task and gave the correct answer. The participant then returned to the eye tracker and was re-calibrated. They had the first task explained to them and they performed five trials of this task. They then had the second task explained to them and performed five trials of that task. These first ten trials are treated as training trials and not used in the analysis or results. After the training, they performed 30 trials without stopping. These trials interleaved the first and second tasks in groups of five. So, participants would perform five trials of the first task followed immediately by five of the second task. This was repeated three times until they had completed all 30 trials. Before each trial, they heard audio instructions which told them which task they needed to perform.

Pilot

Before running the main study, I performed a pilot study on two additional participants. I then analysed these pupil dilation measurements to find the period of maximum difference between conditions. I analysed the pupil dilation relative to the time that participants clicked to give their answer. The maximum difference between conditions turned out to be the 1 second period which started 3 seconds before participants clicked to give their answer. These pilot participants were not counted towards the main experiment.

5.4. Results

As with previous experiments (see chapter 4) all pupil diameter is normalised against the baseline measure. Unless otherwise stated the number of participants (N) is 14.

5.4.1. Hypothesis

The hypothesis proposed that there would be an increase in pupil dilation between the easy and hard conditions when measured in a one second window starting 3 seconds before participants click their answer. 19.8% of the trials lasted less than 3 seconds so I removed all the trials which took less than 3 seconds. This entailed removing 75 trials from the easy condition (Mean 5.36 per participant) and 11 trials from the hard condition (0.79 per participant).

There was a significant difference in the mean pupil dilation for across all participants during this window; $F(1,13)=80.423$, $p<0.001$. There was also an extremely large effect size ($\eta_p^2=0.861$).

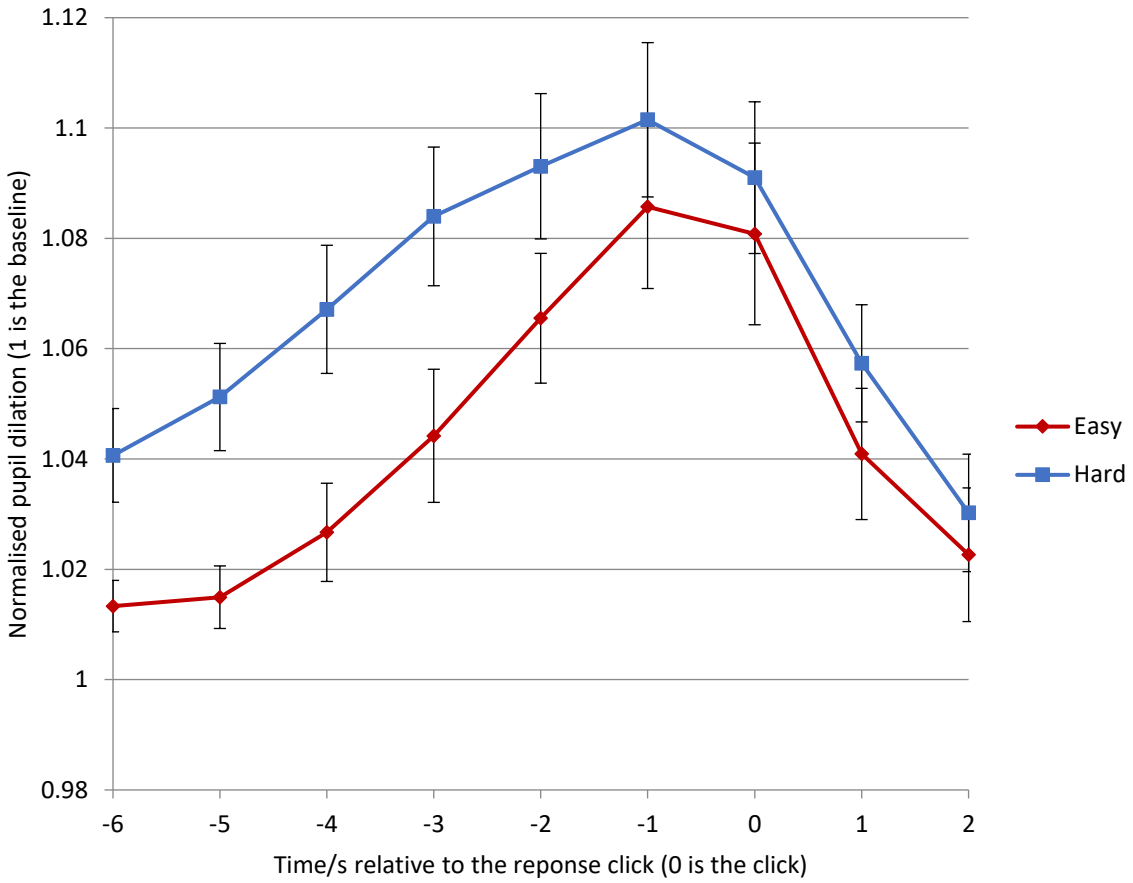

Figure 30 Normalised pupil dilation over the time relative to the time the participant clicks (Error bars show standard error). Data has all trials under 3 seconds removed.

	Hard	Easy	F(1,13)	p value	η_p^2
Time /s	Mean (SD)	Mean (SD)			
-6	1.041 (0.032)	1.013 (0.017)	21.614	<0.001	0.624
-5	1.051 (0.036)	1.015 (0.021)	36.053	<0.001	0.735
-4	1.067 (0.043)	1.027 (0.033)	70.422	<0.001	0.844
-3	**1.084 (0.047)**	**1.044 (0.045)**	**80.423**	**<0.001**	**0.861**
-2	1.093 (0.049)	1.066 (0.044)	32.928	<0.001	0.717
-1	1.101 (0.052)	1.086 (0.055)	9.984	0.008	0.434
0	1.091 (0.051)	1.081 (0.062)	2.309	0.153	0.151
1	1.057 (0.040)	1.041 (0.044)	7.390	0.018	0.362
2	1.030 (0.040)	1.023 (0.045)	1.746	0.209	0.118

Table 23 Normalised pupil dilation over the time relative to the time the participant clicks. Data has all trials under 3 seconds removed.

5.4.2. Time to respond

I looked at the length of time that each participant took to respond to each trial for each condition (See Table 24). There is a significant difference between the length of time that

participants took to respond to each condition. F(1,13) = 35.434, p<0.001. There is a very large effect size (η_p^2=0.732) between conditions.

Condition	Mean/ms	SD/ms
Easy	4701	1317
Hard	9560	3656

Table 24 Time to respond to each condition in milliseconds.

I investigated whether the time that it took participants to complete each trial varied over the time of the experiment. To do this I calculated correlations between the duration of the trial and the trial number for both conditions (See Table 25). There was a significant correlation for both conditions although both correlations were very small.

Condition	Correlation with trial number	p	t	df
Easy	-0.136	0.0233	-2.281	274
Hard	0.150	0.0138	2.477	265

Table 25 Correlations between the duration of the trial and the trial number

5.4.3. Fixations, saccades, saccade amplitude and blinks

I looked at the differences between conditions for the following factors: blinks per second, saccades per second and mean saccade amplitude. There was a significant difference between all of these factors. Effect sizes ranged from large to extremely large. See Table 26.

		Fixations/s	**Saccades/s**	**Mean Saccade Amplitude**	**Blinks/s**
Mean (Standard deviation)	Easy	3.839 (0.493)	3.595 (0.485)	2.011 (0.179)	0.226 (0.152)
	Hard	3.486 (0.467)	3.351 (0.452)	1.642 (0.206)	0.173 (0.158)
	F(1,13)	32.868	18.297	84.779	5.725
	p	<0.001	0.001	<0.001	0.033
	η_p^2	0.717	0.585	0.867	0.306

Table 26 A comparison between conditions of fixations/s, blinks/s, saccades/s and mean saccade amplitude.

5.4.4. Baseline

The baseline pupil dilation for each trial is the mean pupil dilation for the initial two second pause time when participants are doing nothing except looking at a fixation cross.

There is not a significant difference in the baseline pupil dilation between the easy and hard condition. F(1,13)= 3.976, p=0.068. There is a moderate effect size between the two conditions (η_p^2= 0.234).

Condition	Mean	SD
Easy	5076	555
Hard	5147	521

Table 27 Pupil size during the baseline period between two conditions

I was interested to see if the baseline changed over time so I calculated correlations between the baseline and the trial number

Condition	Correlation with trial number	p	t	df
Easy	-0.428	0.0183	-2.506	28
Hard	-0.0168	0.9298	-0.089	28

Table 28 Correlations between the baseline measurement and the trial number

This shows that there was significant moderate correlation between baseline pupil dilation and the trial number for the easy condition. This indicates that the baseline reading decreased as participants progressed through the experiment. However, the hard condition had a smaller correlation which was not significant.

5.5. Discussion

The hypothesis of the experiment proposed that the hard condition would have higher pupil dilation than the easy condition. Specifically, for each condition this compared the mean pupil dilation for a one second window starting at 3 seconds before participants clicked on their answer. This hypothesis was supported with an extremely large effect size.

There is a large difference in the length of time that participants take to click on the answer in the easy and hard conditions. This is significant with a strong effect size. This is not surprising as the hard task needs participants to perform the same process as the easy task (find three dots together) plus an additional task (remove those dots and let others drop down). To compare the pupil dilation between tasks I wanted to look at the moment when participants were trying their hardest at each task. A pilot study found that 3 seconds before participants click is the time of maximum difference between the conditions. This was confirmed by the main experiment. It is difficult to look at the task being done and predict exactly why it happens then although I did come up with a post-hoc explanation which is described below in 5.5.1 Limitations. The pupil dilation during the hard condition is consistently higher than that in the easy condition. In the hard condition, it starts much higher and rises gradually until the answer is clicked. It then decreases rapidly. In the easy condition pupil dilation starts off much lower but increases sharply until the answer is clicked. After the answer is clicked it decreases

rapidly in the same way as the hard condition. For both conditions, the maximum pupil dilation happens around 1 second before the answer is clicked. This is likely to be because participants work out the answer first and it then takes them about a second to move the mouse to click on the correct symbol.

Given that there is such a large difference in the duration of the two conditions I looked to see if there were any patterns in the amounts of time that participants took to complete the task. In particular I looked to see if participants became faster or slower at the task as the experiment progressed. I found that there was a significant correlation between the trial duration and the trial number for both easy and hard conditions. Participants in the easy condition got slightly faster at performing the task over the time of the experiment and participants in the hard condition got slightly slower. However, these correlations are both very small so it is unlikely that fatigue or training effects had a significant effect on the final analysis. Future experiments which might increase the number of trials should monitor the results carefully to ensure that these training and fatigue effects have not become stronger.

Before each trial is a two second "baseline measurement" period. During this period participants look at a fixation cross and do not perform any other task. The mean pupil dilation during this time is used as a baseline to normalise the pupil dilation measurements. I also compared the raw baseline pupil dilation between conditions. I found there was not a significant difference but there was a moderate effect size between conditions. In the first audio-based pupil dilation study I found a significant difference in the baseline between conditions. This task had to completed within a time limit so the difference in baseline may have been due to the additional stress of doing the hard task within a time limit. Although the difference for this experiment is not significant it may still be that participants feel more stressed by the hard task which increases the baseline measure. Whatever the cause, any difference in baseline measurement is factored out from the rest of the analysis by normalising the pupil dilation data against this baseline before analysis.

I measured a range of eye movement measures to get a better picture of participants behaviour during the experiment. I looked at the difference in fixations per second, blinks per second, saccades per second and mean saccade amplitude between conditions. All of these factors were significantly higher in the easy condition than the hard condition. This suggests that participants moved their eyes faster and fixated in more different places in the easy condition than the hard condition. This may be because the easy condition is solely a visual search task so participants spend the whole task searching the grid. The hard task also has a manipulation component which requires fewer eye movements during one segment of the task time. This eye movement data is consistent with the visual nature of task and suggests that participants were performing the task as expected.

In conclusion this experiment found a significant difference in pupil dilation due to the differences in cognitive load used by each condition. This was not substantially affected

by confounds due to other factors which can modify pupil dilation such as changes in emotions, light levels or motor actions. There was an extremely strong effect size with a small number of participants. This suggests that pupil dilation could be an effective measure of changes in cognitive load over the time of a game which may then be a measure of the experience of playing that game.

5.5.1. Limitations

The main limitations of this experiment are the lack of similarity to playing a real game and that it is difficult to say why pupil dilation differences peak when they do. The purpose of this experiment was to investigate whether cognitive load could be measured during a "game-like" task. Although this has been achieved, there are some substantial differences between this task and actually playing a real game. It is very difficult to say how typical the hard task is during a game; players may mentally think ahead before making a move or they may just make moves without much thought or using other heuristics. The experimental task response involves making just one click once participants have come up with the answer. In the real game participants join dots in a more complex response and may be making decisions while they are joining the dots. The real game has more complex goals which may affect players' mental state. Moves happen quickly, one after another, without long pauses in between. These differences would not only affect players' experience of the game but also create new issues with measuring mental state. Although this experiment shows that participants use significantly more cognitive load for a particular hard task which may occur in the game it does not show how often that type of task is actually performed by players of the game.

Both the pilot study and the main study showed that the point of maximum pupil dilation difference between conditions happens 3 seconds before participants click their answer. It is difficult to look at the task being done and predict exactly why it happens then. However, it is possible to examine the data and come up with a possible post-hoc explanation. The pupil dilation for the hard task is consistently higher than the easy task and increases at a steady rate. The easy task pupil dilation starts very low and then increases sharply, so it may be that participants find the initial searching activity is low effort and then they make more of an effort at the end of the task when they need to be sure they have found the right answer. Similarly, the hard task starts with a visual search for a group of 3 dots. Participants then have to mentally remove those dots and then perform a new visual search for the new row of dots. It is possible that the most cognitively demanding part is mentally removing the dots. After this mental removal there is another visual search before participants choose their answer. The need to perform this final visual search may explain why there is a 3 second gap between the peak of pupil dilation and participants clicking to give their answer.

5.6. Exploratory analysis of game pupil dilation

During the training period of this experiment participants played *Two Dots* for five minutes while their pupil data was recorded by the eye tracker. I used this pupil data to carry out exploratory analysis to investigate how different gameplay experiences affected pupil dilation. The main analysis of this experiment showed that the period of maximum difference in pupil dilation was 3 seconds before participants start their move. However, in a real game, participants do not usually think for 3 seconds before making their move, so I needed to look at a different time window. The main analysis also showed a very strong effect (η_p^2= 0.434) between conditions at during the time window which started 1 second before participants start their move. It is much more likely that participants think for at least 1 second before making their move so for this analysis I considered this time window. Similarly, the main analysis has a pause between each trial and after this pause a baseline measurement of pupil dilation is taken for 2 seconds. A real game does not have pauses between moves so I took the baseline measurement as the first 0.5s of each move. Previous pupil dilation experiments (See pd chapter experiment 2) had shown the pupil dilation does not start to change until at least 0.5 seconds after the participants are shown the stimulus. So this baseline is taken when participants have seen the new arrangement of dots but have not yet had a chance to consider their move

If players fail a game level, then they may think harder about the level the second time they try it. I compared the first trial of a level after a success with the first trial after a failure. If players are trying harder after a failure then you would expect the increase in cognitive load to increase the pupil dilation for the trial after a failure. There was no significant difference between these two conditions (Success M=1.018, SD=0.014, Failure M=1.000 SD=0.017, F(1,9) = 4.602,p=0.061), however there was a strong effect size (η_p^2= 0.338). The number of participants in this exploration is small so the lack of significance may not mean that there is no effect. The large effect size makes it more likely that there is a difference, although once again the small sample size means this effect size may not be robust. Initially this looked like it could be a promising sign of differences in cognitive load due to the player experience. However, closer analysis showed that the pupil dilation was stronger after a success than a failure. If this pupil dilation was due to cognitive load then it would indicate that participants put more cognitive effort in after succeeding than failing. This seems unlikely as I expected that players would try harder after failing than after a success. It may be that this difference in pupil dilation is due to positive emotion caused by succeeding at the level. I also considered that players may try harder at the beginning of a level than at the end of a level but found no significant difference in pupil dilation (Success M=1.008, SD=0.014, Failure M=0.998, SD=0.026, F(1,10) = 1.145, p = 0.310, η_p^2= 0.103).

In the final move of a level players generally know whether they will complete the level or fail. I considered that this could lead to changes in pupil dilation, either due to differences in cognitive load or differences in emotion. I looked at the final trial in each level and compared those which led to success with those which led to failure. There was no significant difference between these trials ($F(1,10)= 2.968, p=0.116, \eta_p^2= 0.229$) although the effect size is large. Once again, the trials which led to success (M=1.008, SD=0.018) have a higher pupil dilation than those which let to failure (M=0.982 SD=0.049). Usually when players succeed at a level the final move in the level is fairly straightforward – they have almost hit their target and do not need a special move to get all the way. However, when players think they are going to fail a level they may be more likely to look extra hard for a special move which joins more dots and might lead to success. Because the measured pupil dilation is higher in those trials which led to success this suggests that the pupil dilation is unlikely to be due to cognitive load. As with the investigation into the first move of a level (above) it may be more likely that this increase in pupil dilation is created positive emotion due to succeeding at the level.

In conclusion, this exploratory analysis of pupil dilation found no significant evidence that pupil dilation changes depending on the amount of cognitive load used to play the game. It may be possible that pupil dilation increases due to the emotional reaction of the player to successfully completing a level. To investigate further how game experience can change cognitive load I investigated the effect on pupil dilation of playing different variations on of the same game.

5.7. Experiment Pupil 4: Three variants of *Two Dots*

The previous game puzzle experiment showed a significant difference in pupil dilation between easy and hard tasks taken from the game *Two Dots*. This suggested that pupil dilation could measure changes in cognitive load during game play. I then performed an exploratory analysis on a full game of *Two Dots* looking for links between pupil dilation and the game experience. This found no significant links with cognitive load although there may be some links to emotions that players experience during the game. The reason for this may be that players' experience of the game did not change that much throughout the game so there were no large variations of cognitive load (or experience) to measure.

To make a more thorough investigation into how cognitive load changes with game experience I decided to compare different games which had consistently different experiences (and cognitive load) throughout the duration of the game. To do this I performed an experiment which compared the pupil dilation between three variants of the game of *Two Dots*. One variant was a straight clone of the game of *Two Dots.* The other two were designed to be similar to the full game of *Two Dots* in that the visual display and the actions that players make during the game (i.e. joining dots) are the same but they create a different game experience for the player. These variants are

described in more detail in chapter 3. They are known as the *Full game,* the *No goals* game and the *All dots the same* game. The *No goals* game and the *All dots the same* game both remove gameplay features which make the game engaging so I expected them to have lower player engagement than the *Full game.* In particular the *All dots the same* game is very easy to play because it has no puzzle elements, players can easily find dots of the same symbol to join because all the dots are the same. This is unlike the *Full game* which contains puzzle elements which need to be solved. The *No goals* game does contain similar puzzle elements to the *Full game* but players are not set any targets to meet so they may put less effort into the game and so use less cognitive load. Although the design of the three games suggests that players will have different gameplay experiences, I wanted to confirm that this was the case so I asked players to complete an Immersion Experience Questionnaire (IEQ) Jennett et al. (2008) after playing the game. The reasons for choosing the IEQ are discussed in chapter 3. This experiment was also used for a separate analysis to investigate measuring game attention which is described in chapter 6. This required that in all versions of the game the edges of the screen contained icons from the *Webdings* symbol font. Eye tracking showed that participants were focused on the central area of the screen for a mean of 97.9% (SD = 3.46%) of the trial so it is unlikely that this made any difference to pupil dilation differences between the different games.

Hypothesis

For this experiment I wanted to compare the pupil dilation between the different versions of the game. The previous game puzzle experiment showed that the greatest difference in pupil dilation was at a window starting 3 seconds before participants made a move. However, in a real game of *Two Dots* players do not usually think for 3 seconds before making their move so I decided to compare pupil dilation at the period 1 second before players start to make their move. This gives the following main hypothesis for the experiment.

For a 1 second window starting at 1 seconds before participants press down the mouse to make a move the size of participants' pupils will increase when they are playing the more challenging Full game compared to the other game versions.

The games were designed to give players different experiences. To confirm that this was the case participants filled in an IEQ after playing the game. This led to the secondary hypothesis of the experiment:

There will be a significant difference between immersion scores between the different variants of the game.

5.7.1. Design

This was a between-subjects design with three conditions. The independent variable was the variant of the game that each participant played. The dependent variable is the amount of pupil dilation in each condition. There were three different variations of the game and each participant played one of them. The three variations of the game were:

1. *Full game*, this is a reconstruction of the first ten levels of the game *Two Dots.*
2. *No goals*, this is the same as the full game except that there are no targets to reach or move limits.

3. *All Dots the same*, this is the same as the full game except all the dots are the same symbol.

These are described in more detail in chapter 3. Whilst playing the game participants were recorded by an Eyelink eye tracker as in previous experiments (also described in chapter 3). After playing the game for five minutes players filled in an Immersion Experience Questionnaire (Jennett et al., 2008). In the previous pupil dilation studies all pupil dilation data was normalised against a baseline measure. For this experiment the baseline was taken as the first 0.5 seconds after a new grid of dots had been displayed. Previous pupil dilation studies also had a training period which allowed participants to get used to the task. In this experiment the first ten moves were counted as a training period and removed from the analysis.

I also collected and analysed data which was not linked to the main hypotheses. The exploratory analysis (see section 5.6) showed a strong effect size between the pupil dilation after players had succeeded at a level compared to when they had just failed. To investigate whether this was a robust effect I performed a similar analysis on the pupil data from the *Full game.* To get a fuller picture of participants' experience I also collected and analysed the baseline pupil dilation at the beginning of each move, the saccade amplitude and the number of blinks, saccades and fixations.

Participants

The previous pupil dilation experiment was a within-participants study with two conditions and 14 participants. This experiment was a between participants study with three conditions. I decided to use a similar number of participants per condition so needed a larger sample. 50 students and staff from the University of York took part in the study. However, one was rejected because her high strength glasses prevented the eye tracker from following her eyes. Another was rejected because he was confused about how to play the game and just stared at the screen for several minutes without making any actions. This left 48 (24 female) participants who were considered for the trial. 16 participants performed each condition. Ages ranged from 18 to 62 (mean = 22.3). 10 of the participants were not native speakers of English.

5.8. Procedure

Each participant played one variant of the game. Participants did a consent procedure and were calibrated with the eye tracker. They then played the game for 5 minutes before filling out an Immersion Experience Questionnaire about their game experience.

5.9. Results

Pupil dilation

The main hypothesis of the experiment was: *For a 1 second window starting at 1 second before participants press down the mouse to make a move the size of participants' pupils will*

increase when they are playing the more challenging full game compared with the other game versions.

An ANOVA (F(2,45) = 1.877 p =0.165) showed that there was no significant difference between pupil dilation at this time period (also see Figure 31). The effect size between conditions was small (η_p^2=0.077).

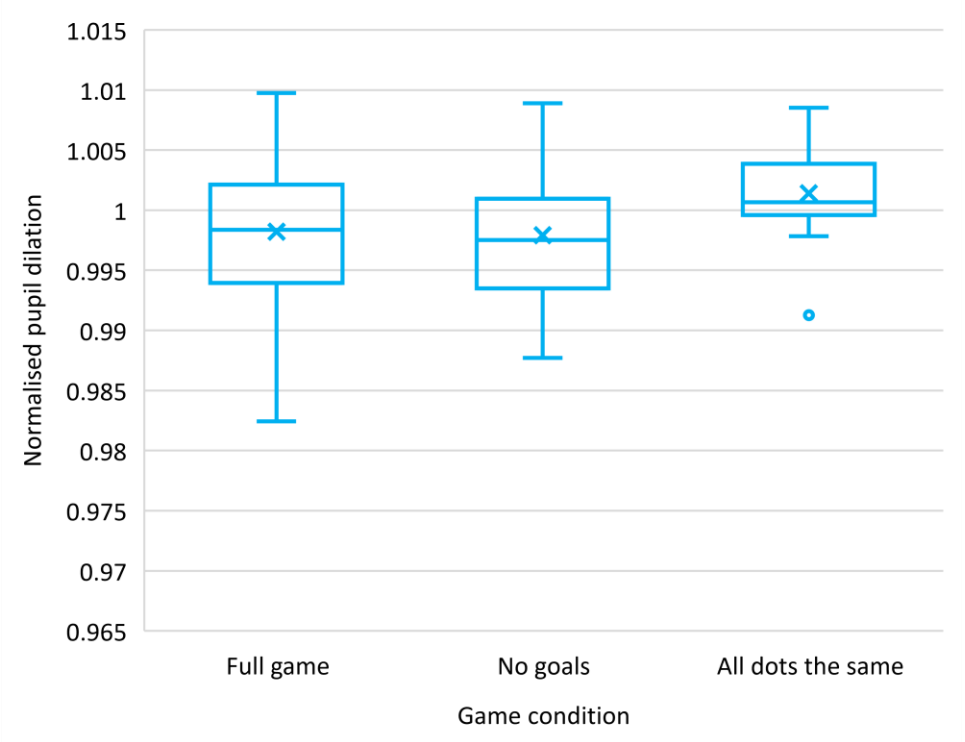

Figure 31 Boxplot of normalised pupil dilation one second before participants start to make their move

I plotted the mean pupil dilation across the time of each move relative to the point when participants first press the mouse button down. This is shown in Figure 32 and the data is shown in Table 29.

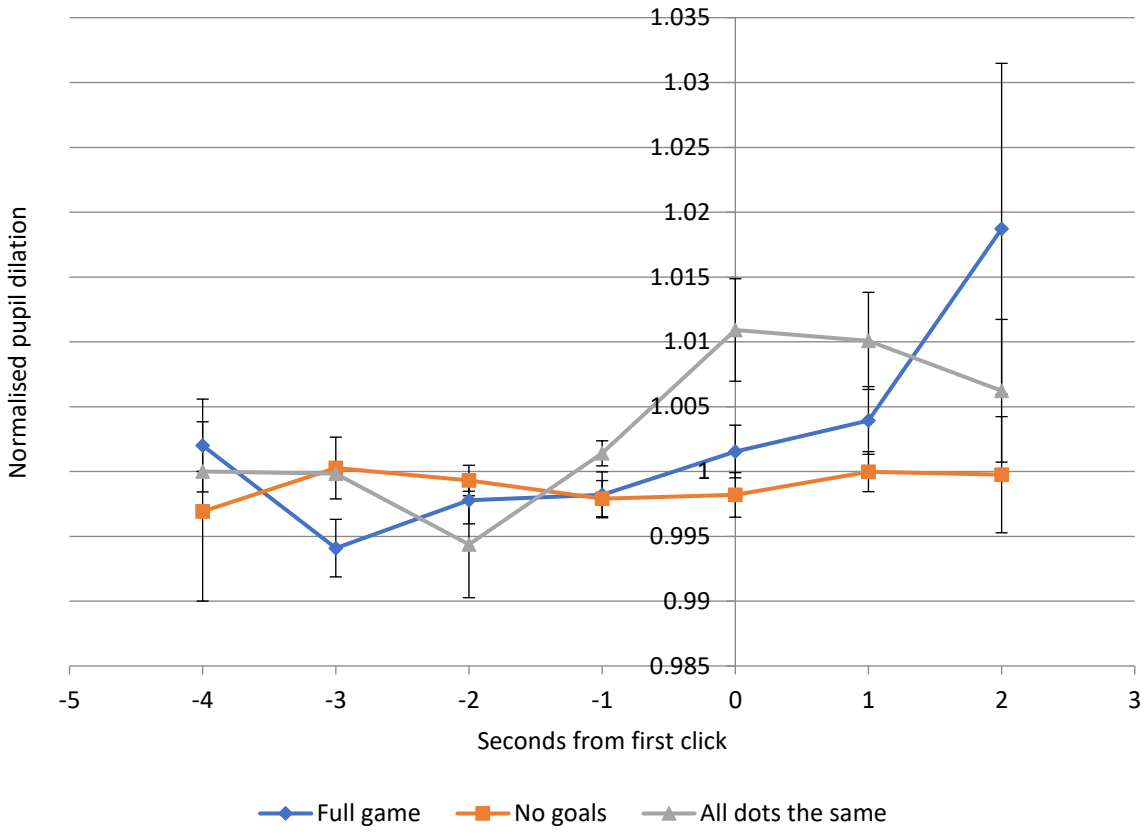

Figure 32 Mean pupil dilation across the time of each move. Error bars show standard error

	Full Game	**No goals**	**All dots the same**	**F(2,45)**	**p value**	η_p^2
Time /s	Mean (SD)	Mean (SD)	Mean (SD)			
-4	1.002 (0.014)	0.997 (0.027)	N/A	0.426	0.519	0.015
-3	0.994 (0.009)	1.000 (0.010)	1.000 (0.000)	1.939	0.161	0.111
-2	0.998 (0.007)	0.999 (0.005)	0.994 (0.015)	1.001	0.376	0.046
-1	**0.998 (0.007)**	**0.998 (0.006)**	**1.001 (0.004)**	**1.877**	**0.165**	**0.077**
0	**1.002 (0.008)**	**0.998 (0.007)**	**1.011 (0.016)**	**5.741**	**0.006**	**0.203**
1	1.004 (0.010)	1.000 (0.006)	1.010 (0.015)	3.330	0.045	0.129
2	1.019 (0.048)	1.000 (0.017)	1.006 (0.022)	1.320	0.278	0.060

Table 29 Comparison of pupil dilation between conditions over time. N/A is used when there were no observations to calculate at that time period.

The graph shows a large difference between conditions at 0 seconds which is the period immediately after the mouse button has been pressed. An ANOVA ($F(2,45) = 8.959$, $p < 0.001$) showed that this difference was significant with a large effect size (η_p^2=0.285). A Tukey's HSD test (see Table 30) showed a significant different between the *All dots the same* game and both the *Full game* (p=0.020) and the *No goals* game (p=0.002). There was no significant difference between the *No goals* game and the *All dots the same* game.

	Full game	No goals
No goals	p=0.394	
All dots the same	p=0.020	p=0.002

Table 30 Tukey's HSD test of pupil dilation at the 0 second time period

Immersion Experience Questionnaire (IEQ) and subfactors

The purpose of the IEQ was to confirm that participants in each condition had had different levels of engagement in the game. There was a significant difference in the immersion scores between the *Full game* (M=107.13, SD=17.00), the *No Goals* game (M= 93.38, SD= 14.05) and the *All dots the same* game (M=93.56, SD=13.38) conditions; F(2,45)= 4.492,p=0.017. There was a moderate effect size between conditions (η_p^2= 0.166). I performed a Tukey's HSD post-hoc test to investigate which conditions were significantly different from each other (See Table 31). This showed that there was significant difference between the *Full game* and the *No goals* game. There was also a significant difference between the *Full game* and the *All dots the same* game. There was no significant difference between the *No goals* and *All dots the same* games.

	Full game	No goals
No goals	p=0.012	
All dots the same	p=0.013	p=0.972

Table 31 Tukey's HSD on immersion score

I looked at the scores between conditions for the subfactors of the IEQ. These are *Cognitive involvement*, *Emotional involvement*, *Real world dissociation*, *Challenge* and *Control* (See Table 32). This showed significant differences in *Emotional Involvement* and *Challenge* but none of the other subfactors. I performed Tukey's HSD post-hoc tests on these subfactors (See Table 33 and Table 34) which showed that the *Full game* was significantly different from the other two variants.

	Mean (Standard deviation)					
	Full game	**No goals**	**All dots the same**	**F(3,45)**	**p**	**Effect size (η_p^2)**
Cognitive involvement	33.94 (5.77)	32.06 (5.54)	29.69 (4.69)	2.533	0.091	0.101
Emotional involvement	17.88 (5.43)	13.81 (4.10)	13.94 (3.61)	4.322	0.019	0.161
Real world dissociation	32.38 (5.54)	28.75 (5.62)	31.13 (5.38)	1.786	0.179	0.074
Challenge	14.38 (1.5)	11.06 (2.44)	10.81 (3.19)	10.349	<0.001	0.315
Control	16.06 (3.77)	14.56 (3.10)	15.25 (2.46)	0.906	0.411	0.039
Immersion	107.13 (16.97)	93.38 (14.05)	93.56 (13.38)	4.492	0.017	0.166

Table 32 Comparison of the IEQ subfactors between conditions

	Full game	No goals
No goals	p=0.013	
All dots the same	p=0.016	p=0.937

Table 33 Tukey's HSD of the *Emotional Involvement* subfactor

	Full game	No goals
No goals	p=0.001	
All dots the same	p=0.001	p=0.956

Table 34 Tukey's HSD of the *Challenge* subfactor

Gameplay differences

Each trial has a *Thinking time* which is the time before participants push down the mouse button and a *Move time* which is the amount of time that they take to make their move. There is a significant difference in both *Thinking time* and *Move time* between conditions (See Table 35).

	Full game	**No goals**	**All dots the same**	**F(2,45)**	**p value**	η_p^2
	Mean (SD)	Mean (SD)	Mean (SD)			
Thinking time /ms	1926 (960)	1277 (590)	686 (259)	15.506	<0.001	0.378
Move time /ms	1143 (369)	733 (241)	4769 (2054)	60.357	<0.001	0.703

Table 35 Comparison of thinking time and move time between conditions

A Tukey's HSD showed that all three conditions have a significantly different *Thinking time* from each other (See Table 36). Another Tukey's HSD shows that the *All dots the same* game has a significantly higher *Move Time* than the other two conditions (see Table 37).

	Full game	No goals
No goals	p=0.001	
All dots the same	p<0.001	p=0.016

Table 36 Tukey's HSD of *Thinking time* between conditions

	Full game	No goals
No goals	p=0.975	
All dots the same	p<0.001	p<0.001

Table 37 Tukey's HSD of *Move time* between conditions

Comparing success with failure

In the previous experiment an exploratory analysis showed a strong effect between pupil dilation for the trial after successfully completing a level and pupil dilation for the trial after failing to complete a level. To see if this was a robust effect, I compared the same game events for the *Full game* condition. There was not a significant difference but there was a medium effect size (F(1,10)= 1.578, p=0.238,η_p^2=0.138). The pupil dilation for the trial after a success (M=1.004 SD=0.020) was higher than the pupil dilation after a failure (M=0.985 SD=0.041).

Fixations, saccades, saccade amplitude and blinks

I looked at the differences between conditions for the following factors: blinks per second, saccades per second and mean saccade amplitude (See Table 38).

	Mean (Standard deviation)					
	Full game	**No goals**	**All dots the same**	**F(3,45)**	**p**	**Effect size (η_p^2)**
Fixations/s	2.741 (0.338)	2.812 (0.305)	2.450 (0.341)	6.001	<0.001	0.211
Saccades/s	2.336 (0.377)	2.407 (0.335)	2.23 (0.368)	0.967	0.388	0.041
Mean Saccade Amplitude	2.204 (0.435)	2.036 (0.284)	2.267 (0.395)	1.602	0.213	0.066
Blinks/s	0.144 (0.102)	0.121 (0.090)	0.176 (0.082)	1.438	0.248	0.060

Table 38 Comparison of eye movements between conditions

There was a significant difference in fixations per second. I performed a post-hoc test to investigate this more fully (See Table 39). The *All dots the same* game has a significantly different number of fixations from both the other games.

	Full game	No goals
No goals	p=0.820	
All dots the same	p=0.034	p=0.006

Table 39 Tukey's HSD of fixations per second

5.10. Discussion

The main hypothesis of the experiment was that there would be a significant difference in pupil dilation at the period one second before participants start their move. This hypothesis was not supported and the effect size between conditions was small (η_p^2= 0.077). I looked at the pupil dilation across the whole time period of the move and found the biggest difference was at the zero second time period – the actual moment of clicking. This difference is significant with a moderate effect size (η_p^2=0.285). A post-hoc test showed that this difference is caused by the pupil dilation for the *All dots the same*

game being much higher than the other two. This was unexpected because, unlike the other two game variants, this game has no puzzle elements so I expected that it would need less cognitive load and have lower pupil dilation. One possible explanation is that for this game participants often tried to join all 25 dots in the grid in one go to make things more interesting. This is illustrated by the large differences in the amount of time it took players to make their moves which is discussed below. It is possible that this required more concentration for the actual act of making the move which then had an effect on pupil dilation.

This experiment used three different variants on the same game. This was to ensure that participants had different game experiences while performing similar tasks. To confirm that this was the case participants completed an IEQ after playing the game to measure their experience. The secondary hypothesis of the experiment was that there would be a significant difference between immersion scores for the three conditions. This hypothesis was supported with a moderate effect size between conditions (η_p^2= 0.166). A post-hoc test showed that there was a significant difference between immersion in the *Full game* and the other two games. There was no significant difference between the *All dots the same* game and the *No goals* game.

The IEQ questionnaire has five subfactors; *Cognitive involvement, Emotional involvement, Real world dissociation, Challenge* and *Control.* Only two of these factors, *Emotional involvement* and *Challenge* were significantly different between games. What is surprising is that there is not a significant difference in *Cognitive involvement*. *Two Dots* is a puzzle game so you would expect that changing the game to significantly alter the *Challenge* and *Emotional involvement* would also involve a change of *Cognitive involvement* but this is not the case. One possible reason is that cognitive involvement is not a key part of the experience of playing *Two Dots.*

Exploratory analysis of pupil data from the previous experiment (see section 5.6) showed a strong effect size in the amount of pupil dilation at different events in the game. Specifically, this meant that in the *Full game* there was a strong effect between those moves which followed a successful level completion and those which followed a failed level completion. This difference was not significant in the exploratory analysis but was worth investigating in this experiment to see if it was a consistent difference. Once again pupil dilation after a success was higher than following a failure but this difference was not significant and the effect size was moderate.

I analysed participants' eye movements and blinks to see if they reflected their experience of playing the game. There were significant differences between conditions in the number of fixations per second. A post-hoc test showed that *All dots the same* condition had a significantly lower number of fixations than the other two conditions. This difference is almost certainly because of the difference in the task that participants were performing in this condition. In the other conditions they have to look for a group of dots which can be joined and then join a small number of them. In the *All dots the same* game all the dots are the same symbol so participants do not have to scan them to

find the right ones to join. The differences in fixations is probably because participants did not have to perform visual search for this condition so they had fewer fixations. So, although this is a clear difference in eye data between the different games, it is probably due to differences in the task that players were performing rather than the experience that they were having.

I analysed other properties of the way that participants played the different games to generate more insight into their experience. The previous game puzzle experiment showed significant differences in the length of time participants took to think about their move. For this experiment I looked at length of time that participants take to think before their move (*thinking time*) and also the length of time they took to actually make their move (*move time*). There was a significant difference in *thinking time* between conditions. A post-hoc test showed that all of the conditions were significantly different from each other with the *Full game* having the longest thinking time and the *All dots the same* game having the lowest. There was also a significant difference in *move time* between conditions. This was due to participants in the *All dots the same* condition taking much longer to make a move. Many players of this game tried to join as many dots as possible in one go, even though this is not needed to succeed. It is possible that they did this to make the game more engaging. Pupil dilation can be caused by motor actions (Richer and Beatty, 1985) so it is possible that if participants in this game made more complex motor actions then they would have increased pupil dilation compared to the other conditions.

5.11. Limitations

The main limitations of this experiment are related to the issues of *timing* and *response* which were identified at the beginning of this chapter (see section 5.1). *Timing* is an issue because participants are in control of when they make a move so it is difficult to identify the point in time when they think about what move to make. Were they planning the move in advance or did they just make a split-second decision? Similarly, participants make their *response* by dragging over a number of dots which makes it difficult to separate the pupil response caused by the motor action from the response associated with deciding which move to make.

The third issue identified at the beginning of the chapter was that variations in the *luminance* of the stimulus may affect pupil dilation readings. It is possible that the small changes in graphics from one stage of the game to another made a difference to the luminance levels reaching participants' eyes and affected the pupil dilation data. However, the graphics for this experiment were almost identical to those of the previous experiment which showed a very strong difference in pupil dilation between conditions so this seems unlikely.

A greater limitation is likely to be the difficulty in measuring a baseline pupil dilation to normalise against. Initially I planned to take one single baseline pupil dilation measure at the start of each game. However, the previous game puzzle experiment showed that

the baseline tends to decrease from the beginning towards the end of the experiment. Using a single baseline measurement taken at the beginning of the experiment would become increasingly inaccurate as the experiment progressed so I needed to record baseline measurement for each move. In previous experiments participants have a rest period at the end of each move which is then followed by a baseline measurement. This would not be possible during a real game because the rest period and baseline measurement would interrupt the flow of the game and confuse players. To measure the baseline before each move, I used the first 0.5 second of each move as the baseline measurement for that move. There is a delay between actions and their corresponding pupil response so it is unlikely that participants' pupils react to the current move in the first half a second of that move. However, it is possible that this baseline measurement could be affected by the previous move as there is no pause between one move and the next.

The final limitation is the large differences in what players actually do during the game, even if they are playing the same variant. Some make a lot of moves, others much fewer. Some make fast progress through the levels; others find it much more difficult. These differences in behaviours make it difficult to systematically pin down what participants are doing at any one time and associate pupil activity with a particular game activity.

5.12. Experiment Pupil 5: Three variations of *Two Dots* with a new hypothesis

In the previous experiment the hypothesis that there would be a significant difference in pupil dilation between game conditions 1 second before the mouse press was not supported. However, exploratory analysis did find a significant difference at the time of the mouse press. The *All dots the same* condition had higher pupil dilation than the other two conditions. Participants in this condition also took significantly longer to make their moves than in the other conditions. I hypothesized that the increase in pupil dilation was down to the longer, more complex moves made by participants in this condition. However, this result was found by exploratory analysis and could be an artefact of the data. To test this hypothesis, I repeated the previous experiment but with a changed hypothesis. For this experiment the design, materials and procedure were exactly the same as the previous experiment. The analysis differed in that I did not analyse pupil fixations or saccades as this had not produced interesting results in the previous experiment. As with the previous experiment this experiment was also used for a separate analysis to investigate measuring game attention which is described in chapter 6. This required that in all versions of the game the edges of the screen contained images of *Disney* characters. Eye tracking showed that participants were focused on the

central area of the screen for a mean of 97.5% (SD = 2.53%) of the trial so it is unlikely that this made any difference to pupil dilation differences between the different games.

Hypothesis

The hypothesis of this experiment is very similar to that of the previous experiment except that the time window of difference in pupil dilation has been changed from 1 second before the mouse click to the actual mouse click time. The other change is that I hypothesize that the *All dots the same* condition will have higher pupil dilation than the other two conditions. The hypothesis of the experiment is:
For a 1 second window starting when participants press down the mouse to make a move there will be a significant difference between the pupil dilation in the different versions of the game. I expect the All dots the same condition to be higher than the other two conditions.

Participants

As this was a replication of the previous experiment, I used a similar number of participants. 54 participants took part in this experiment (18 in each condition). Ages ranged from 18-42 (Mean 21.22). 32 were male and 48 were native speakers of English.

5.13. Results

Pupil dilation

There was a significant difference between the mean normalised pupil dilation for the period of one second after the mouse press ($F(2,51) = 7.861$ $p=0.001$) (*Full game* M=1.001, SD=0.009, *No goals* M=0.995, SD=0.007, *All dots the same* M=1.006, SD=0.010). There was a moderate effect size between conditions. (η_p^2=0.231).
A Tukey's HSD post-hoc test (See Table 40) showed that the *All dots the same* condition is significantly different from *No goals* and that the difference between *All dots the same* and *Full game* is not significant but showing a clear trend (p=0.053).
The pupil dilation for each condition across the time of a move is shown in Figure 33 and Table 41.

	Full game	No goals
No goals	p=0.053	
All dots the same	p=0.303	p=0.001

Table 40 Tukey's HSD test of pupil dilation at the 0 second time period

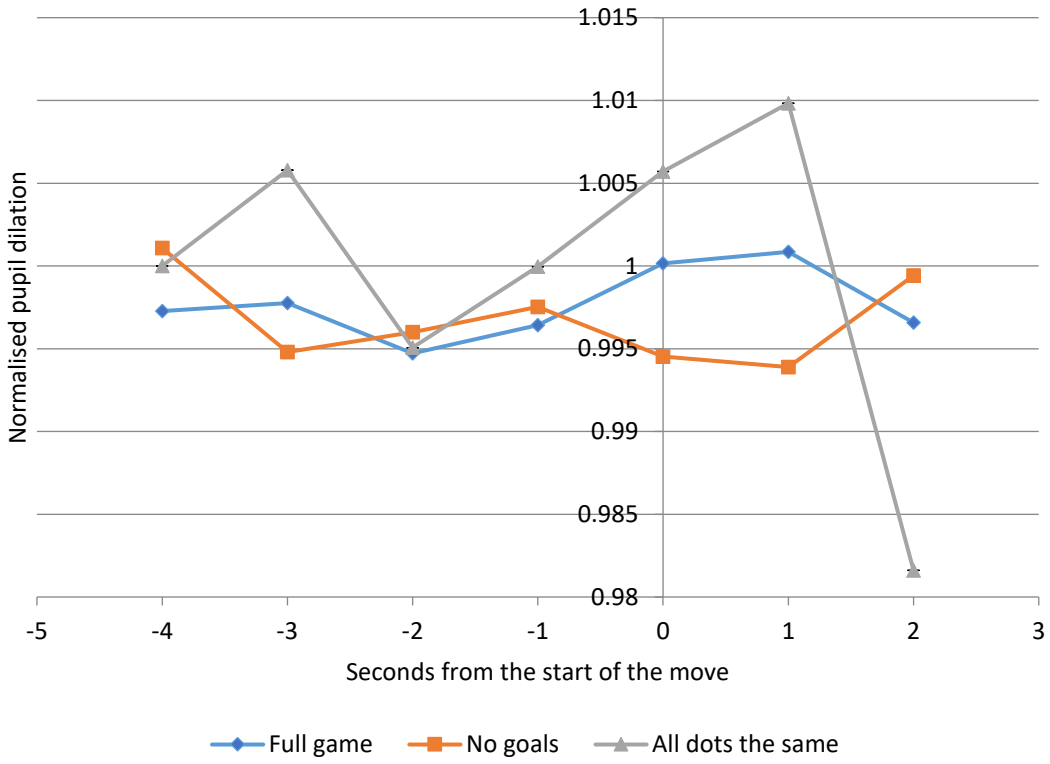

Figure 33 Mean pupil dilation across the time of each move. Error bars show standard error

	Full Game	**No goals**	**All dots the same**	**F(2,45)**	**p value**	η_p^2
Time/s from mouse press	Mean (SD)	Mean (SD)	Mean (SD)			
-4	0.996 (0.013)	1.001 (0.011)	1.006 (0.019)	1.340	0.276	0.075
-3	0.998 (0.009)	0.995 (0.006)	1.005 (0.010)	7.267	0.002	0.222
-2	0.995 (0.012)	0.996 (0.009)	0.995 (0.015)	0.050	0.951	0.002
-1	0.997 (0.009)	0.998 (0.005)	1.000 (0.004)	1.058	0.355	0.040
0	**1.001 (0.009)**	**0.995 (0.007)**	**1.006 (0.010)**	**7.681**	**0.001**	**0.231**
1	1.000 (0.012)	0.994 (0.012)	1.011 (0.018)	6.155	0.004	0.194
2	0.995 (0.016)	0.999(0.014)	0.991 (0.024)	0.717	0.495	0.037

Table 41 Comparison of pupil dilation between conditions over time

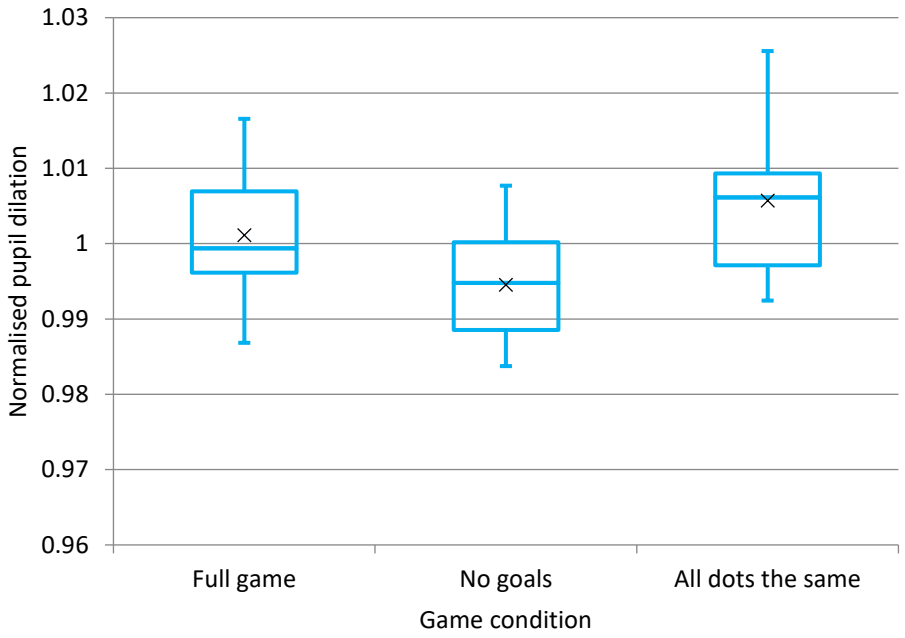

Figure 34 Boxplot showing mean pupil dilation for each condition at the 0 second time window

Immersion Experience Questionnaire (IEQ)

As with the previous experiment the purpose of the IEQ was to confirm that participants in each condition had had different levels of engagement in the game. There was not a significant difference in the immersion scores between the *Full game* (M= 101.94, SD=15.02), the *No Goals* game (M= 98.94, SD= 14.55) and the *All dots the same* game (M=89.61, SD=18.90) conditions; F(2,45)= 2.81,p=0.069. There was a moderate effect size between conditions (η_p^2= 0.099).

	Mean (Standard deviation)					
	Full game	**No goals**	**All dots the same**	**F(3,45)**	**p**	**Effect size (η_p^2)**
Cognitive involvement	32.83 (5.02)	32.44 (4.25)	27.56 (6.68)	5.318	0.008	0.173
Emotional involvement	15.17 (4.73)	14.89 (3.724)	13.28 (4.79)	0.949	0.394	0.036
Real world dissociation	33.00 (5.57)	31.72 (6.29)	31.11 (6.59)	0.440	0.646	0.017
Challenge	13.00 (2.425)	11.56 (2.255)	9.56 (3.185)	7.653	0.001	0.231
Control	15.50 (2.94)	16.00 (3.03)	14.94 (3.70)	0.478	0.623	0.018
Immersion	101.94 (15.02)	98.94 (14.55)	89.61 (18.90)	2.81	0.069	0.099

Table 42 Comparison of the IEQ subfactors between conditions

Gameplay differences

Each trial has a *Thinking time* which is the time before participants push down the mouse button and a *Move time* which is the amount of time that they take to make their move. There is a significant difference in both *Thinking time* and *Move time* between conditions.

	Full game	No goals	All dots the same	F(2, 45)	p value	η_p^2
	Mean (SD)	Mean (SD)	Mean (SD)			
Thinking time /ms	1917 (833)	1377 (552)	840 (274)	14.547	<0.001	0.363
Move time / ms	1120 (332)	762 (228)	4774 (2082)	59.142	<0.001	0.699

Table 43 Thinking time and move time for all three different games

There is a significant difference in both *Thinking time* and *Move time* between conditions. A Tukey's HSD post test showed that all three conditions have a significantly different *thinking time* from each other. A post-hoc test on the move times showed that the *All dots the same* game has a significantly higher *Move Time* compared to the other two conditions.

	Full game	No goals
No goals	p=0.014	
All dots the same	p<0.001	p=0.028

Table 44 Tukey's HSD of the differences in *thinking time* between conditions

	Full game	No goals
No goals	p=0.316	
All dots the same	p<0.001	P<0.001

Table 45 Tukey's HSD of the differences in *move time* between conditions

5.14. Discussion

The main hypothesis of the experiment was that there would be a significant difference in pupil dilation between conditions for the 1 second time window starting when the mouse button is pressed. This hypothesis was supported with a moderate effect size (η_p^2 =0.231). A boxplot and post-hoc tested showed that this difference between conditions is down to the *All dots the same* condition being higher than the other two. (Significantly higher than the *No goals* condition). As in the previous experiment there was a significant difference between conditions in the length of time that participants take to make their move (η_p^2 =0.699) which was due to participants in the *All dots the same* condition taking longer to make a move. The results from the previous experiment supported the hypothesis that the increase in pupil dilation during the *All dots the same* condition was due to the longer and more complex moves that participants made. The

results from this experiment support this position. The pupil dilation across the time of the experiment is consistently higher in the *All dots the same* condition than in the *Full game* condition even though the *All dots the same* condition has no puzzle elements and should be far less demanding than the *Full game.* If sustained cognitive load was needed for the *Full game* then you would expect to see higher pupil dilation than the *All dots the same* game which has no puzzle elements. This suggests that sustained cognitive load is not required to play the *Full game* and that pupil dilation as a measure of cognitive load is not likely to be an effective measure of game experience.

This experiment used the same three variants of one game as the previous one. The reason for this was to ensure that participants had different game experiences while performing similar tasks. To confirm that this was the case participants completed an immersion questionnaire (IEQ) after playing the game. However, the difference in immersion between games was not significant in this case although there was a moderate effect size between conditions (η_p^2= 0.099). Looking at the subfactors of the IEQ only *Cognitive Involvement* and *Challenge* were significantly different between conditions. This differs from the last experiment in which *Cognitive Involvement* has no significant difference but the *Emotional Involvement* subfactor was significantly different.

5.15. Limitations

This experiment has similar limitations to the previous experiment concerned with measuring cognitive load during a game of *Two Dots. Timing* is still an issue because participants are in control of when they make a move so it is difficult to identify the point in time when they think about what move to make. Similarly, participants make their *response* by dragging over a number of dots which makes it difficult to separate the pupil response caused by the motor action from the response associated with deciding which move to make. As with the previous experiment all pupil dilation is normalised against a baseline pupil dilation which is measured during the first 0.5 second of each move. It is possible that this baseline measurement is affected by pupil dilation from the previous move. Despite these potential limitations the hypothesis there would be a significant difference in pupil dilation between conditions and that the *All dots the same* condition would have the highest pupil dilation was supported. This suggests that although these limitation factors may add noise to the pupil data this noise is not so great to prevent significant differences between conditions being detected.

Participants in this experiment and the previous one played one of three different variations on the game of *Two Dots.* The difference in games was to ensure that participants in different conditions had a different game experience. In the previous experiment this was confirmed by a post-game immersion questionnaire which showed a significant difference between conditions. In this experiment there was no significant difference in immersion between the conditions even though participants played the same three games. This would be a problem if the main hypothesis of this experiment was that pupil dilation would be different due to difference in game experience. However, the hypothesis of this experiment was that pupil dilation would be different

due to differences in motor actions between conditions. This hypothesis was supported despite the lack of difference in game experience. This suggests that significant differences in pupil dilation can occur without significant differences in game experience which supports the finding that pupil dilation is not a good measure of game experience.

5.16. Chapter conclusions

The goal of this chapter was to show that pupil dilation can be used to measure the changes in cognitive load which happen when participants are playing a real self-paced game. The motivation behind that goal was that changes in cognitive load would reflect changes in the player experience and that pupil dilation could be used as a measure of the experience of playing self-paced games.

The first experiment in this study showed that a "game-like" task can create large differences in pupil dilation due to differences in cognitive load used. The task in this experiment was very similar to that performed in a game like *Two Dots*. This experiment contained a training phase in which participants played the *Full game* of *Two Dots* while their pupil dilation was measured. The data from this training phase was then used for an exploratory analysis of how different game events affect pupil dilation. This analysis found a moderate effect size between pupil dilation due to success or failure at the game. The difference was not significant but the next experiment attempted to replicate the effect and found a small effect size which was also not significant. It is possible than an experiment with more participants might show that this is a robust effect. The pupil dilation is higher after a success than a failure which is the reverse of what I expected. I anticipated that players would try harder after a failure and show more pupil dilation than after a success. This suggests that if there is a robust effect then it looks more likely to be due to the player's positive emotion at completing the level than because they are using more cognitive effort to play the game.

The second experiment attempted to use pupil dilation to measure the difference in cognitive load between three different versions of the same game. The three different games were designed to give players a different game experience even though they used a similar stimulus (a grid of dots) and had a similar task (joining those dots). The *All dots the same* game had no puzzle elements so I expected that it would require significantly less cognitive load to play than the *Full game* of *Two Dots* which is fully featured puzzle game. An immersion questionnaire showed that players of the different games did indeed have significantly different game experiences but this was not reflected in a significant difference in the pupil dilation and my initial hypothesis was not supported. Looking at the pupil dilation across time the only significant difference was at window starting from when the mouse button was pressed. Surprisingly this difference showed the *All dots the same* condition had higher pupil dilation than the *Full game* which is the opposite of what I expected. I expected the higher cognitive load required in the *Full game* would lead to higher pupil dilation than the *All dots the same* game but this was not the case. Players of the *All dots the same* game spent significantly

longer making their moves than players of the other games. This led me to hypothesize that the difference in pupil dilation was mainly due to players concentrating more on making long complex moves than increased cognitive load due to the difficulty of the game.

The third experiment tested this hypothesis by repeating the previous experiment on three different variations of the game. The hypothesis was supported by significant difference in pupil dilation with the *All dots the same* game showing higher pupil dilation than the other two conditions. Participants in this experiment also completed an immersion experience questionnaire after the game but this did not show a significant difference in game experience between conditions. This experiment found a significant difference in pupil dilation but it is most likely to be due to differences in the physical moves that participants made rather than their experience of the game.

One possible explanation for this is that the different games did require different amounts of cognitive load but my experiments did not detect it due to other confounds on the data. Playing an actual game is much more complex and varied than the game puzzle which participants had to solve in the first experiment. Participants decide themselves when to move and make longer and more varied motor actions. Players may occasionally be using significant cognitive effort for some moves but not for the rest. This could mean that small differences in pupil dilation may exist but my exploratory analysis which compared pupil dilation during different events in the game did not pick it up. The third experiment in this chapter found a significant difference in pupil dilation between games but this is most likely due to participants making longer more complex motor actions in the *All dots the same* game rather than them using more cognitive load to solve the game's puzzles. It is possible that even if players are using sustained cognitive load to play the game that this pupil dilation is masked by the additional pupil dilation produced by making moves in the game. However, comparing the pupil dilation variation in the first puzzle experiment with the other two game experiments shows that the pupil dilation in the puzzle experiment was much higher than the game experiment. Typically, the pupil dilation in the puzzle experiment hard task increases by an additional 10% above the baseline whereas pupil dilation in the game experiment varied by values in the order of around 0.4%. This would suggest that if participants were using similar amounts of cognitive effort to play the game as to solve the hard task puzzle then it would be clearly visible in the analysis. Even if there are small changes of pupil dilation due to the game experience, then differences this small are unlikely to be a useful measure of game engagement. Tests between different games would need a large number of participants for a significant result and would be unable to pick up subtle differences between games.

Another possible explanation is that participants are not using significant cognitive effort when playing any of the versions of the game, including the *Full game* version which requires puzzles to be solved. This would explain why there is not a significant difference between the different games. This was unexpected because the first game puzzle experiment did show that participants use significant cognitive effort to solve

the hard game puzzle. As the hard puzzle is derived from the game of *Two Dots* this suggested that they would do the same during the game. However, the pupil dilation analysis shows no sign of this level of cognitive effort which may be because participants in the game experiments have more choice over which moves and strategies to use than participants in the puzzle experiment. In the puzzle experiment participants are given a particular move to find and the only way to find that move is to perform the cognitively demanding task of mentally removing dots and letting them drop down. In the game tasks participants can choose which moves to make during the game and which strategies to use to find them. As the choice is up to them, they choose to use easier strategies which do not need sustained cognitive load.

If participants are not using much cognitive effort this would fit with Load theory (Lavie, 2005, Lavie et al., 2004) which found that high cognitive effort makes it easier to distract people. Games like *Two Dots* are highly successful at keeping attention. If they required too much cognitive effort then Load theory predicts that players would be more easily distracted. The idea that game players are not using significant cognitive effort would also explain why people can play games for long periods: if games really did require substantial cognitive effort, they would be exhausting. Instead these results suggest, games require cognitive engagement but not substantial effort and hence can be a source of activity for a long period of time. This cognitive engagement is created by the design of the game which may activate mechanisms such as curiosity (Silvia, 2006, Loewenstein, 1994) or closure (Webster and Kruglanski, 1997, Berenbaum et al., 2008) to create sustained engagement without the need for substantial effort.

Whichever explanation is correct (and it may be a combination of the two) it looks unlikely that pupil dilation as a measure of cognitive load will prove to be an effective measure of the experience of playing self-paced games. If pupil dilation does change in response to game experience then confounds on the data may make these changes too difficult to measure reliably. Likewise, if sustained cognitive load is not a key part of the experience of playing puzzle games then measuring that cognitive load will not prove an effective measure of that experience. These experiments show that pupil dilation can be used to measure cognitive load during controlled situations but the complexity of a real game, even one as simple as *Two Dots* makes measurement of cognitive load during a game uncertain and most likely impractical for most uses. Added to that is evidence that sustained cognitive load may not be a key part of the experience of playing puzzle games so I felt that I should try a different approach to finding a new measure of the experience of playing self-paced games. I decided to look at how games manage to hold players' attention and my investigations into this are described in the next two chapters.

6. Measuring game attention

The previous chapter concluded that measuring pupil dilation is unlikely to provide a useful measure of the experience of playing self-paced games. This left me with the need to find an alternative measure of this experience. Ideally this would be an "on-line" measure which measured experience directly rather than relying on participants' self-reports of the experience. The advantage of "on-line" measures is that they can indicate peaks and troughs in the experience and they are also not subject to bias due to participants being unable or unwilling to report their full experience. To find other measures I considered other aspects of the game playing experience. If a particular aspect of the experience is common to all self-paced games and has the potential to be measured, then it may be a candidate to become a useful measure of the experience of playing games.

One of the defining features of the game playing experience is the way they hold our attention and stop us getting distracted by external events. This applies to most successful games, not just self-paced games but for self-paced games this feature is more remarkable because players could stop at any time to attend to distractions but generally their attention is held by the game and they ignore distractions. Intuitively it seems that "good" well designed games will hold our attention better than "bad" less engaging games. So, if we had a reliable method of measuring how well a game holds our attention then it would be useful for testing and designing better games. It could also give insights into how self-paced games manage to keep people's attention even though there is nothing in the game which prevents them stopping at any time. This chapter describes a series of experiments which seek to measure how well games hold players' attention and stop them being distracted.

As described in the literature review (chapter 2) researchers such as Brockmyer et al. (2009) and Jennett (2010) have already used distraction techniques to measure how engaged players are in a game. Brockmyer et al. (2009) played the same audio distractor to game players several times and noted whether they responded to the distractor and how quickly. Jennett (2010) played 9 different audio distractors to participants during a 10-minute game.

There are several sources of evidence to suggest that visual distractors may also be suitable for measuring attention and be more memorable than audio distractors. Raveh and Lavie (2015)'s work on attention and distraction suggests that the attention system may treat visual and audio distraction in a similar way. Paivio and Csapo (1971) found that people were more likely to remember words if they were concrete nouns, which could be represented visually, than if they were abstract concepts. This suggests that visual recall may be more reliable than audio recall. Standing (1973) tested how reliable visual recall could be by showing participants a large number of images in turn and then testing how well they recalled those images. He found that participants remembered around 90% of the images they were shown. Brady et al. (2008) repeated the study but tested participants on the details of the pictures and found a similarly high rate of retention and recall. In contrast to this high level of visual recognition, Miller and Tanis (1971) found that participants only remembered 75% of an audio stimulus consisting of common words. This suggests that visual recall may be more reliable than audio recall and so using visual distractors may be a more reliable measure of attention that audio distractors.

There is another reason why visual distractors may be superior to audio distractors. A picture can be seen and comprehended almost immediately but an audio clip takes some time to play so players may need to be distracted from the game for several seconds to hear the whole clip. This may lead to situations where participants only notice half of the audio distractor which would make it difficult to decide if they heard it or not. Because visual distractors take up less time than audio distractors it is possible to have a larger number of them spread throughout the game experience. Having a larger number of distractors makes it less likely that the measure could be skewed if a particularly engrossing section of the game happens to coincide with the distractor. A measure of attention based on distractors would test players after the game to see how many distractors they remember. Having more distractors allows the measure to be finer grained than if I was using a small of number of distractors. For example, if participants are only shown 5 different distractors during a game then the number that they can remember will between 0 and 5 and the measure can have 6 different steps. However, if they are shown 60 different distractors then potentially the measure has 61 different values so is finer grained and may be more sensitive to small changes in game experience. I decided to use visual distractors because they are more memorable than audio distractors and can be used in greater numbers which should make a finer grained measure.

6.1. Distractor recognition paradigm

Having decided to use visual distractors rather than audio distractors I needed to design an experimental method to put this into practice. I called this method the *distractor recognition paradigm* (DRP). This paradigm has two parts, the first is the presentation of the distractors during the game and the second is the test after the game to see how many distractors participants recognise seeing during the first presentation.

6.1.1. Presentation of distractors

Mobile casual games like *Two Dots* are designed to be played on a small (4-6") mobile phone screen. When converted to play on a wide screen desktop monitor the game is usually in the middle of the screen with large gaps on either side. I decided to use the middle half of the screen for the game and surrounding quarters for the distractors. If players remember seeing a distractor at any one point in the game then their attention was not strongly held by the game, conversely if they do not remember the distractor then their attention was probably on the game. If a distractor stays on the screen for too long, then participants are likely to get used to it and it becomes less distracting. A useful measure of game attention would give an indication of how strongly players' attention is held at different points in the game, so each point in the game needs to be associated with a different distractor. In Standing (1973)'s study participants saw each image for 5 seconds at a time. I decided to follow this example and display the distractors for 5 seconds before they change to another one. To provide a constant level of distraction, the distractors also need to have a similar level of interest. For the initial experiments I used icons taken from the *Webdings* typeface. This was chosen because the symbols are different from each other but all have a similar level of interest. The symbols were pre-screened so that they were all reasonably different from each other. For example, *Webdings* has several rain cloud symbols, only one was used. For the experiments which used this paradigm, participants played the game for 5 minutes, during this time the distractors changed every 5 seconds without repetition, so participants were shown a total of 60 distractors. The *Webdings* icons are all different shapes and also vary in size. The visual attention system is very sensitive to differences in size and shape (Wolfe, 2014) which can cause objects to "pop" out of a visual search. I could have just displayed a single large icon on either size of the screen, but this would have emphasised these size and shape differences and may have made some distractors more distracting than others. To reduce this effect, I displayed the distractors as 18 small repeating "tiles" on each side of the screen. This makes the overall visual display of each distractor more similar to the other distractors and may make the level of distraction more constant. See Figure 35.

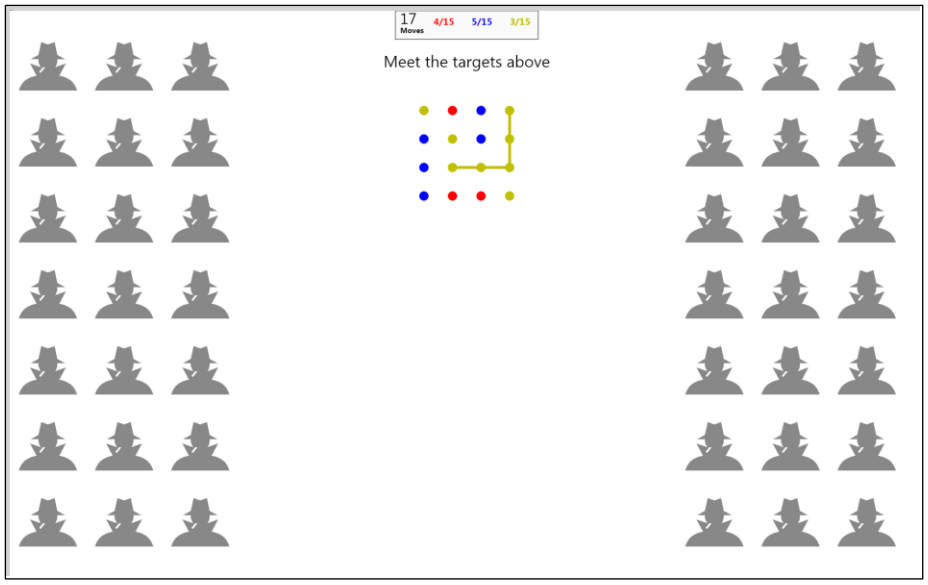

Figure 35 The Two Dots game surrounded by distractor symbols.

Experimental setup

The stimulus was shown on a 24" inch monitor with screen dimensions of 51.5 x 32.5cm. During the game play participants kept their chin in a chin-rest which was positioned 95cm from the screen. This meant that the screen display filled 31.5° of the participant's field of view. The chin rest was used to make sure that all participants had the same field of view of the screen and could not look away from the screen and see what was happening in the surrounding room. Participants could adjust the height of their chair to make sure that they were comfortable, but the chin rest was in the same place for all participants. Both games ran full screen and were controlled by a mouse.

6.1.2. Recognition of distractors

To measure how well the game held participants' attention I needed some way of knowing how many of the surrounding distractors they had noticed. Kinoshita (1995)'s work showed that if participants are paying attention to a stimulus then they are more likely to recognise it if shown it again. So, the simplest measure is just to show participants all of the distractors and ask them which they had seen before. When participants are shown a stimulus and asked if they have seen it before this is known as *recognition* (Baddeley, 2013). In Kinoshita's experiment she showed participants a list of words, some of which had been shown during the experiment and others which had not. She then asked participants to indicate which they had been shown before. A problem with this approach is that the results can be changed by how confident participants feel. Participants who do not feel confident are less likely to indicate that they remember a distractor. Similarly, very confident participants are more likely to indicate that they have seen a distractor and are more likely to also indicate that they have seen something they have not been shown. To avoid these issues, I used a *forced choice test* similar to that used by Standing (1973). This means that during the recognition test participants are shown one distractor and one dummy image that they have not been shown but is similar to the other distractors. Participants have to indicate

which of these images they have been shown during the experiment. Because they are forced to make a choice it does not matter how confident they are in that choice, all they are being asked to indicate is which image they are more confident about. For the experiments in this study participants were shown 60 different distractors. To keep the testing phase short, I only tested participants on even numbered distractors, so they were tested on 30 different distractors. As this is a multiple-choice test with two options, they have a 50% chance of getting the correct answer by chance. As there are 30 questions the expected score, if participants guess randomly for each question, is 15. Testing was done on screen and participants selected their answer by clicking with the mouse (See Figure 36).

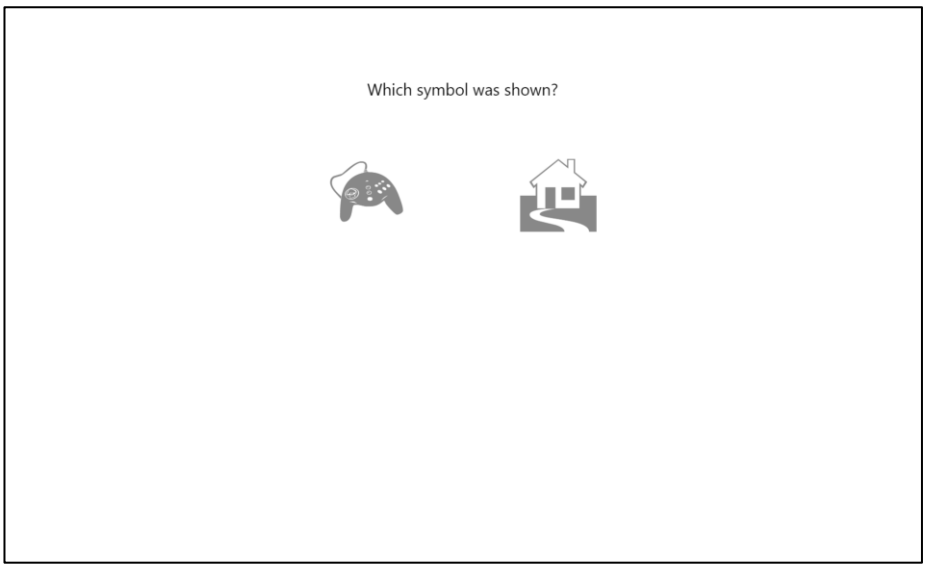

Figure 36. A screen from the distractor recognition test. One of these pictures has been shown to the participant during the experiment. The other is a dummy which has not been shown before.

Eye tracking

Another way of testing whether participants have been distracted by visual distractors is to use eye tracking to see whether they are looking at the game or the distractors. There have been many studies (Deubel and Schneider, 1996, Shepherd et al., 1986) which show that the area that your eyes focus on corresponds to your area of attention. Shepherd et al. (1986) also found that although it is possible to change your focus of attention without moving your eyes, moving your eyes always results in a shift in attention. I used eye tracking to measure distraction in two experiments described in this chapter - *6.4 Experiment attention 3: More similar games with* eye tracking and *6.5 Experiment attention 4: More distracting distractors* memory test.

6.1.3. Experimental plan

Before testing whether the distractor recognition paradigm could be used to measure attention in a real game, I wanted to validate it to make sure that participants would recognise the distraction images in a situation where there was no game for them to play. This was important to set a "ceiling" on the number of distractors which could be

recognised by participants. Standing (1973) and Kinoshita (1995)'s results suggested that participants *would* remember distractors if their attention was not distracted by a game but I wanted to test this by running an experiment which used the paradigm without a game.
After validating the distractor recognition paradigm, I planned to explore how effective it was at measuring attention in real games. The next experiment compared two very different games to see if the distractor paradigm could detect a difference in how well they held attention. After that I compared three more similar games and also looked at whether eye tracking could be used to create a more sensitive measure of attention. I then tried changing the distractors to be more interesting to see if this would create a stronger distraction effect. Finally, I looked at creating a more natural ecologically valid environment by removing the chin rest and leaving the participant alone in the experiment room during the experiment.

6.2. Experiment attention 1: Validating the distractor recognition paradigm

Aims

The main aim of this experiment is to validate the distractor recognition paradigm by investigating how many distractors would be remembered if there was no game to play so that participants' full attention was on the distractors. The number of distractors remembered gives an idea of the maximum which could possibly be given even if participants were completely disengaged from the game. For an effective measure this "ceiling" would be high to show that participants would remember most of the distractors if their attention was not on the game.
This experiment aimed to find out how many distractor images participants could remember.

Design

This was a single condition experiment. All participants were shown 60 distractor images and then tested to see which ones they could recognise as described in *6.1 Distractor recognition paradigm.*

Participants

10 staff and students from colleges in York took part in the study. 8 were male, 8 were native speakers of English with ages ranging from 18-27 (Mean 20.3)

Materials

This experiment uses the distractor recognition paradigm as described in section 6.1. The images shown in the presentation phase and used as dummies were randomly chosen from a pool of 90 icons which was different for each participant. They were also presented in a random order which differed for each participant.

Procedure

Participants performed a consent procedure and were then told to watch the screen and that they would be asked some questions afterwards. They were then shown 60 distractor images in the presentation phase of the experiment. Immediately afterwards they performed the testing phase of the experiment as described in *6.1 Distractor recognition paradigm* and were tested on 30 images to see whether they recognised seeing them.

Results

The mean number of distractor images recognised was 27.5 out of 30 (SD=2.88). This is equivalent to recognising 91.6% of the images.
I plotted the probability that each distractor would be correctly recognised which can be seen in Figure 37.

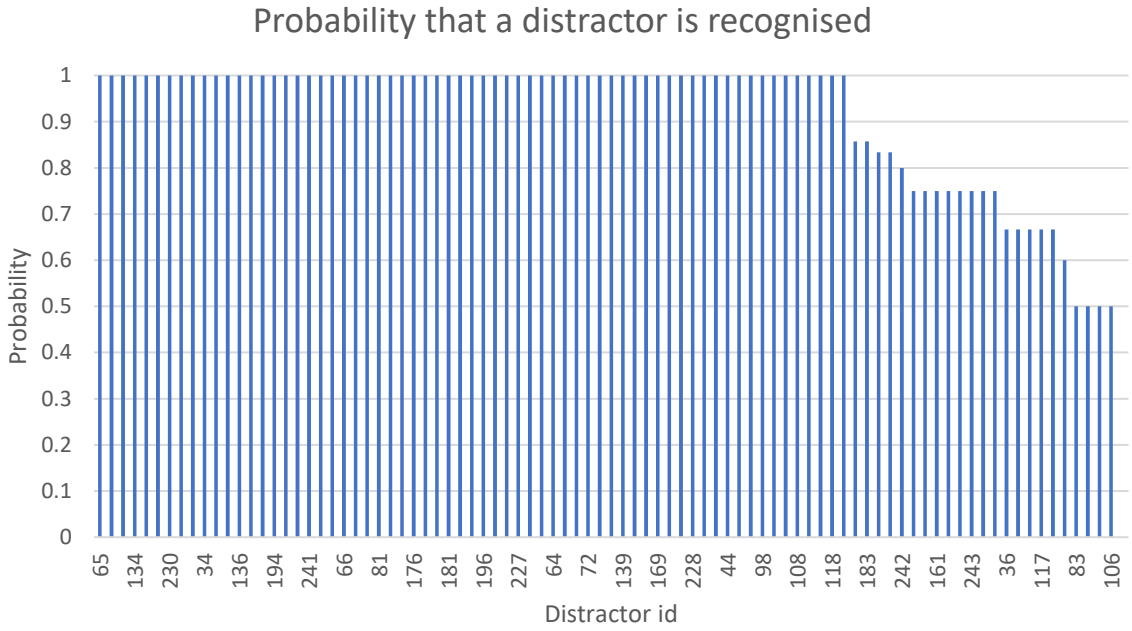

Figure 37 The probability that each different distractor will be recognised. The distractor ids are ordered by probability – highest to lowest.

I also plotted the probability that each distractor would be recognised against the time through the experiment. This is shown in Figure 38.

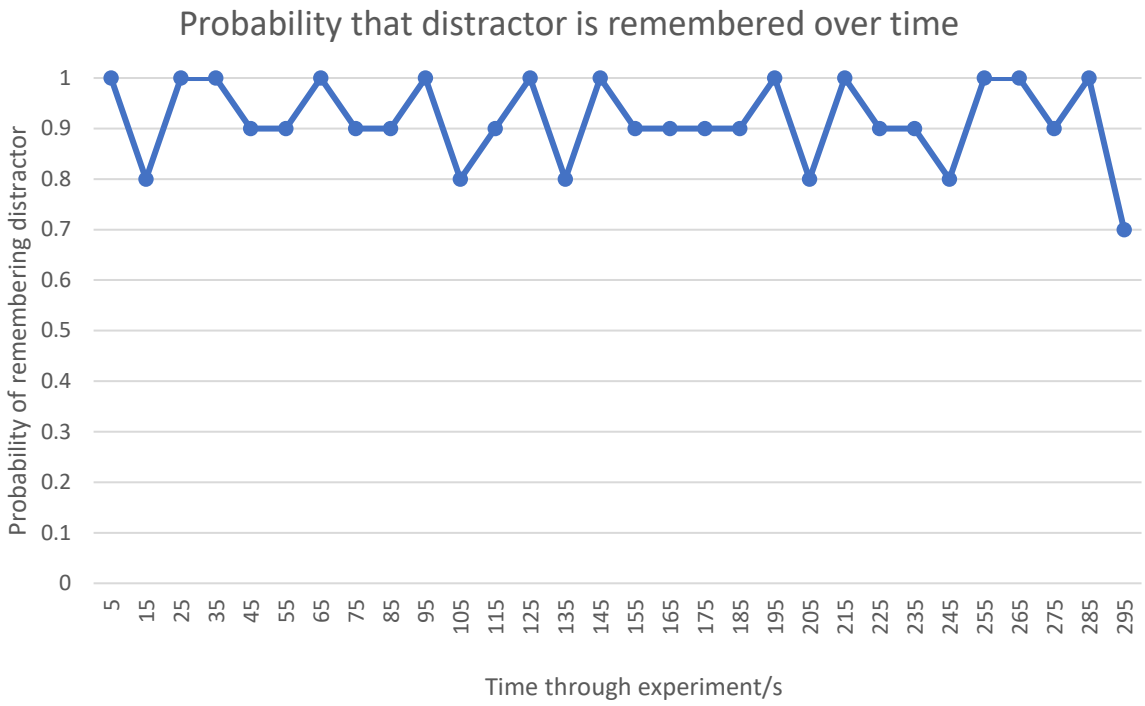

Figure 38 The probability that a distractor will be recognised against the time it is displayed

Conclusions

The results of this experiment support Standing (1973)'s finding that participants remembered over 90% of the images they were shown. The results indicate that participants have a high level of recognition of the set of 60 *Webdings* characters when they are giving them their full attention. That means that if these characters were used as distractors then participants are likely to recognise them if they are paying attention to the distractors rather than the game. To see whether any of the distractors were more likely to be recognised than others I plotted the probability of each them being recognised on a graph. This shows that it is difficult to see any particular pattern as almost all the distractors were recognised every time which is not surprising as average recognition rates were over 90%. The probability curve of those that were not recognised each time is roughly in line with random distribution. So, although it is possible that there are some small differences between the recognition rates of individual distractor images this experiment does not show any firm evidence that this is the case. I also plotted the probability of each distractor being recognised against the time they were shown in the experiment. This also did not show any definite pattern of recognition changing over time. Although once again, as average recognition rates were over 90% the overall variation was small, so small changes over time were unlikely to be visible.

The high rate of recognition, coupled with the lack of variation in recognition due to the distractor image displayed or the time through the experiment, validate that this distractor recognition paradigm may be an effective measure of game attention.

6.3. Experiment attention 2: Distraction between two games

Aims

The main aim of this next experiment is to test the feasibility of using the *distractor recognition paradigm* to measure the difference in how well games hold attention. This is an unproven paradigm with no previous data to form an estimate of likely effect sizes. I decided to use two very different games which were likely to give a large difference in how well they held attention. The two games used in the experiment were the *Full game* and the *Bad game* (see chapter 3 for more details). These games had a good chance of producing a significant difference in how well they held participants' attention. If there was not a significant difference in distractors recognised then it would indicate that I may have to change the paradigm to be able to measure attention effectively. I also aimed to measure how immersed players were in the two different games. Jennett (2010)'s experiments suggest that immersion is a form of selective attention and I wanted to compare the measures of immersion and attention to see how they related to each other.

Hypothesis

The hypothesis of the experiment was:
The number of distractors that participants remember will be higher for the *Bad game* condition than the *Full game* condition. The null hypothesis is that there will be no difference between the number of distractors remembered.

6.3.2. Method

Design

This was a between-subjects design with two conditions. The independent variable was the game each participant played – either the *Full game* or the *Bad game*. The main dependent variable was the number of distractors that participants recognise after the activity. Another, secondary dependent variable is the Immersion Experience Questionnaire (IEQ) score for each participant's experience of the activity.

Participants

20 students and staff from the University of York took part in the study. 12 were men. Ages ranged from 21 to 50 (mean = 30.7). Game experience and attitudes varied between the participants, ranging from those who played games less than once per month, to those who played several times a week. All participants received chocolate for their participation.

Materials

Participants either played the *Full game* or the *Bad Game* which are described chapter 3. For both games, attention was measured using the distractor recognition paradigm as described in section 6.1. As described in this paradigm, participants completed an on-screen test of how many distractors they remembered.

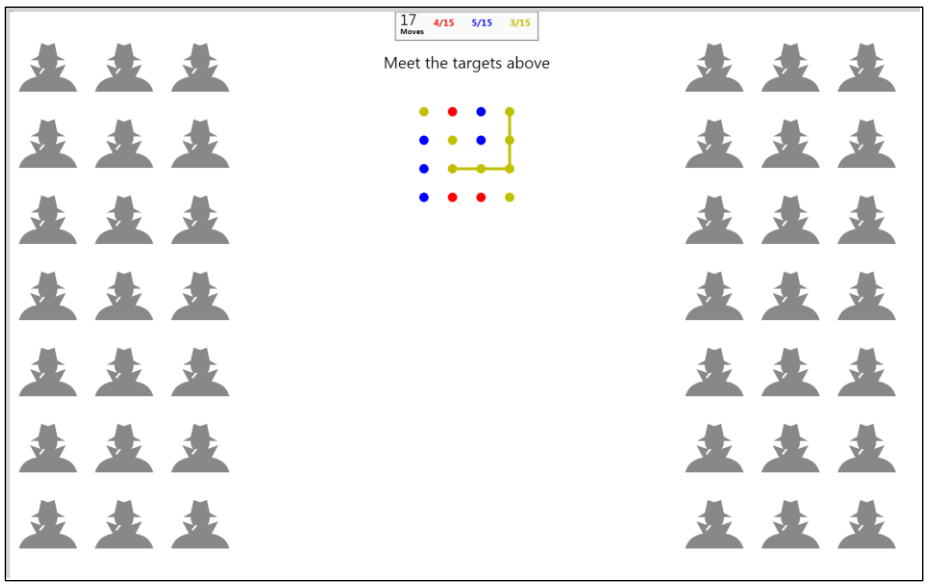

Figure 39 The *Full game* surrounded by distractor symbols.

6.3.3. Experimental setup

The experimental setup was as described in in *6.1 Distractor recognition paradigm.*

6.3.4. Procedure

Participants began by completing a consent form. They then played either the *Full game* condition or the *Bad game* condition. Both games included a short tutorial to teach participants how to play. Participants played the game for 5 minutes. After this time the game stopped automatically and the participants then completed the on-screen distractor recognition test. This was done immediately after the game so that all participants in both conditions would not forget any distractors or have their memories confused by other tasks. After the distractor test participants came away from the chin rest and screen and filled in a paper based demographic questionnaire and an IEQ about their experience of the game.

6.3.5. Results

Distractors recognised

There was a significant difference in the number of distractors correctly remembered between the *Full game* (M=13.0, SD=3.1) and the *Bad game* (M=19.4, SD= 3.9) conditions; F(1,18)= 16.28, p=0.001, η_p^2= 0.475.

Each distractor appeared at a particular time in the activity so the time of a distractor can be plotted against the chance of it being recognised. This is shown for both conditions in Figure 43.

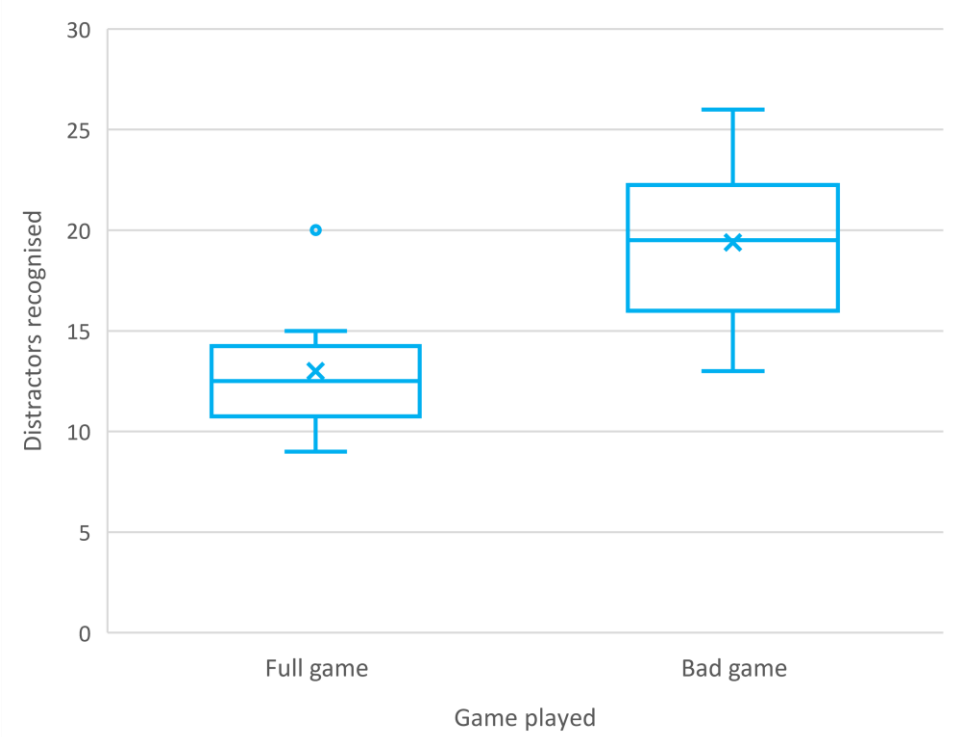

Figure 40 Boxplot showing the number of distractors recognised for both conditions

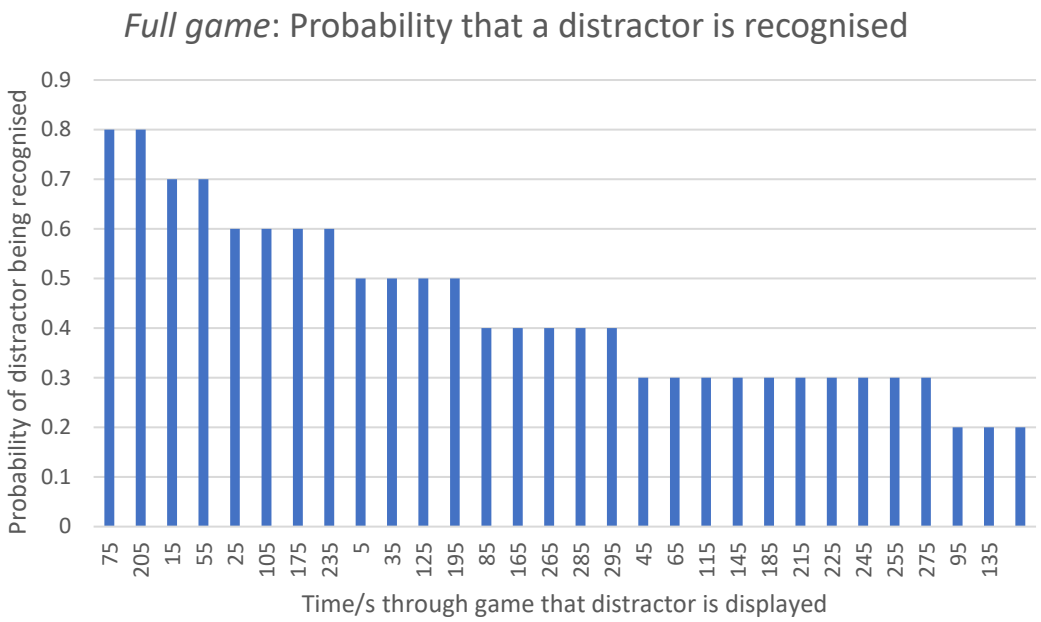

Figure 41 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest.

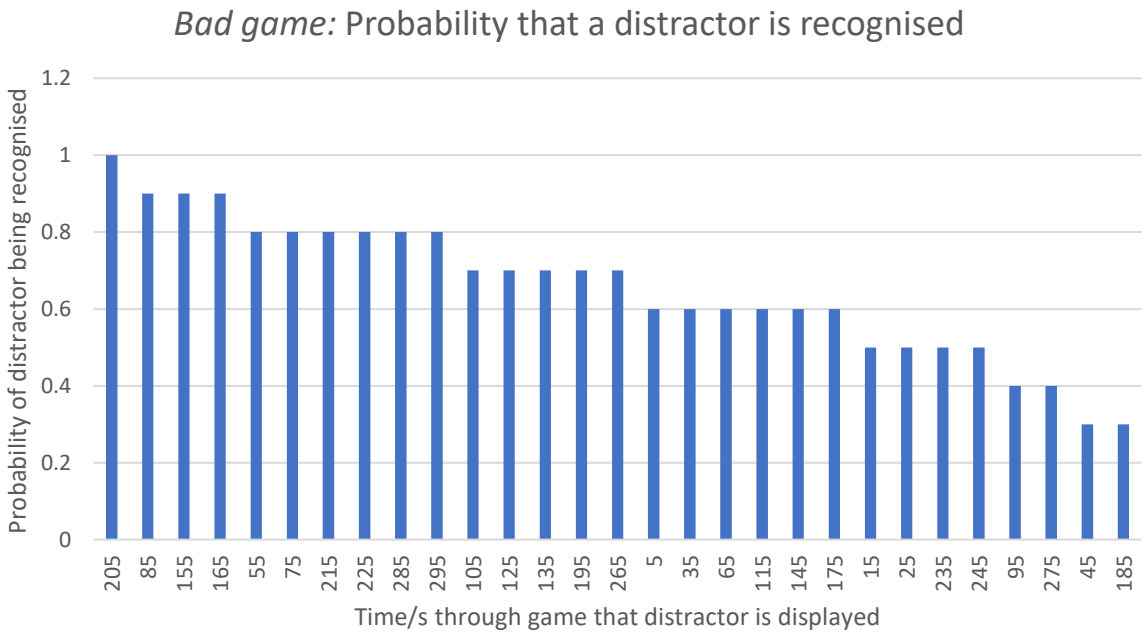

Figure 42 The probability that each different distractor shown in the *Bad game* will be recognised. The distractor times are ordered by probability – highest to lowest.

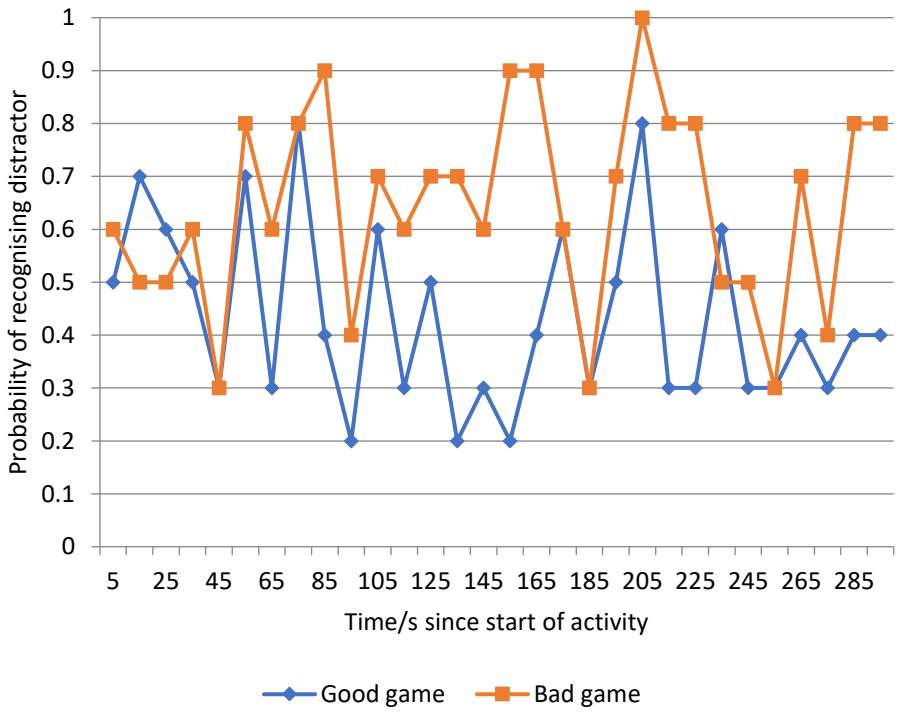

Figure 43 The probability of a distractor being recognised correctly plotted against the time it was shown during the game.

Immersion Experience Questionnaire (IEQ)

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was a significant difference in the immersion scores between the *Full game* (M=110.2, SD=14.4) and the *Bad game* (M= 84.6, SD= 10.8) conditions; $F(1,18)= 20.30, p<0.001, \eta_p^2= 0.530$.

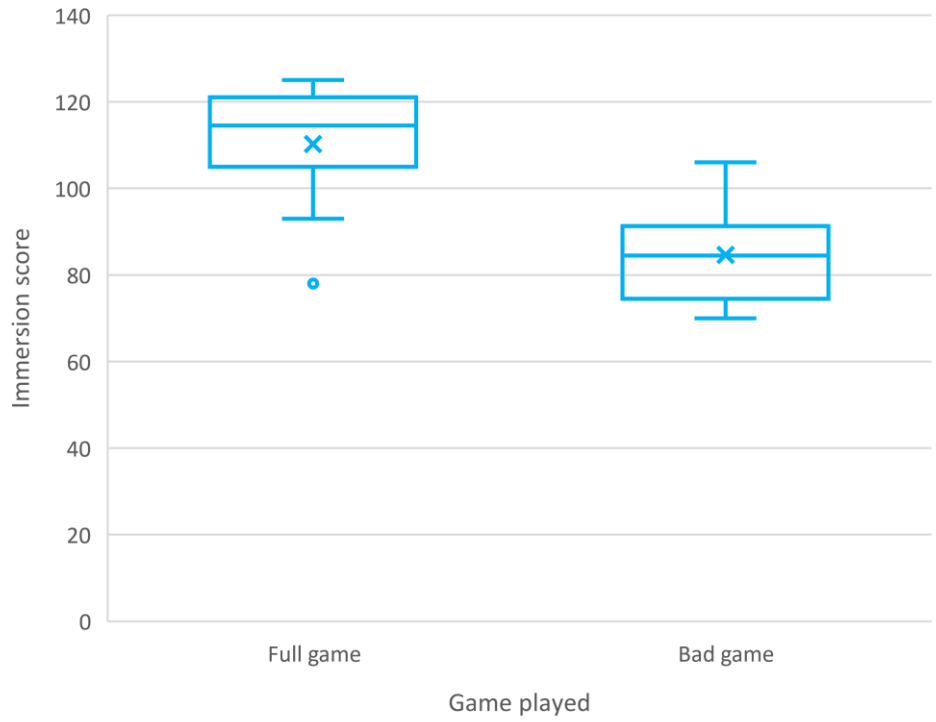

Figure 44 Boxplot showing IEQ scores for both conditions

IEQ scores can be broken down into five categories; *Cognitive involvement, Emotional involvement, Real world dissociation, Challenge and Control*. As can be seen from Table 46 there was a highly significant difference between the three conditions in the scores for *Cognitive involvement*, *Emotional involvement* and *Challenge*. Although there was a difference in the other scores, *Real world dissociation* and *Control* it was smaller and only not significant.

	Good game Mean (SD)	**Bad game** Mean (SD)	**Effect size** η_p^2	**p value**	**F(2,19)**
Cognitive involvement	34.3(5.5)	26.4(4.6)	0.356	0.005	9.950
Emotional involvement	19.5(6.2)	12.2(3.6)	0.489	0.001	17.246
Real world dissociation	32.2(7.6)	27.5(6.7)	0.153	0.088	3.259
Challenge	14(1.7)	9.4(2.9)	0.444	0.001	14.384
Control	17.8(4.2)	15.5(2.0)	0.148	0.094	3.130
Immersion	110.2(14.4)	84.6(10.8)	0.530	p<0.001	20.297

Table 46 Results from the IEQ

Correlations

There was not a significant Pearson's correlation between the number of distractors recognised and the immersion score for either the *Full game* (r= 0.197, t(8)= 0.570,p= 0.585) or the *Bad game* (r= 0.328, t(8)= 0.982,p= 0.355).
I also calculated Pearson's r correlations between the number of distractors recognised and the *Real World Dissociation* component of the IEQ and found no significant correlations for either the *Full game* (r=-0.269, t(8) = -0.791,p=0.452) or the *Bad game* (r=0.472, t(8) = 1.514,p= 0.168).

6.3.6. Discussion

The hypothesis that the number of distractors remembered by participants who played the *Bad game* would be higher than the number remembered by those in playing the *Full game* was supported. The level of immersion for participants who played the *Full game* was significantly higher than for those who played the *Bad game.* This indicates that these participants had a significantly different game experience. These two results indicate that the distractor recognition paradigm may be a useful metric for measuring how well self-paced games hold players' attention.

Jennett (2010) compared two games, one with feedback and one without. She found a significant difference in both immersion (IEQ) scores and how likely players were to be distracted by audio distractors. This is similar to results reported in this study. Looking at the individual components of the IEQ score there are significant differences in *Cognitive Involvement, Emotion Involvement* and *Challenge.* There are also trends towards significance in *Control* and *Real World Dissociation.* This breakdown is consistent with the difference between the two games. The *Bad game* had different gameplay and was a less involving experience which is reflected in the three significant factors. Many participants tried to make the *Bad game* more interesting by making long complex moves and trying to join all the dots at once which may explain the difference in *Control.* Differences in *Real World Dissociation* may be due to differences in how well the different games held participants' attention.

One of the aims of the study was to measure differences in attention across time. These differences are plotted in Figure 43. Visual inspection of the graph does not show any strong patterns across time. This may be because there is a 10 second gap between the symbols which participants are tested on. For a pattern to show up, one of the activities would need to have patterns of interest which varied over time frames longer than this. The *Full game* had occasional events which may have caused a change in attention, such as when players finished a level. But these generally take less than a couple of seconds. The *Bad game* had an almost constant level of interest all the way through with almost no game events which would change attention. Given this lack of change in the activities it is not surprising that the graph does not show strong patterns of attention change.

I considered that some distractors may be more memorable than others. Each distractor was shown at the same point in time in each condition. To see whether any of the distractors were more likely to be recognised than others I plotted the probability of each them being recognised on a graph. The probability curve for each condition is roughly in line with what would be expected due to random variation. However, each symbol was always presented at the same time during the activity so there may be an interaction affect with what is happening in the game. To avoid this issue, the subsequent experiments randomised the order in which distractors were displayed.

Jennett (2010) suggested that immersion is a form of directed attention. This would suggest that total immersion would result in complete attention on the game with no attention to distractors. Reduced immersion would thus lead to reduced attention on the game and more attention paid to distractors. To investigate this, I calculated correlations between the immersion score and number of distractors recognised but found no significant correlation in either condition. The *Full game* condition does not have a significant correlation between the immersion score and number of symbols remembered. The low number of symbols remembered and low correlation with the immersion score suggests that even those participants who were not particularly immersed in the game were still sufficiently distracted so they did not remember any distractors. In the *Bad game* the correlation between immersion and distractors remembered is higher but still not significant. If the distractors had been more distracting, then there might have been a stronger correlation between symbols remembered and immersion score. This would allow the distractors remembered count to discriminate more finely between the different experiences that players have of the game.

Limitations

These two games were very different which produced large differences in engagement. The large difference in immersion levels (η_p^2= 0.530) indicate that participants were particularly un-immersed in the bad game activity. One of the main reasons for measuring game experience is to compare different game designs. The *Bad game* has had so many game elements removed that it cannot really be considered a game. Subsequent experiments compared more similar games with smaller design differences to investigate differences in their ability to hold players' attention.

Participants only played the games for a short period of time. Participants performed both the games in this study for 5 minutes. This is a much shorter amount of time than the typical self-paced game would be expected to hold a player's interest. A longer study might be more ecologically valid and produce more distraction. However, participants were already reporting that they were bored by the *Bad game* so initial future work should probably focus on the level of distraction rather than making the task longer.

It is also possible that some participants were distracted by the distractors but did not remember them afterwards. Subsequent experiments will use eye tracking to measure

the amount of time they spent looking at the distractors and activity could provide an alternative measure of attention which should be investigated.

All participants saw the same distractor at the same time for both conditions. There is no evidence that participants remembered one symbol more than another but if they did this could distort the data if all participants saw a particularly memorable symbol at the same point in the game. For subsequent studies the order of distractors was randomised for each participant.

6.4. Experiment attention 3: More similar games with eye tracking

One of the main limitations of the last experiment was that the two games compared were very different. A useful measure of engagement would be able to differentiate between more similar game experiences. So, this next experiment used three different games which were more similar to each other. These were the *Full game,* the *No goals game* and the *All dots the same* game (see chapter 3). The other main difference from the previous experiment is that this one also used eye tracking to determine how much participants were looking at the game and how much they were looking at the distractors. This experiment was also used for a separate analysis to investigate measuring cognitive load using pupil dilation which is described in section 5.7.

Aims

This experiment aimed to see whether visual distractors could be used as a measure of attention for more similar games. It also aimed to see whether tracking participants' gaze could be used as a measure of attention.

Hypotheses

The hypotheses of the experiment were

1) The number of distractors that participants remember will be higher in the two game variants than in the full game.
2) The amount of time that participants spend looking at the game rather than the distractors will be higher for the full game rather than the game variants.
3) The Immersion score of the two game variants will be less than that of the full game.

6.4.2. Method

Design

This was a between-subjects design with three conditions. The independent variable was the game each participant played – either the *Full game,* the *No goals* game or the *All dots the same* game. The main dependent variable was the number of distractors that participants recognise after the activity. Another dependent variable was percentage of time that participants were fixated on the central game area rather than the surrounding

distractors. The final dependent variable is the Immersion Experience Questionnaire (IEQ) score for each participant's experience of the game.

Participants

I performed a power calculation to estimate how many participants to have in the study. The effect size of the previous experiment was η_p^2= 0.475. This is equivalent to a Cohen's f of 0.9512. I expected this experiment to have a smaller effect size as the games were more similar. So, I divided the previous f value by 2 to give an expected f of 0.4756. Using this effect size in a power calculation with a power of 0.8 (80%), an alpha of 0.05 and 3 conditions gives 15.24 participants per condition which I rounded up to 16 for each condition.

48 students and staff from the University of York took part in the study. 24 were male. Ages ranged from 18 to 62 (mean = 22.3). Game experience and attitudes varied between the participants, ranging from those who played games less than once per month, to those who played several times a week. In the previous experiments participants were paid in chocolate but this made it more difficult to recruit them in sufficient numbers. So, in this and subsequent experiments all participants were paid £6.

Materials

Participants played one of three games, either the *Full game, No goals game or All dots the same game.* The size of their pupils was also recorded for an experiment discussed in section 5.7. To reduce possible noise from light changes I used the monochrome versions of the game (see chapter 3). All three versions of the game were surrounded by distractor symbols as shown in Figure 45. The experiment used the same distractor recognition paradigm (see section 6.1) as in the previous study. The only difference in the distractors from the previous study was that each participant was shown a random set of 60 distractors chosen from a set of 90 and presented in a random order.

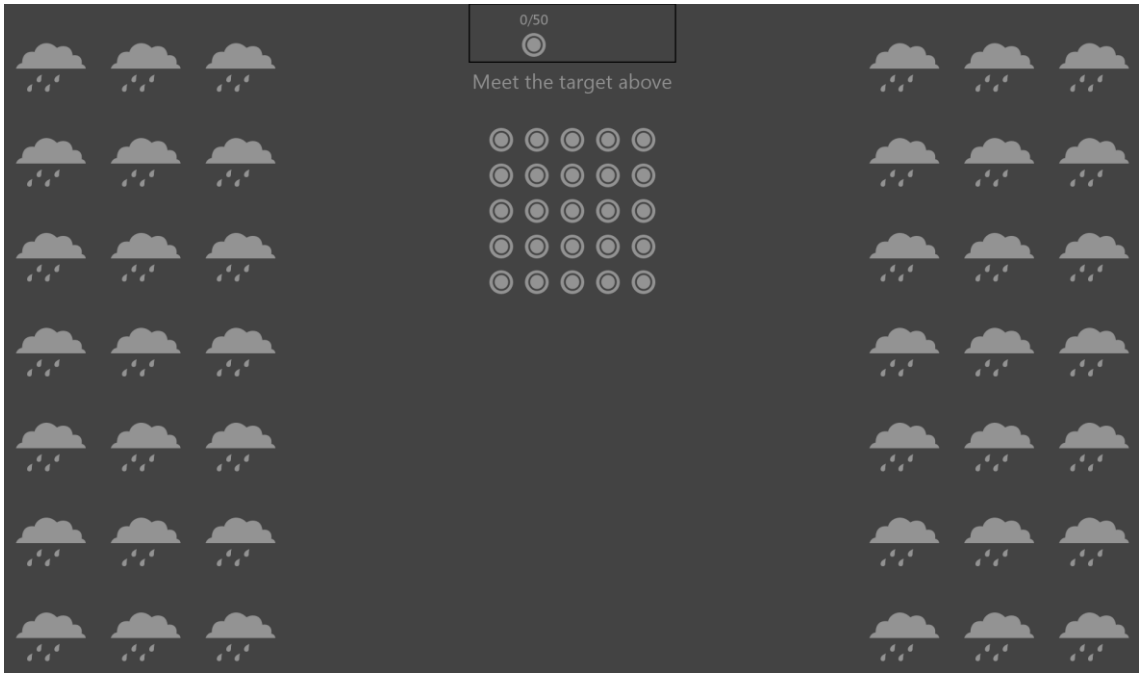

Figure 45 The *All dots the same* condition surrounded by distractors

6.4.3. Experimental setup

The experimental setup was the same as for the previous experiment as described in the distractor recognition paradigm (see section 6.1).

6.4.4. Procedure

Participants began by completing a consent form. They were then configured with the eye tracker (see chapter 3Experimental setup). After that the procedure was exactly the same as for the previous experiment.

6.4.5. Results

Distractor recognition

There was a significant difference in the number of correct distractors recognised between the *Full game* (M=14.13, SD=3.00), the *No goals* game (M=16.94, SD= 2.93) and the *All dots the same* game (M=16.31, SD=2.60) conditions; F(1,46)= 4.336, p=0.019, η_p^2= 0.162. A boxplot shows one outlier in the *No goals* condition. Removing this outlier still gives a significant difference between conditions with a stronger effect size; F(1,44)= 5.691, p=0.006, η_p^2= 0.206

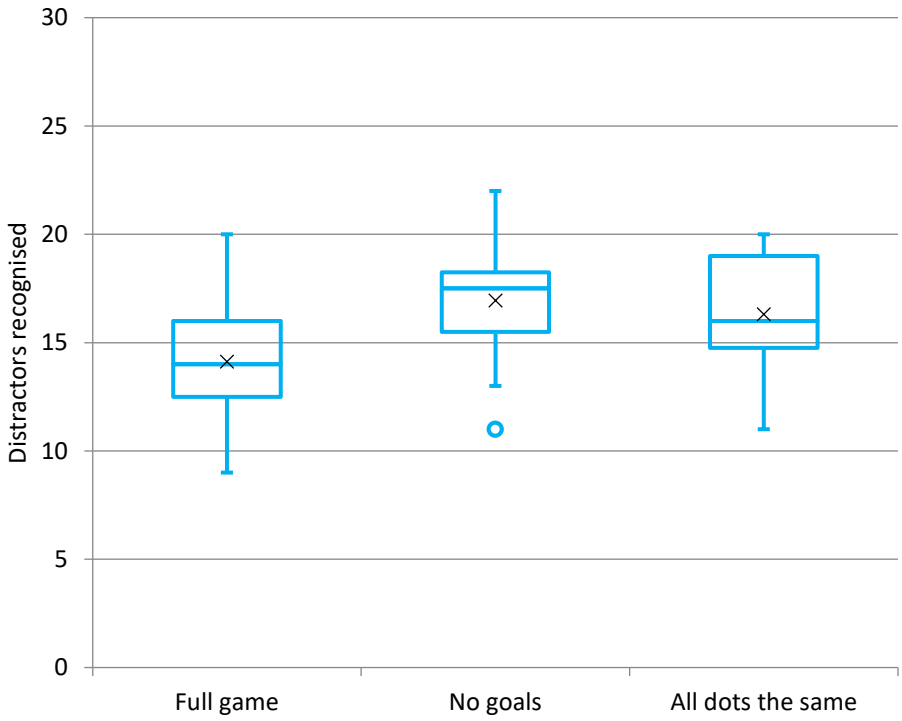

Figure 46 Boxplot showing the number of distractors remembered for all three conditions

I performed a Tukey's HSD post-hoc test to investigate which conditions were significantly different from each other.

Condition	*No goals*	*All dots the same*
Full game	p=0.020	p=0.085
No goals		p=0.808

Table 47 Tukey's HSD comparing distractors remembered for all conditions

This shows that there was a significant difference between the *Full game* and the *No goals* game. The difference between the *Full game* and the *All dots the same* game was not significant but tended towards significance (p=0.085). There was no significant difference between the *No goals* and *All dots the same* games. I compared the number of times that each different distractor was successfully recognised (see Figure 47, Figure 48 and Figure 49) and found no consistent difference between the different symbols. I also looked at number of distractors remembered across time for each different condition and found no definite pattern (see Figure 50).

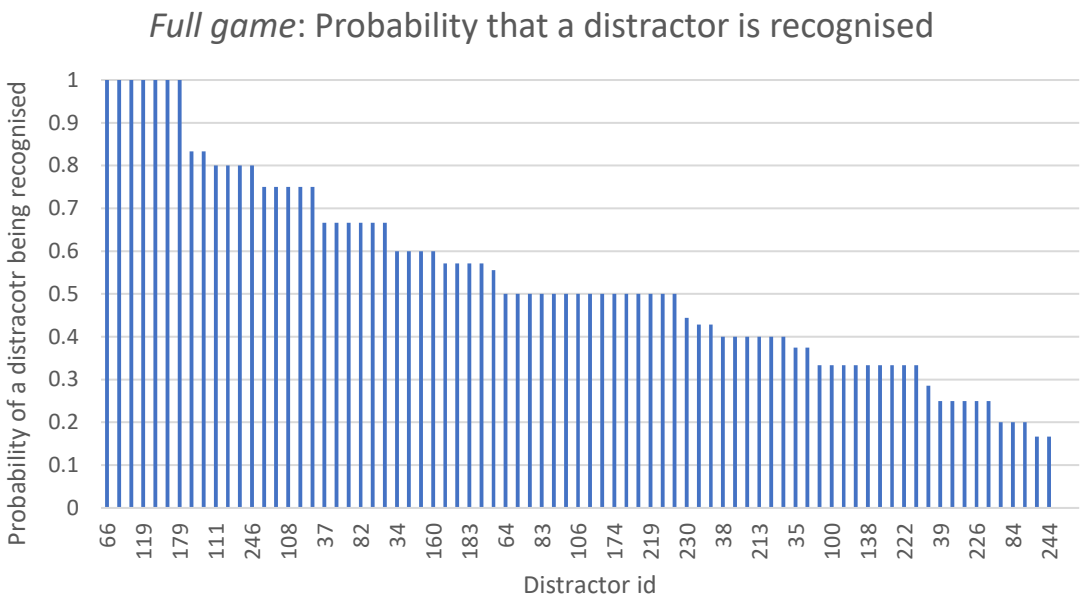

Figure 47 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest

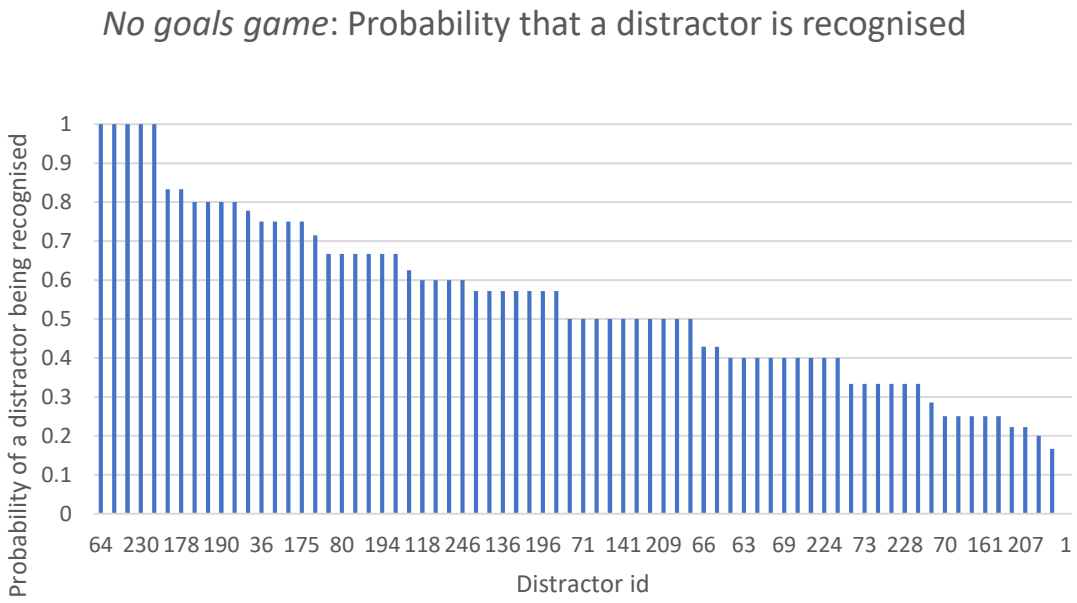

Figure 48 The probability that each different distractor shown in the *No goals game* will be recognised. The distractor times are ordered by probability – highest to lowest

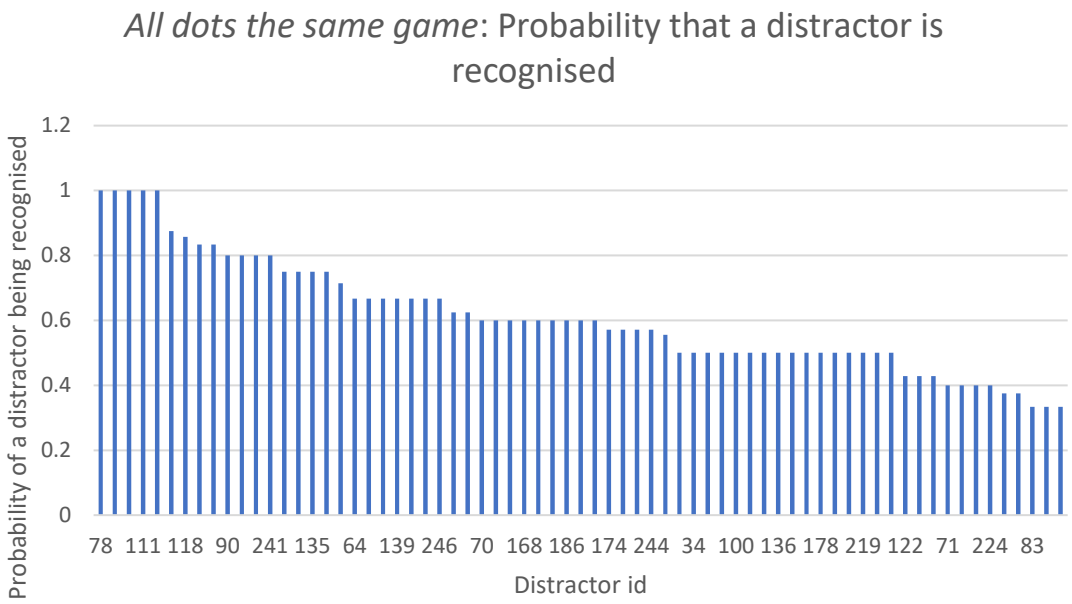

Figure 49 The probability that each different distractor shown in the *All dots the same* game will be recognised. The distractor times are ordered by probability – highest to lowest

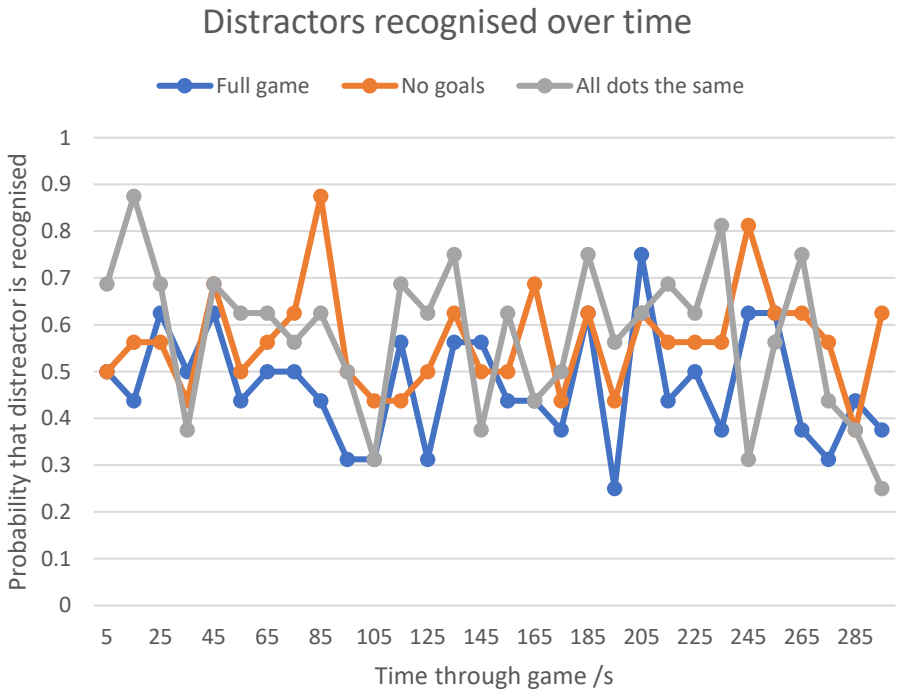

Figure 50 The probability that a distractor will be recognised over the time of the game.

Correlations

There was not a significant Pearson's correlation between the number of distractors recognised and the immersion score for either the *Full game* (r= -0.053, t(14)= -0.200,p=

0.844) the *No goals game* (r= 0.315, t(14)= 1.240, p = 0.235) or the *All dots the same game* (r= 0.205351, t(14)= 0.785, p= 0.446).

I also calculated Pearson's r correlations between the number of distractors recognised and the *Real World Dissociation* component of the IEQ and found no significant correlations for either the *Full game* (r=-0.009, t(14) = 0.034, p=0.973), the *No goals game* (r=0.149, t(14) = 0.563,p= 0.582) or the *All dots the same game* (r= 0.250, t(14)= 0.965, p = 0.351).

Eye tracking

I defined an Area of Interest (AOI) which filled the middle third of the screen that contained the game but not the distractor images.

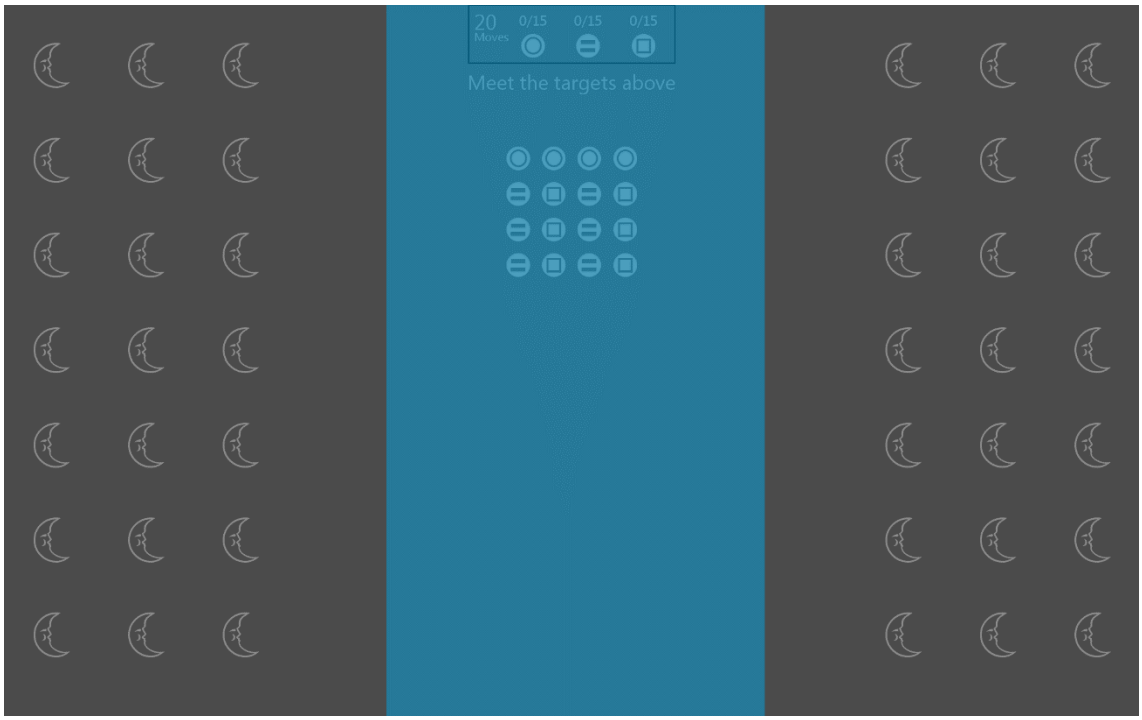

Figure 51 Game screen with distractors. The Area of Interest is shown in blue taking up the middle third of the screen. (This AOI is not visible to participants)

I then calculated the percentage of time that participants spent fixated at the central area rather than looking at the distraction images. The difference between conditions was not significant; the *Full game* (M=98.42, SD=2.70), the *No goals* game (M=98.47, SD= 2.43) and the *All dots the same* game (M=96.20, SD=3.99); F(2,46)= 2.754, p=0.074, η_p^2= 0.109.

I was concerned that this analysis may be affected by a ceiling affect as the maximum value is 100% and many of the values are very close (less than 1 standard deviation) to this value. So I changed the size of the interest area to be the middle quarter of the screen. Once again the difference between conditions was not significant; the *Full game*

(M=97.41, SD=4.46), the *No goals* game (M=98.30, SD= 2.28) and the *All dots the same* game (M=95.35, SD=4.83); F(2,46)= 2.249, p=0.117, η_p^2= 0.091.

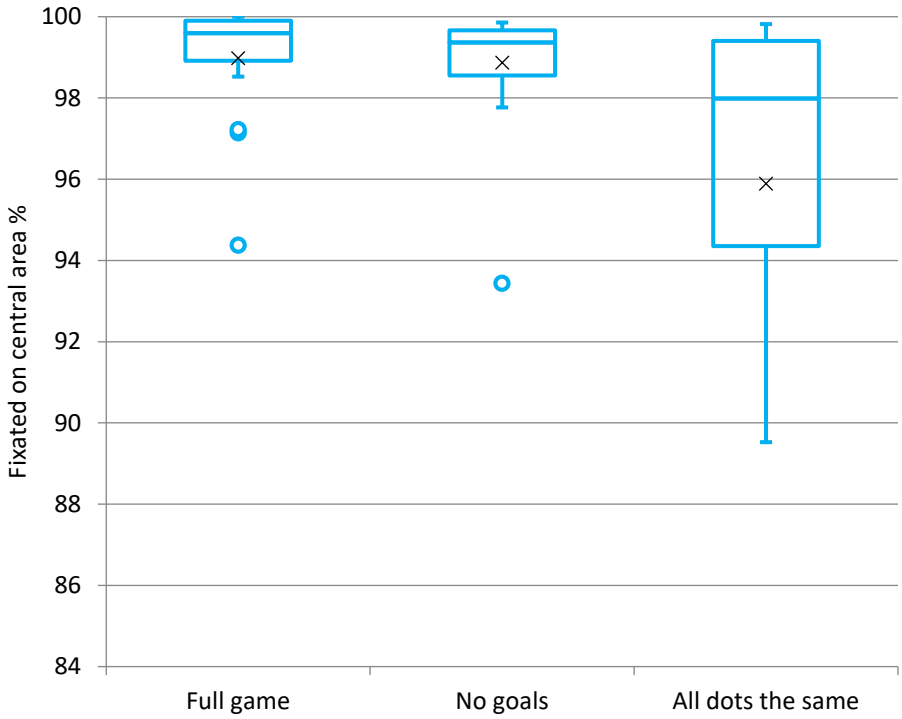

Figure 52 Percentage of time that participants fixated on the central area containing the game

Immersion Experience Questionnaire (IEQ)

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was a significant difference in the immersion scores between the *Full game*
(M=107.13, SD=17.00), the *No Goals* game (M= 93.38, SD= 14.05) and the *All dots the same* game (M=93.56, SD=13.38) conditions; F(2,45)= 4.492,p=0.017, η_p^2= 0.166.
The *All dots the same* game had three outliers. I repeated the analysis without these outliers to make sure that they were not skewing the results too much. There was still a significant difference in the immersion scores between the *Full game* (M=107.13, SD=17.00), the *No Goals* game (M= 93.38, SD= 14.05) and the *All dots the same* game (M= 91.46, SD=6.11) conditions; F(2,42)=6.051, p=0.005, η_p^2= 0.224.

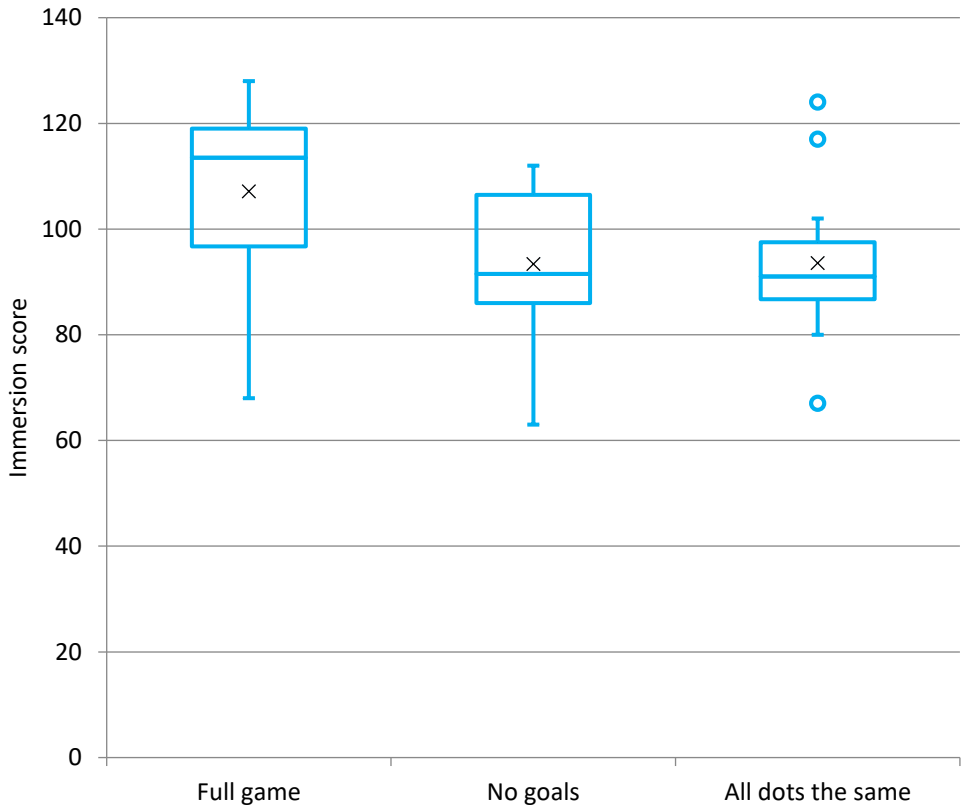

Figure 53. Boxplot showing IEQ scores for all three conditions

I performed the Tukey's HSD post-hoc test to investigate which conditions were significantly different from each other.

Condition	*No goals*	*All dots the same*
Full game	p=0.032	p=0.039
No goals		p=0.999

Table 48 Post-hoc test on IEQ scores between three game variants

This shows that there was significant difference between the *Full game* and the *No goals* game. There was also a significant difference between the *Full game* and the *All dots the same* game. There was no significant difference between the *No goals* and *All dots the same* games.

	Full game **Mean (SD)**	**No goals** **Mean (SD)**	**All dots the same Mean (SD)**	**Effect size** η_p^2	**Significance (p value)**	**F(3,46)**
Cognitive involvement	33.94 (5.77)	32.06 (5.54)	29.71 (4.65)	0.101	0.091	2.533
Emotional involvement	17.88 (5.43)	13.81 (4.10)	14.12 (3.60)	0.161	0.019	4.322
Real world dissociation	32.38 (5.54)	28.75 (5.62)	31.12 (4.65)	0.074	0.179	1.786
Challenge	14.38 (1.5)	11.06 (2.43)	11.18 (3.51)	0.315	<0.001	10.349
Control	16.06 (3.77)	14.56 (3.10)	15.35 (2.48)	0.039	0.411	0.906
Immersion	107.13 (16.97)	93.38 (14.05)	94.24 (13.66)	0.166	0.017	4.492

Table 49 A comparison of the different components of the IEQ across game conditions

Correlations

I investigated whether there was a correlation between the IEQ scores and the number of distractors recognised for each condition. For each condition I calculated the Pearson's correlation coefficient (r).
Full game: r=-0.053, t(14) = -0.200, p = 0.844
No goals: r= 0.315, t(14)=1.240, p = 0.235
All dots the same: r=0.205, t(14) = 0.785, p = 0.446

There were no significant correlations between IEQ and distractors recognised for any of the conditions.

6.4.6. Discussion

The first hypothesis that participants would remember more distractors for the game variants than the *Full game* was supported. A post hoc analysis showed a significant difference between the *Full game* and the *No goals* game. The difference between the *Full game* and the *All dots the same* game approaches significance (p=0.085) but there was no difference between the two reduced variants of the game. There was a large effect size between conditions (η_p^2= 0.162).

The second hypothesis that participants would fixate more on the distractors during the game variants was not supported although it does approach significance (p=0.074). There was a moderate effect size (η_p^2= 0.109). Participants fixated on the distractors more during the *All dots the same* game than during the other two games. This suggests that participants may be more easily distracted from the *All dots the same* game but as their mental attention is still focused on reaching the target, they do not remember any more distractors.

The third hypothesis that levels of immersion would be lower in the two game variants was supported. A post-hoc analysis showed a significant difference between the *Full game* and both other games. There was no significant difference between the *No goals* and *All dots the same* game variants. This suggests that both challenge and uncertainty may contribute equally to immersion in games. This is a surprise because some researchers such as Denisova et al. (2017) have seen challenge as the most important feature of digital games. This may be because although the *All dots the same* game nominally has challenge because players are set a target to reach, players do not find this challenge meaningful (Sinclair et al., 2007) as it does not relate to their skill level. There was a large effect size between conditions (η_p^2= 0.166) which is of similar magnitude to the effect size between distractors remembered (η_p^2= 0.162). The IEQ can be subdivided into five different components. Of these only *Emotional involvement* and *Challenge* are significantly different between conditions. Of the other non-significant conditions *Control* is similar between the different games as players use the same method of controlling the game and the questions for *Real world dissociation* are not as meaningful for 2d self-paced games. I was initially surprised that *Cognitive involvement* is not significantly different between conditions as *Two Dots* is a puzzle game, however this may support the suggestion from previous chapters that players can have a significantly different game experience without experiencing differences in the cognitive load required to play the game.There were no significant correlations between IEQ and the number of distractors remembered for any of the conditions. This may be because the variation in the number of distractors remembered is very small which does not give enough range to create a significant correlation.

These results suggest that visual distractors can be used to measure the differences in attention between more similar games. However, they also suggest that incorporating eye tracking does not provide a more accurate measure. Eye tracking does provide some differentiation between game conditions but the effect is smaller than distractor recall. This is probably because there are many levels of attention (Rensink, 2015) and although participants eyes may be looking at the distractors their mental attention is still partly on the game and they do not remember any more distractors.

Limitations

In the introduction to this chapter I discussed how one of the advantages of an "on-line" measure such as distractor recognition over "post-game" measures such as questionnaires is that, in theory, they could be used to measure the changes in experience over the period of a game. For this experiment there are no obvious changes over time in any of the conditions. This may be because the game experience was similar for the full length of the gameplay time or it could be because the distractors recognition measure was not sensitive enough to pick up the patterns that were there. Although the effect size for differences in distractor recognition is comparable to that of the immersion questionnaire it is much lower than in the initial experiment. From the players' point of view the games are very different, and it may be possible to make better measure of experience which reflects this difference with a stronger effect size.

Another limitation is that the eye tracking equipment added occasional pauses into the game which may decrease immersion. The eye tracking equipment also requires considerable setup and configuration for each participant. At no point do I mention that participants' eyes are being tracked (I told them that the camera was there to measure changes in the size of their pupils) but they are well aware that I am tracking where they are looking which may make them more likely to obey the instructions to play the game, rather than looking at the distractors. Some participants told me "I saw the distractor images but thought they were designed to put me off the game so I tried to ignore them".

6.5. Experiment attention 4: More distracting distractors memory test

The previous experiment was largely successful in its aims to measure game experience using the distractor recognition paradigm. For the next experiment I looked at whether this paradigm could be improved to make a better measure of experience. The previous experiment produced a significant difference in the number of distractors recognised in each condition. However, in the *Full game* condition the number of distractors that participants recognise is very low, the same as they would recognise by chance. This puts a floor effect on the number recognised and may reduce potential sensitivity and the effect size of this measure. I considered that one way of reducing this floor effect this would be to make the distractor images more interesting which may help participants recognise more of them. The previous experiments used *Webdings* icons which are fairly abstract and bland, so players may not find them more interesting than the game. Any new set of distractor images had to contain at least 90 different images, 60 for the distractors surrounding the game and 30 for the dummy answers to the test. I decided to use images of Disney characters because they're bright and interesting and participants may even have some emotional reaction to them depending on which films that they have seen. There are also a large number of different characters so it was possible to get 90 of them from Disney fan sites on the internet. Before using them in a distraction experiment, I wanted to see whether they were more memorable than the *Webdings* icons. Being more memorable might make the distraction experiment more sensitive but being less memorable could have the opposite effect.

Aims

The main aim of this experiment was to validate the use of Disney characters rather than *Webdings* characters in the distractor recognition paradigm. To validate this change in the paradigm I wanted to find out how many Disney distractor images participants could recognise. As with the previous validation experiment (See section 6.2) this sets a baseline for the number which could be expected if players' attention was completely distracted away from the game.

Method, Materials and Procedure

The methods and materials for this experiment were the same as the previous validation experiment (See section 6.2). The materials were identical except that the distractors were Disney characters rather than Webdings icons. The procedure was identical except that participants competed the immersion questionnaire on a 10″ iPad rather than on paper. This speeded up analysis of data and there is no evidence that it made a difference to the results obtained.

Figure 54 Disney characters replacing the *Webdings* distractors. Participants were instructed to ignore the text in the centre of the screen and not given a mouse or keyboard to interact with the computer. © Disney Corporation

Participants

10 staff and students from colleges in York took part in the study. 9 were male, 9 were native speakers of English with ages ranging from 21-52 (Mean 26.7)

Results

The mean number of distractor images recognised was 27.6 out of 30 (SD=2.07). This is equivalent to remembering the 92% of the images. For comparison the previous experiment found that the mean number of *Webdings* icons recognised was 27.5 out of 30 (SD=2.88) which is equivalent to remembering 91.6% of the images.

I plotted the probability that each distractor would be correctly recognised which can be seen in Figure 55.

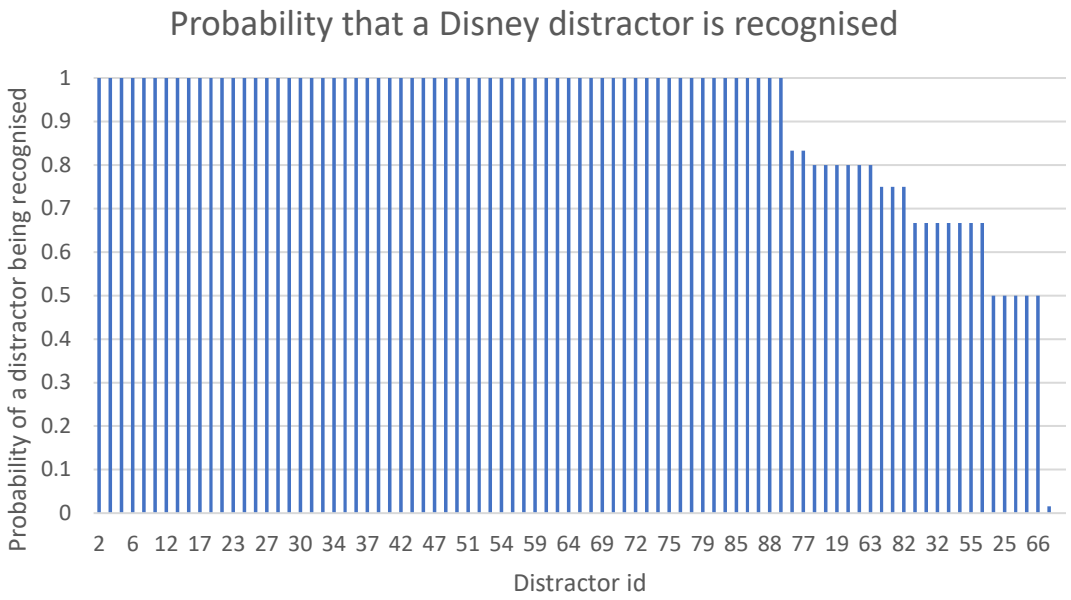

Figure 55 The probability that each different distractor will be recognised. The distractor ids are ordered by probability – highest to lowest.

I also plotted the probability that each distractor would be recognised against the time through the experiment. This is shown in Figure 56.

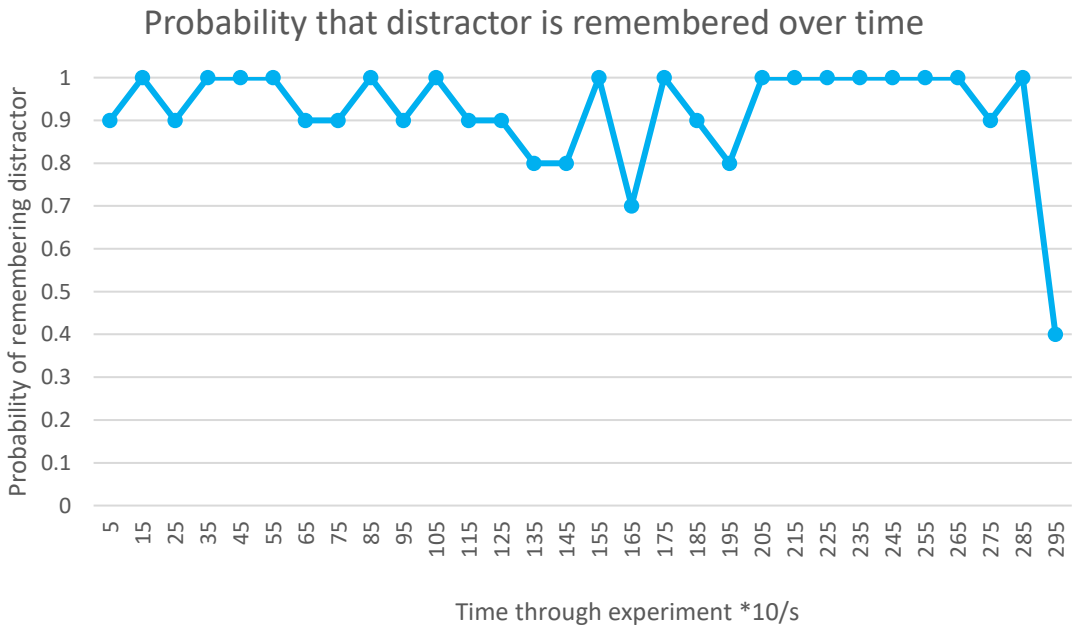

Figure 56 The probability that a distractor will be recognised against the time it is displayed

6.5.2. Conclusions

There is no apparent difference in how memorable Disney distractors are compared to *Webdings* icons. Participants in this experiment were older than in the previous icons experiment (Mean 26.6 years compared to 20.3) which may have made some difference

but it seems likely that the memorability of both sets of distractors is very close. As with the previous validation experiment, I was interested to see whether particular distractors were more memorable than others. I plotted both the recognition rates for each distractor and the recognition rates over time. Neither of these plots showed any evidence for particular images or times being more or less memorable.
Both this experiment and the previous validation experiment suffer from the limitation that they use a small number of participants (n=10) and so are not suitable for more sophisticated analysis which might pick up small differences in how memorable the different set of distractors were. However, these distractors will be used for experiments which also use a small number of participants and expect large effect sizes so even if there is a small difference between the distractors it is unlikely to make a difference to those results.

The high rate of recognition of Disney characters which is almost the same as that for *Webdings* characters suggest that Disney characters may give similar results to *Webdings* characters when used to measure game attention in a distractor recognition task. The next experiment was designed to investigate this in more detail.

6.6. Experiment attention 5: Three games with Disney distractors

The aim of this experiment was to investigate whether using Disney characters as distractors would make a better measure of game attention. The previous experiment showed that Disney character were no more or less memorable than *Webdings* icons. However, it is possible that the bright colours and more interesting characters would make them more distracting. Participants may be more tempted to look away from the game to see the Disney characters because they may be more interesting than the game. I therefore repeated the previous experiment (see section 6.4) except that I used Disney characters as distractors rather than *Webdings* icons.

The hypotheses, aims, methods and procedure were identical. The materials were identical except that the distractors were Disney characters as shown in Figure 57. Participants played the game whilst being recorded by the eye tracker as before. This experiment was also used for a separate analysis to investigate measuring cognitive load using pupil dilation which is described in section 5.12.

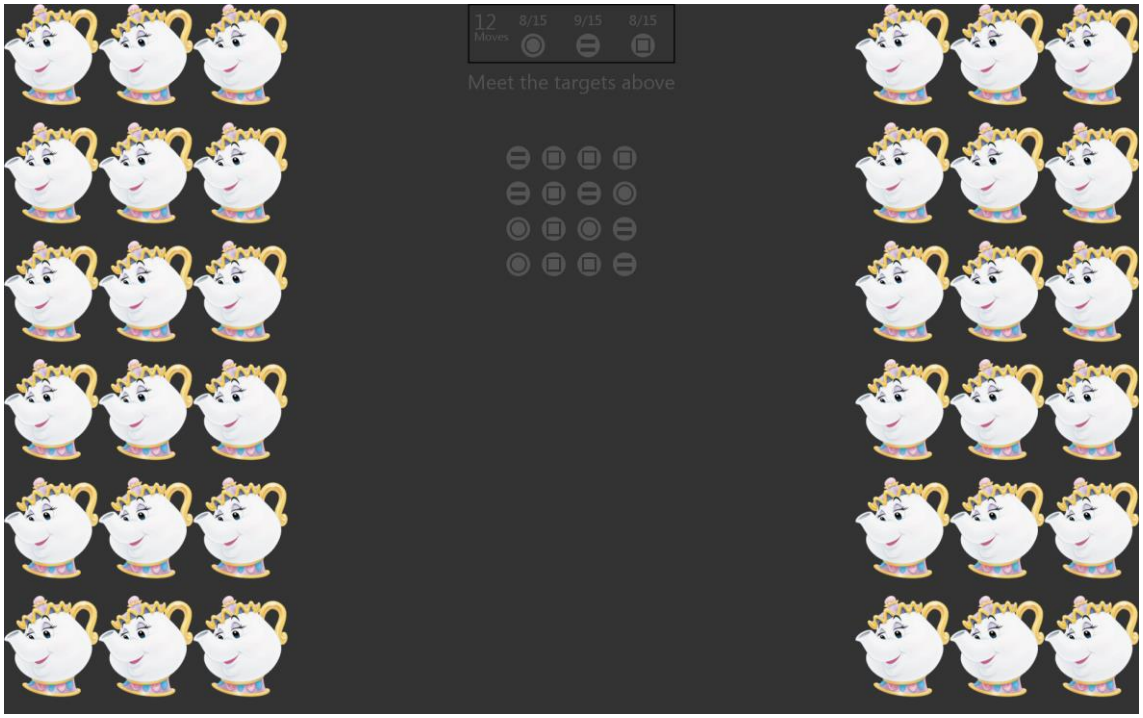

Figure 57 *Two Dots* game surrounded by Disney distractors

6.6.1. Participants

I performed a power calculation to estimate how many participants to have in the study. In the previous distraction experiment the difference in the number of distractors remembered had an effect size of η_p^2= 0.162. This is equivalent to a Cohen's f of 0.4397. I was hoping that this experiment would have the same or a larger effect size than the previous one. Using this effect size in a power calculation with a power of 0.8 (80%), an alpha of 0.05 and 3 conditions gives 17.65 participants per condition which I rounded up to 18 for each condition.

54 participants took part in this experiment (18 in each condition). Ages ranged from 18-42 (Mean 21.22). 32 were male and 48 were native speakers of English.

6.6.2. Results

Distractor Symbols

There was not a significant difference in the number of correct distractors remembered between the *Full game* (M=16.78, SD=2.98), the *No goals* game (M=16.67, SD= 2.93) and the *All dots the same* game (M=18.67, SD=3.71), conditions; F(2,52)= 2.184, p=0.123, η_p^2= 0.079. The *All dots the same* condition has 4 outliers. Removing these outliers from the *All dots the same* game (M=18.43, SD=1.604) gives a very similar result which is also not significant. *F*(2,49)= 2.095, p=0.134, η_p^2= 0.082.

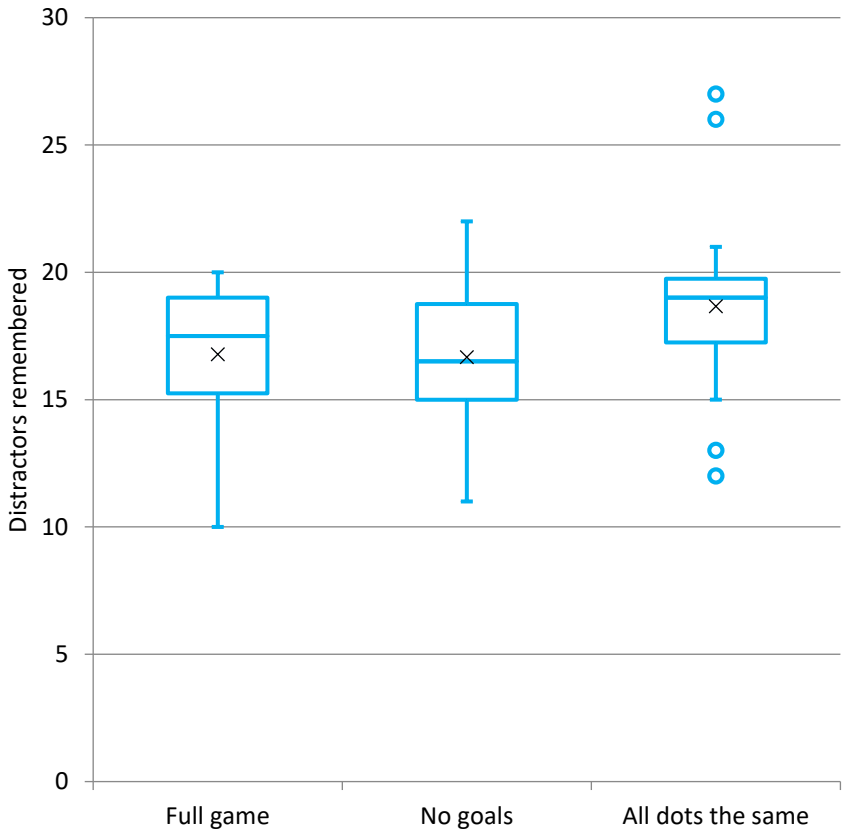

Figure 58 Boxplot showing the number of distractors remembered for all three conditions

I compared the number of times that each different distractor was successfully remembered (see Figure 59, Figure 60 and Figure 61) and found no consistent difference between the different symbols. I also looked at number of distractors remembered across time for each different condition and found no definite pattern (see Figure 62).

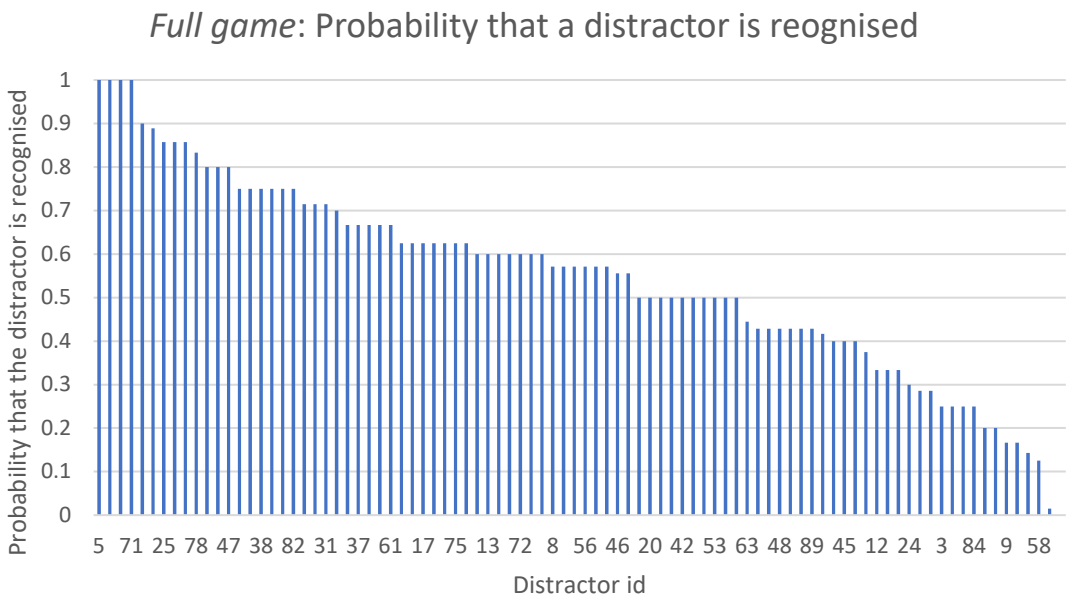

Figure 59 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest

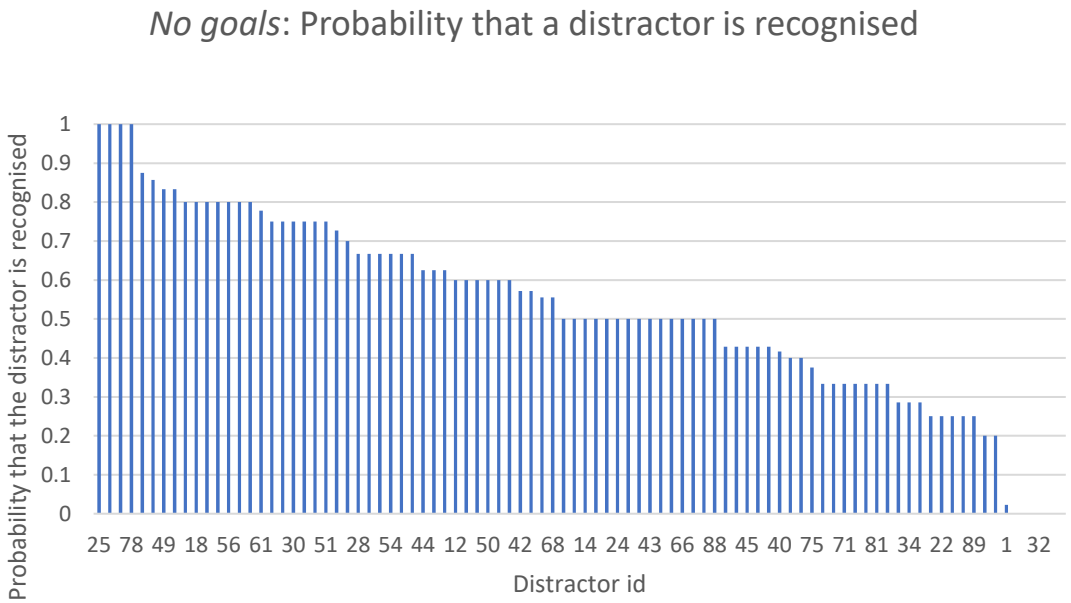

Figure 60 The probability that each different distractor shown in the *No goals* will be recognised. The distractor times are ordered by probability – highest to lowest

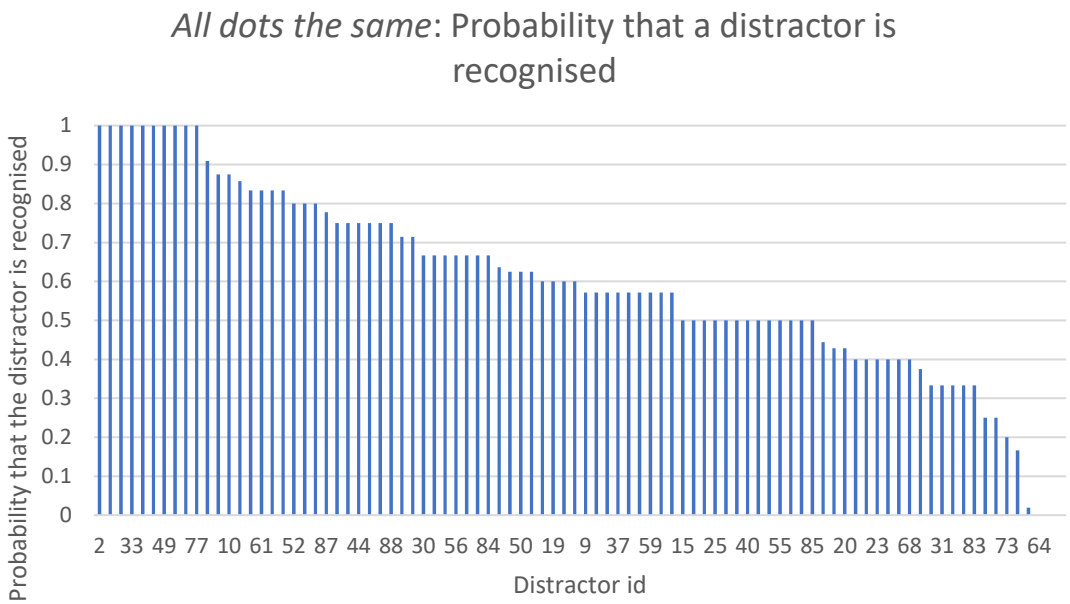

Figure 61 The probability that each different distractor shown in the *All dots the same* game will be recognised. The distractor times are ordered by probability – highest to lowest

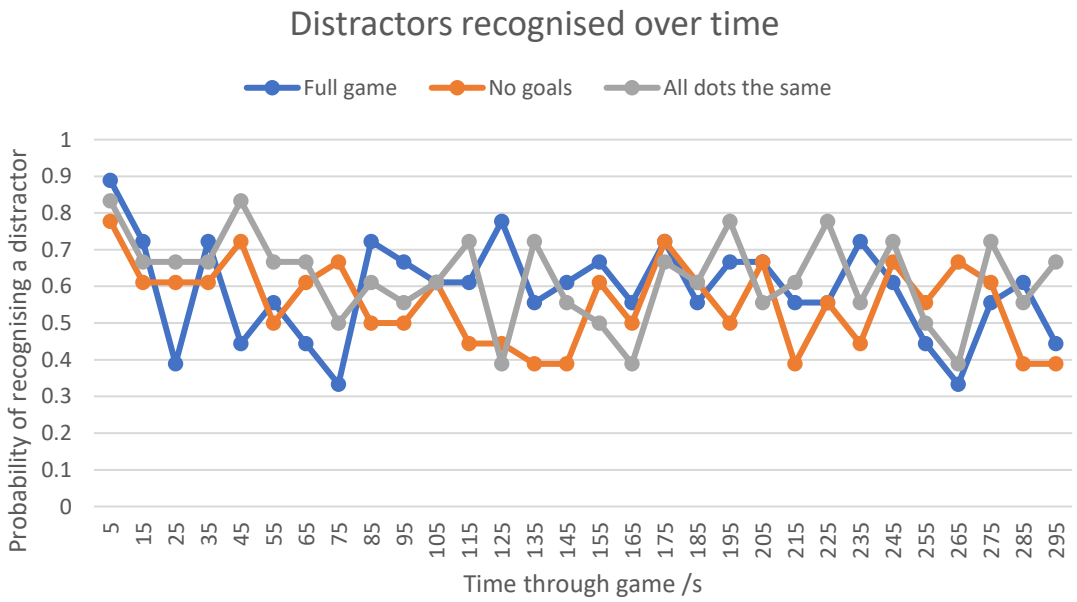

Figure 62 The probability that a distractor will be recognised against the time it is displayed

Immersion Experience Questionnaire

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was not a significant difference in the immersion scores between the *Full game* (M=101.94, SD=15.02), the *No Goals* game (M= 98.94, SD= 14.54) and the *All dots the same* game (M=89.61, SD=18.90) conditions. However, there was a trend towards significance (p=0.069). $F(2,45)= 2.811, p=0.069, \eta_p^2= 0.099$.

	Full game Mean (SD)	**No goals** Mean (SD)	**All dots the same** Mean (SD)	**Effect size** η_p^2	**p value**	**F(3,52)**
Cognitive involvement	32.83 (5.02)	32.44(4.25)	27.56 (6.68)	0.173	0.008	5.318
Emotional involvement	15.17 (4.73)	14.89 (3.72)	13.28 (4.79)	0.036	0.394	0.949
Real world dissociation	33 (5.57)	31.72 (6.29)	31.11 (6.59)	0.017	0.646	0.440
Challenge	13 (2.43)	11.56 (2.26)	9.56 (3.19)	0.231	0.001	7.653
Control	15.5 (2.94)	16 (3.03)	14.94 (3.70)	0.018	0.623	0.478
Immersion	101.94 (15.02)	98.94 (14.55)	89.61 (18.903)	0.099	0.069	2.811

Table 50 The Immersion scores and sub-factors for each of the three conditions

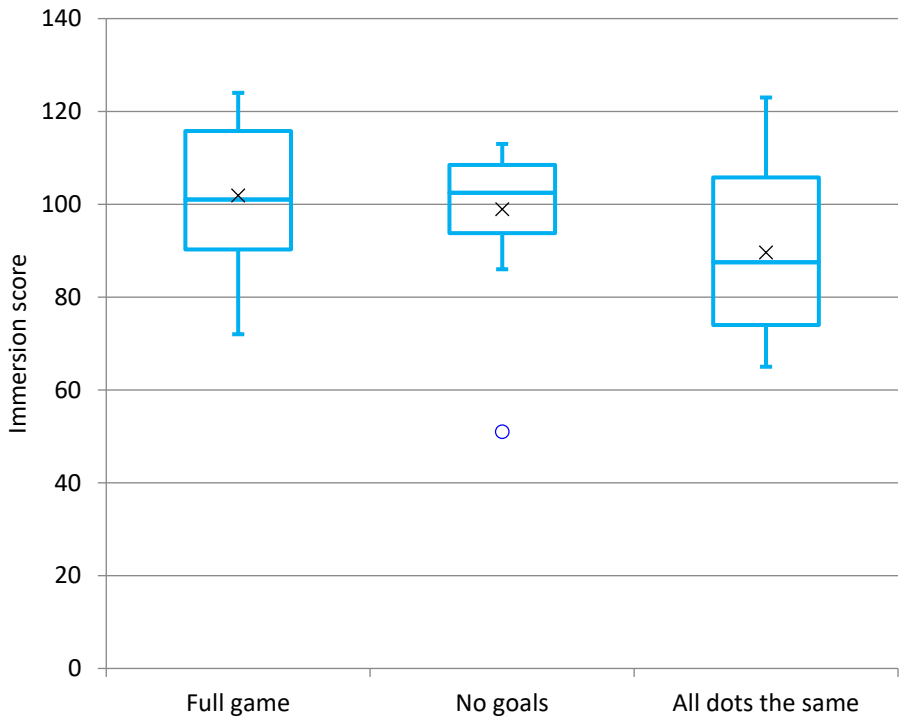

Figure 63 Boxplot showing the Immersion scores for all three conditions

Correlations

I investigated whether there was a correlation between the IEQ scores and the number of distractors remembered for each condition. For each condition I calculated the Pearson's correlation coefficient (r).

Full game: r= -0.116, t(16) = -0.467, p = 0.647

No goals: r= -0.341, t(16)= -1.452, p = 0.166

All dots the same: r=-0.314, t(16) = -1.323, p = 0.205

There were no significant correlations between IEQ and distractors remembered for any of the conditions.

Eye tracking

As in the previous experiment I defined an Area of Interest (AOI) which filled the middle third of the screen that contained the game but not the distractor images.

I then calculated the percentage of time that participants spent fixated on the central area rather than looking at the distraction images. The difference between conditions was not significant; F(2,52)= 1.074, p=0.349, η_p^2= 0.040.
There were four outliers in this analysis, in particular one participant in the *Full game* fixated on the central area for only 42% of the time. I removed these outliers and to see if they were skewing the result. This did not show a significant difference although there was a trend towards significance (p=0.053). F(2,49)= 3.122, p=0.053, η_p^2=0.117.
This analysis still faces the problem that that there is a ceiling affect which may skew the ANOVA. To investigate whether this was affecting the results I performed another post-hoc analysis. In this analysis I changed the Area of Interest so that it only covered the middle quarter of the screen. This removed the ceiling affect. This analysis did not show a significant difference between conditions. F(2,52)= 1.252, p=0.295, η_p^2=0.047.
Descriptive statistics for all three analyses are shown in Table 51.

	Mean (Standard deviation)		
Analysis	*Full game*	*No goals*	*All dots the same*
AOI is 33% of the screen	94.44 (12.25)	98.31 (1.38)	95.41 (5.15)
AOI 33% of the screen but without outliers	97.51 (2.66)	98.69 (0.85)	96.18 (4.09)
AOI is 25% of the screen	94.69 (0.12)	98.25 (0.017)	94.66 (0.06)

Table 51 Descriptive statistics for different approaches to analysing the eye tracking data

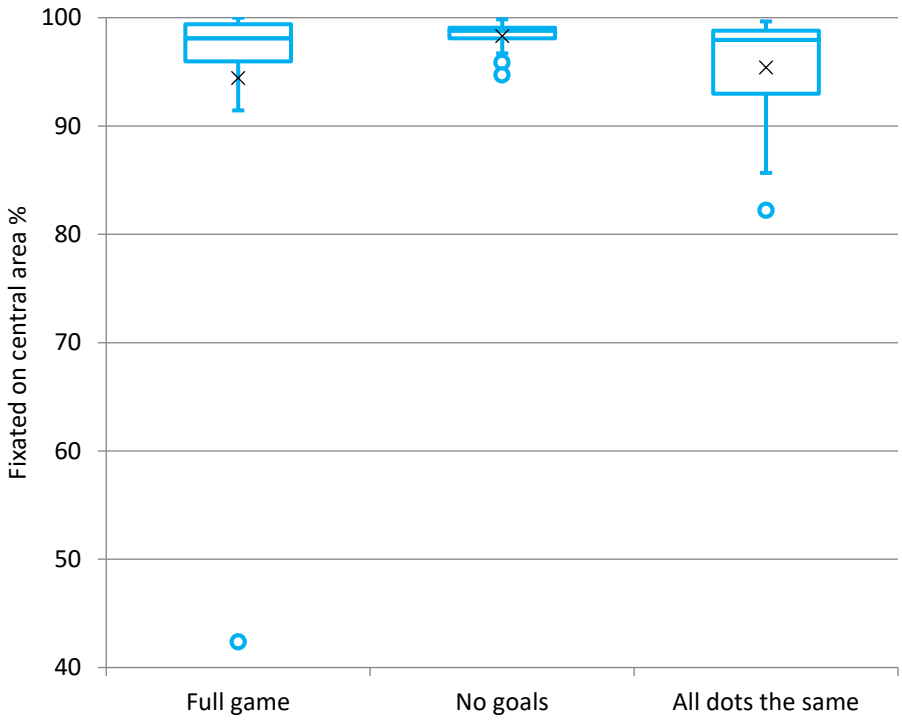

Figure 64 The percentage of fixations on the central game area for each condition

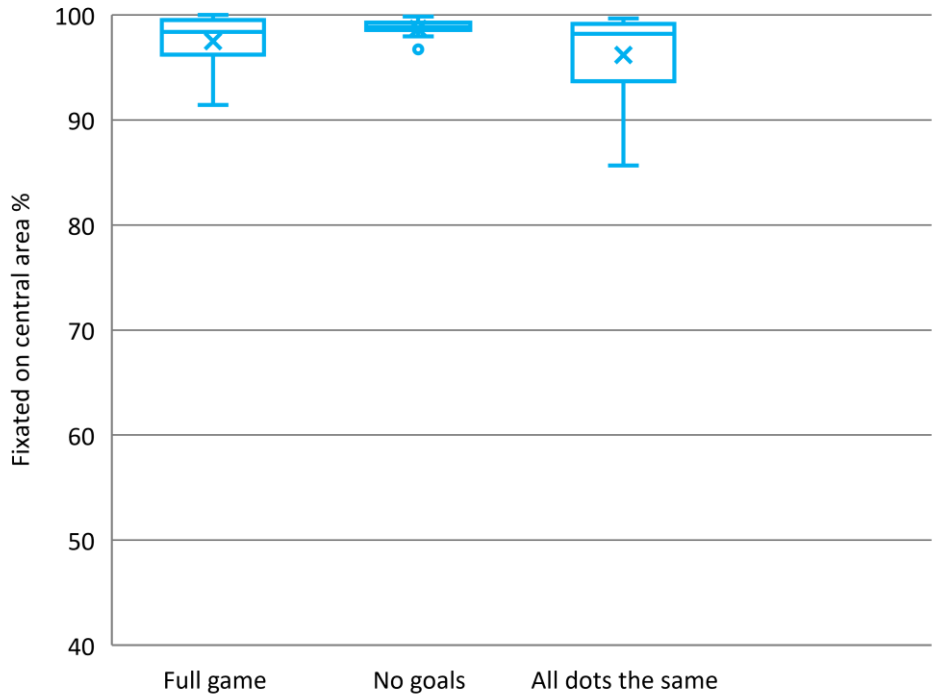

Figure 65 The percentage of fixations on the central area for each condition with the initial set of outliers removed

6.6.3. Comparisons with previous experiments

One of the goals of this experiment was to compare using Disney distractors with *Webdings* distractors. To do this I compared the results of this experiment with those *of 6.4 Experiment attention 2: Distraction* between two games. Table 52 compares the number of distractors recognised for each condition.

Condition	*Webdings* distractors	Disney distractors
Full game	14.13 (3.00)	16.78 (2.98)
No goals	16.94 (2.93)	16.67 (2.93)
All dots the same	16.31 (2.60)	18.67 (3.71)

Table 52 Comparing the mean number of distractors recognised for *Webdings* distractors and Disney distractors (SD in brackets)

Table 53 compares the percentage of time that participants fixated on the Area of Interest which consisted of the middle third of the screen.

Condition	*Webdings* distractors	Disney distractors
Full game	98.42 (2.70)	94.44 (12.25)
No goals	98.47 (2.43)	98.31 (1.38)
All dots the same	96.20 (3.99)	95.41 (5.15)

Table 53 Comparing the mean percentage time fixated on the central game area for Webdings distractors and Disney distractors (SD in brackets).

Table 54 compares the immersion scores for each condition.

Condition	*Webdings* distractors	Disney distractors
Full game	107.13 (17.00)	101.94 (15.02)
No goals	93.38 (14.05)	98.94 (14.54)
All dots the same	93.56 (13.38)	89.61 (18.90)

Table 54 Comparing the immersion scores for the Webdings distractors and Disney distractors

6.6.4. Discussion

The motivation behind this experiment was that using more distracting distractors would mean that participants remember more distractors and create a greater difference between conditions. This may create a more sensitive measure of game attention than using *Webdings* icons as distractors.

The first hypothesis is that there would be a significant difference between the number of distractors remembered. This was not supported. Participants did remember more distractors in this experiment than the previous distraction experiment however the difference between conditions was smaller and not significant (η_p^2= 0.079). When speaking to participants after the experiment some of them mentioned that a particular

character was more memorable than others. However, they all mentioned a *different* character as being more memorable. Plotting the distribution of how memorable the different characters were shows no pattern of any character being more memorable than others other than you would expect by chance.

The second hypothesis, that participants would spend a significantly different amount of time looking at the game was not supported. Participants did fixate less on the central game area for all three games but the difference between conditions was not significant (η_p^2= 0.040).

The third hypothesis, that there would be a significant difference in immersion between games was also not supported (η_p^2= 0.099) although there was a trend towards significance (p=0.069). More distracting distractors also had an effect on the overall immersion which was reduced for two for the games although curiously it increased for the *No goals* game. I investigated to see whether there was a correlation between distractors recognised and immersion but found no significant correlations for any of the conditions.

It appears that the effect of the Disney distractors is to reduce the immersion of the *Full game* and the *All dots the same* game and make the overall experience of playing them more similar. This may be because players know they have goals to meet in those games and being distracted from those goals by the Disney characters makes this harder and reduces immersion. However, the immersion score for the *No goals* game is higher, this may be because freed from the need to reach goals participants see the Disney characters as just an extra source of engagement that they can look at if they like without interfering with the rest of the game. Having said that it is difficult to compare the results of two different experiments so these differences may just be down to random variation. Looking at the subcomponents of the IEQ shows that the differences in *Cognitive Engagement* and *Emotional Engagement* are the most reduced by adding Disney distractors. As in previous experiments plotting the recognition rates for each distractor icon showed no evidence that any of the icons was more or less likely to be recognised than any other.

In conclusion changing distractors to use Disney characters has had a similar effect on distractors recognised, immersion and eye movements which is to reduce the difference between these measures across conditions so reduce the resulting effect size. This is the opposite result to the aim of this experiment.

Limitations

This experiment had similar limitations to the previous icon distractor experiment such the awkwardness of the eye tracking equipment. Another limitation is the number of outliers in the data which may have skewed the analysis. Removing the outliers from the eye tracking data showed a trend towards significance in a post-hoc ANOVA (p=0.053) which compared conditions. Future experiments could investigate whether this is a reliable effect or a random artefact due to chance.

One limitation of the Disney distractors is that there are a number of reasons why they could be more distracting. They're all characters rather than things, they are colourful and participants may recognise them or have emotional reactions to some of them. It is difficult to say which of these factors make them more distracting and it is possibly a combination.

6.7. Experiment attention 6: Distraction test without eye tracking

This experiment builds on the previous experiments but makes changes to the experimental design to investigate the effect on the distractor recognition paradigm as a measure of game attention. During the previous distraction experiments, participants' eye movements were recorded using an eye tracker. This is an intrusive process involving a setting up and calibrating the eye tracker as well as a chin rest (See chapter 3 Experimental setup). During the experiment I never refer to eye tracking instead calling it "a camera to measure how your eyes react to light". However, most participants are well aware that the camera can tell where they are looking and this may make them reluctant to be distracted because they know that they have been told to play the game. As well as inhibiting distraction the whole eye tracking setup makes the experiment less like a real game playing experience and so reduces ecological validity.
Another factor which may affect participants' behaviour is the presence of the experimenter. During eye tracking experiments I sit in front of the eye tracker screen which is behind the participant. Participants almost certainly feel that they are being watched which may also make them less likely to be distracted. For this experiment I removed the eye tracking and left participants alone in the room to play the game. I hoped that this more relaxed setting would make participants feel more comfortable about being distracted from the less interesting games and create a larger effect size between conditions.

The hypotheses, aims, methods and materials were identical to the previous *Webdings* icons distractor experiment (See 6.4 Experiment attention 3: More similar games with eye tracking). The procedure differed as described below.

6.7.1. Participants

For this experiment I was hoping for an effect size which was larger or equal to the previous distraction experiment. Therefore, I used the same number of participants. 54 participants took part in this experiment (18 in each condition). Ages ranged from 18-28 (Mean 20.67). 27 were male and 51 were native speakers of English.

6.7.2. Procedure

Participants began by completing a consent form. They then started playing the game and the experimenter left the room. At the end of 5 minutes the game ends and sounds a "game over" message. The experimenter then returned to the room guided the participant through the recognition y test of distractors and the immersion experience questionnaire. The IEQ was performed on an iPad with a 10" screen.

6.7.3. Results

Distractor Symbols

There was not a significant difference in the number of correct symbols remembered between the *full game* (M=16.39, SD=3.35), the *No goals* game (M=16.44, SD= 2.26) and the *All dots the same* game (M=17.83, SD=3.35), conditions; F(2,52)= 1.316, p=0.277, η_p^2= 0.049

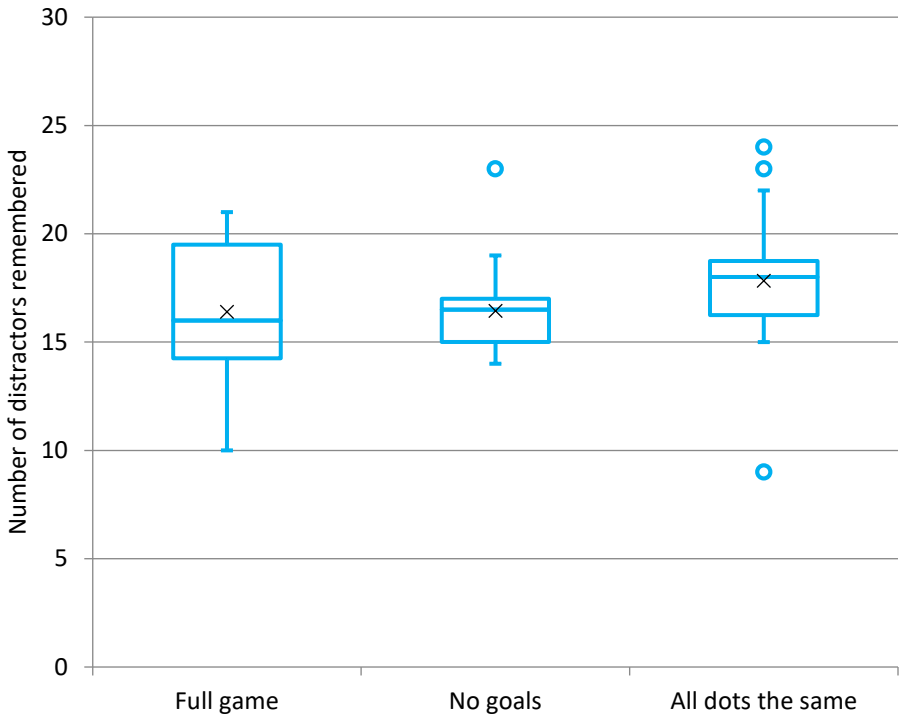

Figure 66 Boxplot showing the Immersion scores for all three conditions

The boxplot in Figure 66 shows that there are several outliers in the *No goals* and the *All dots the same* conditions. I was concerned that this might skew the analysis so I repeated the ANOVA without these outliers. There was still not a significant difference in the number of correct symbols remembered between the *full game* (M=16.39, SD=3.35), the *No goals* game (M= 16.06, SD=1.600) and the *All dots the same* game (M= 17.83, SD=1.84), conditions; F(2,48)= 1.912,p=0.159, η_p^2= 0.075).

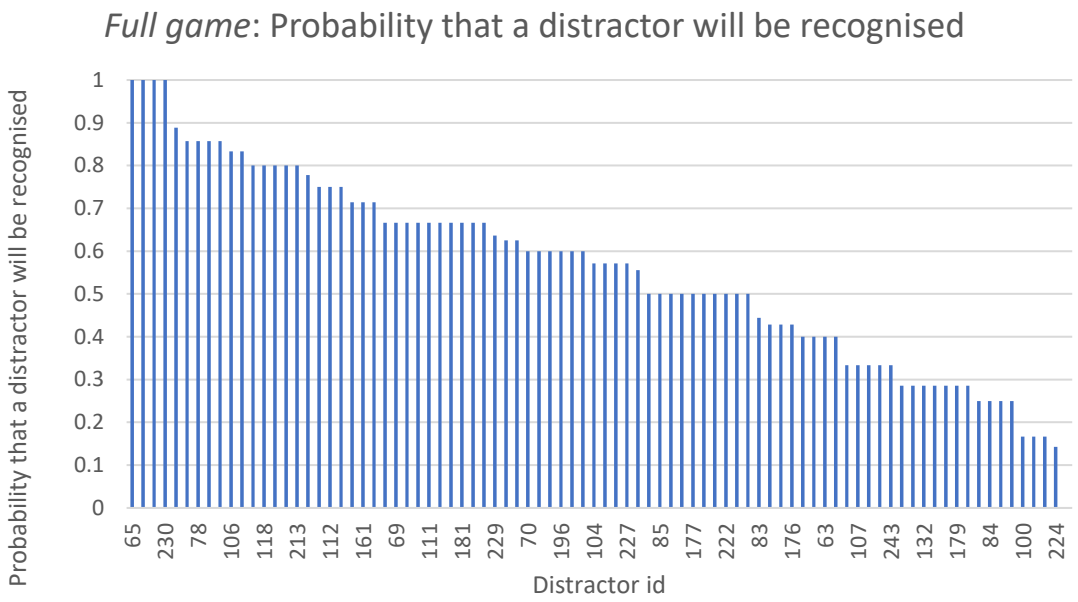

Figure 67 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest

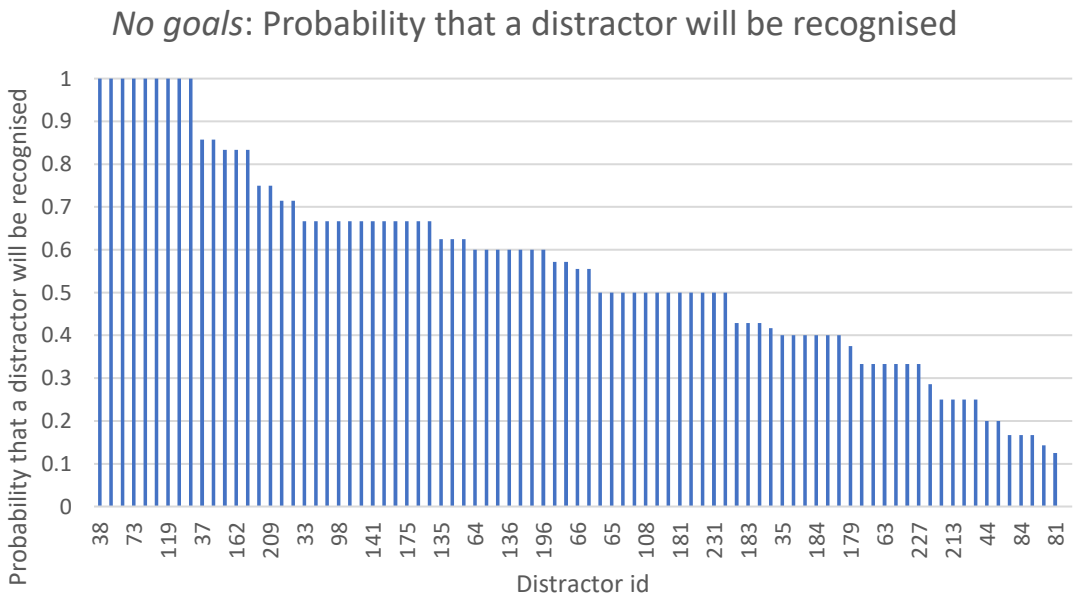

Figure 68 The probability that each different distractor shown in the *No goals* game will be recognised. The distractor times are ordered by probability – highest to lowest

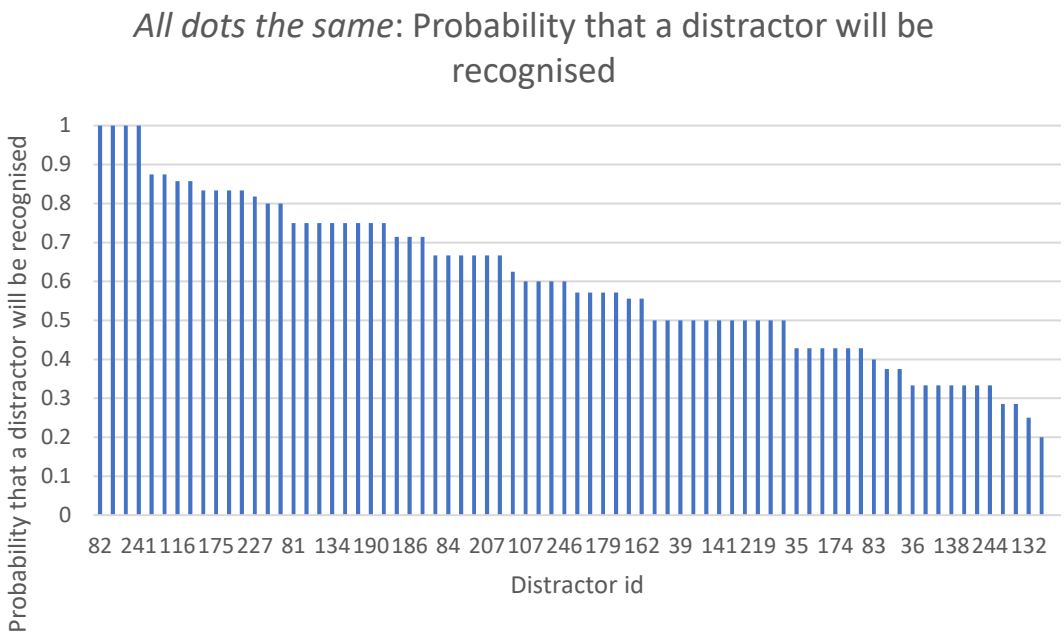

Figure 69 The probability that each different distractor shown in the *All dots the same* game will be recognised. The distractor times are ordered by probability – highest to lowest

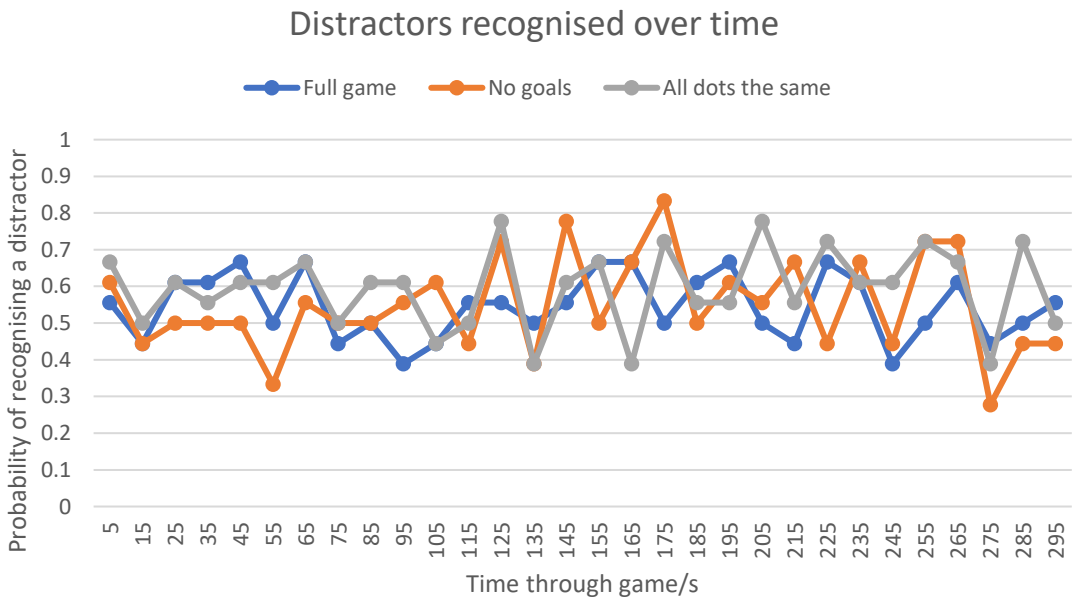

Figure 70 The probability that a distractor will be recognised against the time it is displayed

Immersion Experience Questionnaire

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was a significant difference in the immersion scores between the *Full game* (M=111.39, SD=9.78), the *No Goals* game (M= 107.83, SD= 13.31) and the *All dots the same* game (M=89.56, SD=11.13) conditions; $F(2,45)= 18.683, p<0.001, \eta_p^2= 0.423$

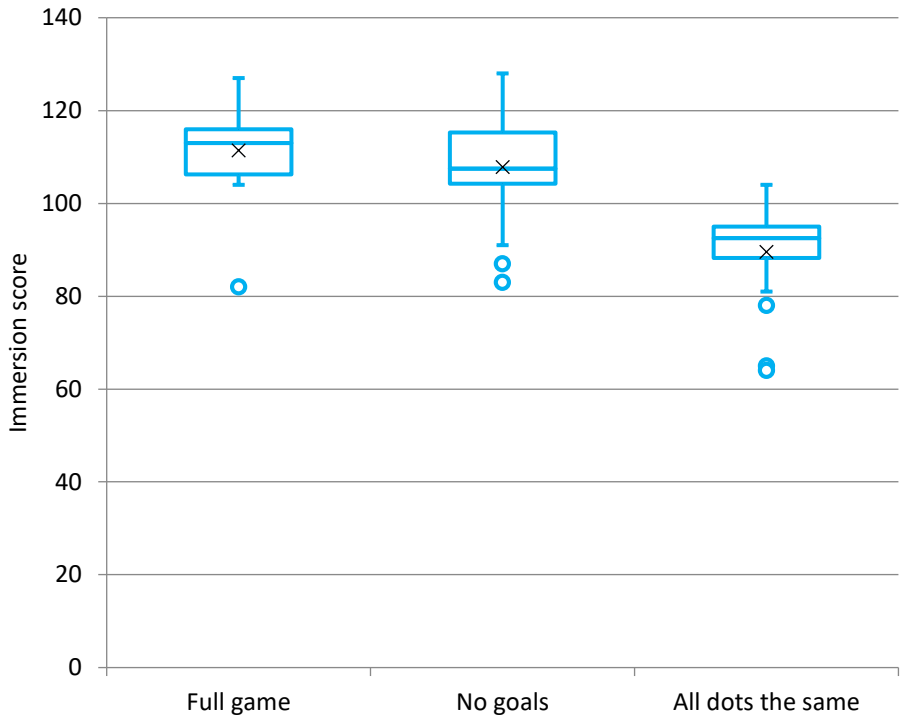

Figure 71 Boxplot showing the Immersion scores for all three conditions

The boxplot shown in Figure 71 contains 6 outliers. I was concerned that these might skew the ANOVA so I removed the outliers and repeated the results. There was still a significant difference in the immersion scores between the *Full game* (M=113.12, SD=6.67), the *No Goals* game (M= 110.69, SD= 11.04) and the *All dots the same* game (M=93.67, SD=5.75) conditions; F(2,45)= 26.150, p<0.001, η_p^2= 0.538
I also calculated the different IEQ components which are shown in Table 55.

	Full game Mean (SD)	**No goals** Mean (SD)	**All dots the same** Mean (SD)	**Effect size** η_p^2	**p value**	**F(3,52)**
Cognitive involvement	36.94 (3.57)	36.11 (5.50)	28.28 (5.10)	0.413	<0.001	17.932
Emotional involvement	18.56 (3.94)	18.22 (3.75)	12.78 (3.00)	0.366	<0.001	14.710
Real world dissociation	32.5 (4.13)	31.72 (4.98)	30.94 (5.63)	0.017	0.644	0.444
Challenge	14.67 (2.09)	11.5 (1.72)	9.44 (2.79)	0.492	<0.001	24.726
Control	16.11 (3.45)	17.72 (2.19)	15.5 (2.50)	0.108	0.053	3.103
Immersion	111.39 (9.78)	107.83 (13.31)	89.56 (11.13)	0.423	<0.001	18.683

Table 55 IEQ results for all three conditions

Correlations

I investigated whether there was a correlation between the IEQ scores and the number of distractors recognised for each condition. For each condition I calculated the Pearson's correlation coefficient (r).
Full game: r=--0.086, t(16) = -0.344, p = 0.735
No goals: r= 0.315, t(16)= -1.544, p = 0.142
All dots the same: r=0.273 t(16) = 1.1332, p = 0.274

There were no significant correlations between IEQ and distractors recognised for any of the conditions.

6.7.4. Comparisons with previous experiments

One of the goals of this experiment was to compare the results of removing eye tracking with the previous similar experiment which did use eye tracking. To do this I compared the results of this experiment with those of 6.4 Experiment attention 3: More similar games with eye tracking.

Condition	Webdings distractors with eye tracking	Webdings distractors without eye tracking	**Effect size η_p^2**	**p value**	**F(1,33)**
Full game	14.13 (3.00)	16.39 (3.35)	0.119	0.046	4.314
No goals	16.94 (2.93)	16.44 (2.26)	0.009	0.584	0.306
All dots the same	16.31 (2.60)	17.83 (3.35)	0.028	0.348	0.908

Table 56 compares the number of distractors recognised for each condition. Table 57 compares the immersion scores.

Condition	Webdings distractors with eye tracking	Webdings distractors without eye tracking	**Effect size η_p^2**	**p value**	**F(1,33)**
Full game	14.13 (3.00)	16.39 (3.35)	0.119	0.046	4.314
No goals	16.94 (2.93)	16.44 (2.26)	0.009	0.584	0.306
All dots the same	16.31 (2.60)	17.83 (3.35)	0.028	0.348	0.908

Table 56 Comparing the mean (SD) of number of distractors remembered with and without eye tracking

Condition	Webdings distractors with eye tracking	Webdings distractors without eye tracking	**Effect size η_p^2**	**p value**	**F(1,33)**
Full game	107.13 (17.00)	111.39 (9.78)	0.025	0.369	0.829
No goals	93.38 (14.05)	107.83 (13.31)	0.229	0.004	9.488
All dots the same	93.56 (13.38)	89.56 (11.13)	0.063	0.153	2.148

Table 57 Comparing the mean immersion scores (SD) with and without eye tracking.

6.7.5. Discussion

The motivation behind this experiment was that removing the eye tracking and leaving participants unattended would make them feel more comfortable about being distracted and produce a larger difference in the number of distractors recognised.
The first hypothesis is that there would be a significant difference between the number of distractors recognised. This was not supported. Participants did recognise more distractors in this experiment than the previous *Webdings* icons experiment however the difference between conditions was much smaller and not significant with a small effect size (η_p^2= 0.049). Participants recognised more distractors in the *Full game* ($p<0.05$) and the *All dots the same* conditions but fewer for the *No goals* condition. The standard deviations were also larger for the *Full game* and *All dots the same* conditions. It seems likely that participants have several different motivations which may affect the results. Removing the eye tracking and experimenter may make them more likely to be distracted but without the chin rest they can look all over the room, not just at the distractor images. As the distractor images are not that interesting, they recognise a few more for the *All dots the same* game but not substantially more than the other games. As in previous experiments plotting the recognition rates for each distractor icon showed no evidence that any of the icons was more or less likely to be recognised than any other.

The second hypothesis, that there would be a significant difference in immersion between games was supported with a very strong effect size (η_p^2=0.423). This was a substantially larger effect size than with eye tracking which was only η_p^2=0.166. Overall immersion is much higher and the differences between games larger. There were also significant differences in many of the sub factors of the IEQ including *Cognitive Involvement, Emotional Involvement* and *Challenge*. There is also a trend towards significance in *Control.* This increase in the effect size is probably due to removal of the eye tracking equipment which is intrusive and also adds a small delay to the mouse movement at intervals which reduced immersion. The *No goals* game had much larger immersion but the *All dots the same* immersion was slightly reduced. It seems likely that removing the pressure of being watched makes players more relaxed about the lack of challenge in the *No goals* game. During the previous experiments they may have felt confused by the lack of explicit goals and been concerned that they were "doing it wrong" but in this experiment it is possible they felt less self-conscious and more likely to enjoy the experience. However, it should be noted that comparing two different experiments like this is subject to confounds and these results could be down to random fluctuations in the data.

It is clear that this experiment was not successful in its aim to increase the difference in the number of distractors remembered. Removing eye tracking and the experimenter does mean that participants recognise more distractors after the game. However, it also increases the amount of engagement experienced for the *No goals* game and makes it

much more similar to the *Full game*. This may explain why the number of distractors remembered for the *No goals* game and the *Full game* is almost the same. Looking across the last three experiments it seems that immersion in the *No goals* game is particularly sensitive to changes in the environment around the game. In the original *Webdings* distractors experiment the immersion was similar to the *All dots the same* game. When I changed the distractors to Disney characters the immersion levels rose although the immersion levels for the other two games were reduced. Finally, for this experiment removing the intrusive eye tracking equipment and sense of being watched significantly ($p<0.05$) increased the immersion during the *No goals* game so that it was very close to that of the *Full game.* Looking at how contextual factors affect immersion was not the main focus of this series of studies but these results do seem to show that immersion in low challenge games is particularly sensitive to changes in context. Future studies could look at this in a more systematic way.

6.7.6. Limitations

One of the key limitations of this experiment is because I was not actually present in the room while participants were playing the game it is difficult to say exactly what they were doing. I did record logs of the game activity but this would require complex analysis to build up a picture of activity. Even if I did this then I may miss important information about what participants were actually doing. Alternatives would be to use screen recording software, cameras and/or one-way window observation rooms. Another limitation is because I removed both the eye tracker and myself from the experiment it is difficult to tell which of these changes caused the changes in the number of distractors recognised. Without eye tracking there is no record of where participants were actually looking, so we do not know whether they were looking at the game, the distractors or somewhere else in the room. Although removing intrusive recording methods such as eye tracking may make players feel less self-conscious and make them behave more naturally the downside is that without them it is difficult to record what has happened.

6.8. Chapter conclusions

The initial goal of this chapter was to investigate whether measuring attention could serve as a measure of game engagement. To do this I tested a measure based on showing participants irrelevant distractors while they played a game. After the game they were tested to see how many of those distractors they recognised. This is known as the distractor recognition paradigm (DRP).

My initial experiment showed that participants will recognise a high number of distractors if their attention is not on a game. The next experiment compared two very different games and found that the DRP gives a significant difference in the number of distractors recognised with a very high effect size. After that I compared three more similar games. Again, I found a significant difference in distractors recognised. These results indicate that the DRP can serve as an effective measure of game attention.

I was interested in the link between attention and immersion. Although the previous experiments showed a significant difference in immersion between games there were no significant correlations between the number of distractors recognised and the immersion score. This could be due to floor effects in the DRP which restrict the variance of the results or it could be that there is not a strong relationship between attention and immersion. To avoid these floor effects, I tried replacing the *Webdings* icons used in the previous DRP experiments with Disney characters. I then performed an experiment which showed these characters are as memorable as the *Webdings* characters. I then used them in a new experiment to measure the difference in attention between three similar games. This experiment showed that participants did recognise more distractors but that this changed the immersion levels between games which made the experience of each game more similar. This reduced the effect size of both the differences in immersion and the differences in distractors recognised. I concluded that using more interesting distractors such as Disney characters has too much effect on the game experience to be useful in measuring that experience.

I also used eye tracking to measure whether participants were looking at the distractors or the game. Although there were differences between conditions, they were not significant and the effect sizes between conditions were also smaller than in the DRP. This suggests that participants will not recognise the distractors just because they are looking at them, their internal attention needs to be focused on the distractor as well. I was concerned with the ecological validity of the experimental setup. The eye tracker makes participants feel "watched" and the chin rest may make them feel constrained. The presence of the experimenter in the room may also make participants feel watched and pressured to obey instructions by playing the game. To test this, I repeated the experiment with three similar games but no eye tracking and the experimenter out of the room. The results show higher immersion and a larger difference in immersion between games. However, the DRP does not show a significant difference in distractors recognised and the effect size is much lower. This is probably because when participants become bored of the game, they do not look at the distractors, they are more likely to look around the room.

Of all of these results, I was particularly interested in the finding that eye tracking was not a more useful measure of attention because the attention that players pay to a game is an *internal* mental state rather than being purely indicated by where they are looking. This suggested that I could put the distractors within the game itself and the amount of attention that players paid to them would be related to how engaging they found the game. This also had the potential to become a measure of engagement in self-paced games and my investigations into this idea are described in the next chapter.

7. In-game distractors

This chapter investigates another method of measure game engagement using distractor recognition. The previous chapter showed that game attention could be measured using distractors located around the game, the more immersed players were in the game the less attention they paid to the distractors. In this chapter, I look at what happens if the distractor images are within the game rather than around the sides.
The experiments in the previous chapter showed that players were more likely to be distracted from less immersive games. What was particularly interesting is that the nature of this distraction was not just that players were looking away from the game, there was a difference in their focus of attention within their field of view. The evidence for this is that using eye tracking to see whether players were looking at the game or the distractors was a weaker measure of how immersed they were in the game compared with testing how many distractors they recognised. This suggests that players were looking at the distractors, but because their mind was on the game, they did not register seeing them or recognise them afterwards.

This looks very similar to the phenomenon of *inattentional blindness* which is discussed in the literature review (see chapter 2). In these situations, participants do not see a stimulus, even though they are looking directly at it, because their attention is focused elsewhere. There is also evidence that during *inattentional blindness* participants really do not see the stimulus rather than seeing it but not remembering it. Investigations into this phenomenon have shown that it increases when participants are experts in the task and under high perceptual load. Both of these conditions are likely to happen during digital games in which players quickly gain expertise in the task and also have to process complex visual stimuli.

This suggested that the distractor recognition paradigm would still be effective as a measure of game attention if the distractor images were inside the game rather than around it. If players were only paying attention to the game, they would ignore the distractor images because they were irrelevant to the game. On the other hand, if player's attention wandered from the game, they would be sure to see the distractor images because they were looking directly at them. Players who are less engaged in the game are more likely to let their attention wander and look at the distractor images. Testing participants on how many distractors they recognise could thus form a measure of how engaged they were in the game. This chapter describes experiments which seek to test this hypothesis and explore how it works when applied to a real game.

7.1.1. Experimental plan

The experiments used in this chapter are very similar to those in the last chapter on measuring game attention. I decided to stay with the game of *Two Dots* which would allow me to use a similar experimental procedure and also to compare the results to previous experiments. Similarly, the distractor recognition paradigm used in the last chapter was shown to be an effective measure of game attention. So, for these experiments I used a similar paradigm. The experiments in the previous chapter successfully used icons from the *Webdings* font as distractor images to measure immersion. The previous experiments found that participants were able to recognise which ones they had seen before with a high level of accuracy – around 91%. However, they are not so distracting as to interfere with participants' experience of playing the game. Because of this, I decided to use these icons in the next series of experiments. I needed some way of putting them inside the game in such a way that participants would be looking straight at them throughout the game but that they would not interfere with the gameplay. I decided to increase the size of the dots and to put the same icon inside every dot (See Figure 72). This is described in chapter 3.The dots are different colours and players need to join dots of the same colour – the icon inside the dot is irrelevant to playing the game so this is likely to lead to players' learning to ignore them and if they are immersed in the game, they will become effectively blind to which icon they have seen. As with the previous distractor recognition paradigm all the icons on the screen change to a different icon every 5 seconds and players are then tested after the game on how many of them they recognise.

The progression of experiments also follows a similar plan to previous chapter. All the experiments used variants of the game of *Two Dots* as described in chapter 3. The first experiment was an initial exploration to determine if in-game distractors could be effective in measuring game attention. This uses a modest number of participants and compares two games with each other. The second experiment then explored how robust these findings were by replicating the first experiment but comparing three different games and a larger number of participants. The third and final experiment in the chapter investigates the effect of changing the instructions used in the distractor recognition paradigm. Unlike previous distraction experiments, participants were told before the experiment that they would be tested on the distractors afterwards.

7.2. Experiment attention 7: In-game distractors initial exploration

Aims

This experiment aimed to be an initial exploration of whether in-game distractors could be used to measure game engagement between two different games in the same way as I had previously used distractors around the game. To do this I made changes to the

distractor recognition paradigm (see chapter 6) so that the distractors were inside the game rather than around the game. Participants either played the *Full game* or the *All dots the same* game. Both of these games contained regularly changing visual distractors. At the end of the activity participants were tested to see how many distractor symbols they recognised. This was an initial study designed to see if there would be any kind of effect between conditions. As I was not sure if there would be any effect at all I did not want to commit too much time to running the experiment and used a modest number of participants which made it likely that the study was under powered. As an initial exploration, finding a significance difference between conditions was not the primary goal of the experiment. I was more interested in the effect size between conditions and the direction of the effect (if any).

Hypothesis

The main hypothesis of the experiment was:
The number of distractors that participants recognise will be lower for the *Full game* condition than the *All dots the same* condition. The null hypothesis is that there will be no difference between the number of distractors recognised.
An additional secondary hypothesis is that the immersion score will be higher for the *Full game* than the *All dots the same* game.

7.2.2. Method

Design

This was a between-subjects design with two conditions. The independent variable was the game each participant played – either the *Full game* or the *All dots the same* game. The main dependent variable was the number of distractors that participants recognise seeing after the activity. Another, secondary dependent variable is the Immersion Experience Questionnaire (IEQ) score for each participant's experience of the activity.

Participants

As previously discussed, this was an initial exploratory study with a modest number of participants. 24 staff and students took part in the study. 12 were male, 20 were native speakers of English with ages ranging from 18-36 (Mean 21.46). Participants were paid £6 for their participation.

Materials

This study used two different games; the *Full game* and the *All dots the same* game. These are described in more detail in chapter 3. Both of these games had irrelevant distractor images added into the game. This was done by increasing the size of the dots and putting the same distractor image into each dot. This is shown in Figure 10 and described in more detail in chapter 3.

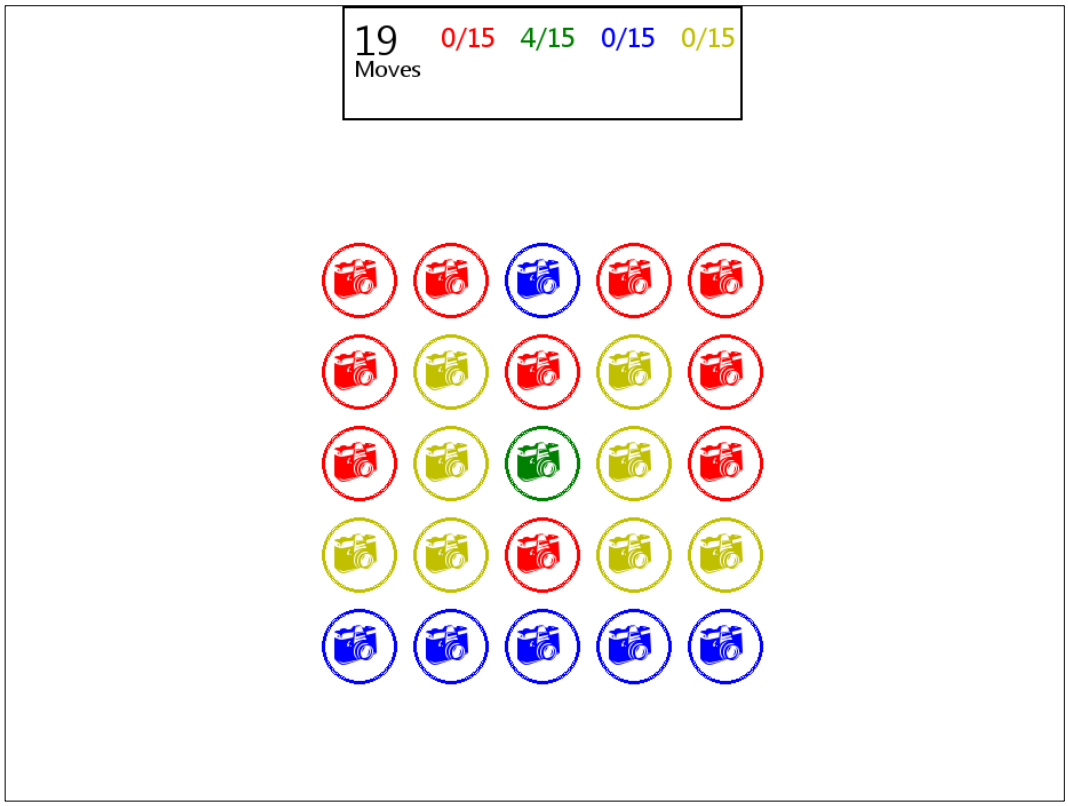

Figure 72 A screen from the Full game. Players have to join dots of the same colour. The images inside the dots change every 5 seconds.

Once players had finished the game, they were tested on which distractor pictures they recognised using the same *distractor recognition paradigm* described in chapter 6.

Procedure

The procedure was identical to previous experiments performed using the distractor recognition paradigm described in chapter 6.

7.2.3. Results

Distractor Symbols

There was a significant difference in the number of correct distractors recognised between the *Full game* (M=17.92, SD=2.71) and the *All dots the same* game (M=21.25, SD= 3.36) conditions; F(1,23)= 7.149, p=0.014, η_p^2= 0.245. I plotted the number of distractors recognised on a boxplot (Figure 73). I also plotted histograms of the probability of each different distractor being recognised for both the *Full game* (Figure 74) and the *All dots the same* game (Figure 75). I also potted the probability of each distractor being recognised over the time of the game (Figure 76).

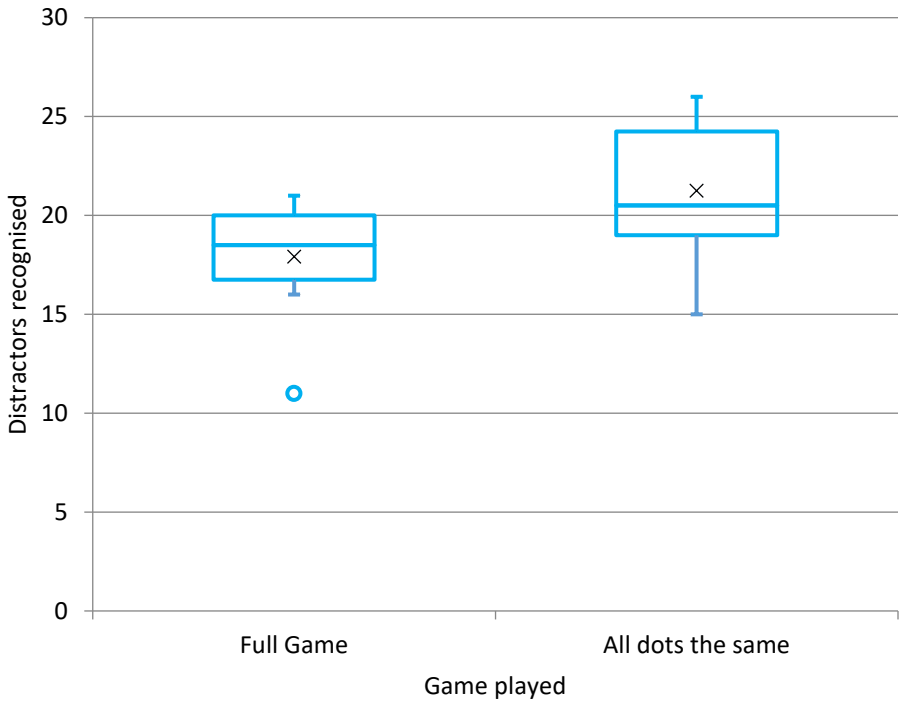

Figure 73 Boxplot of the number of distractors recognised for each game

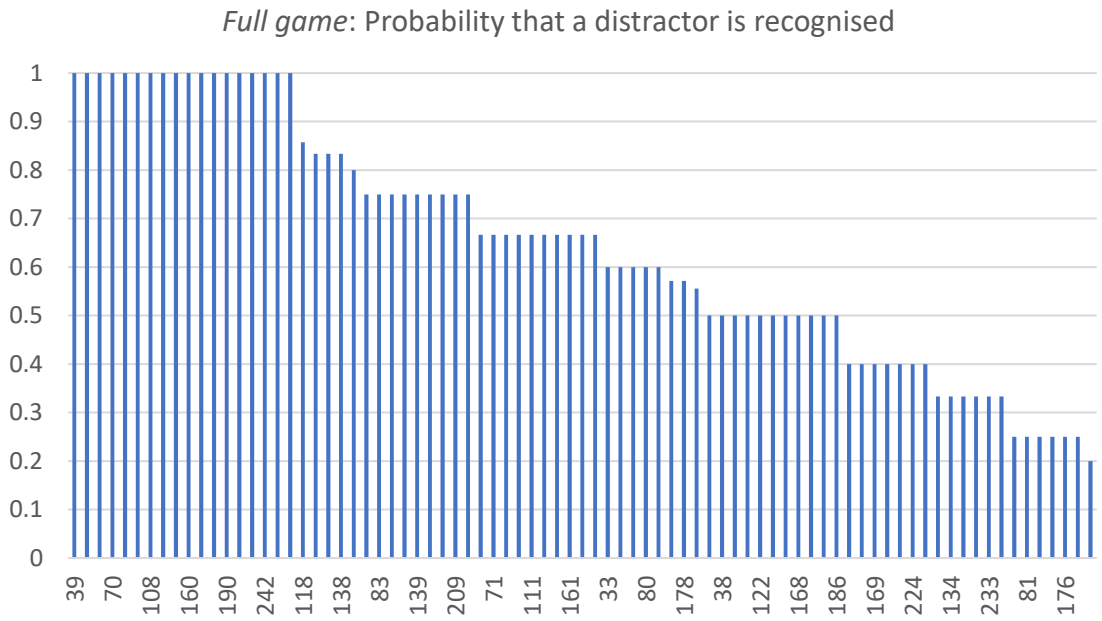

Figure 74 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest.

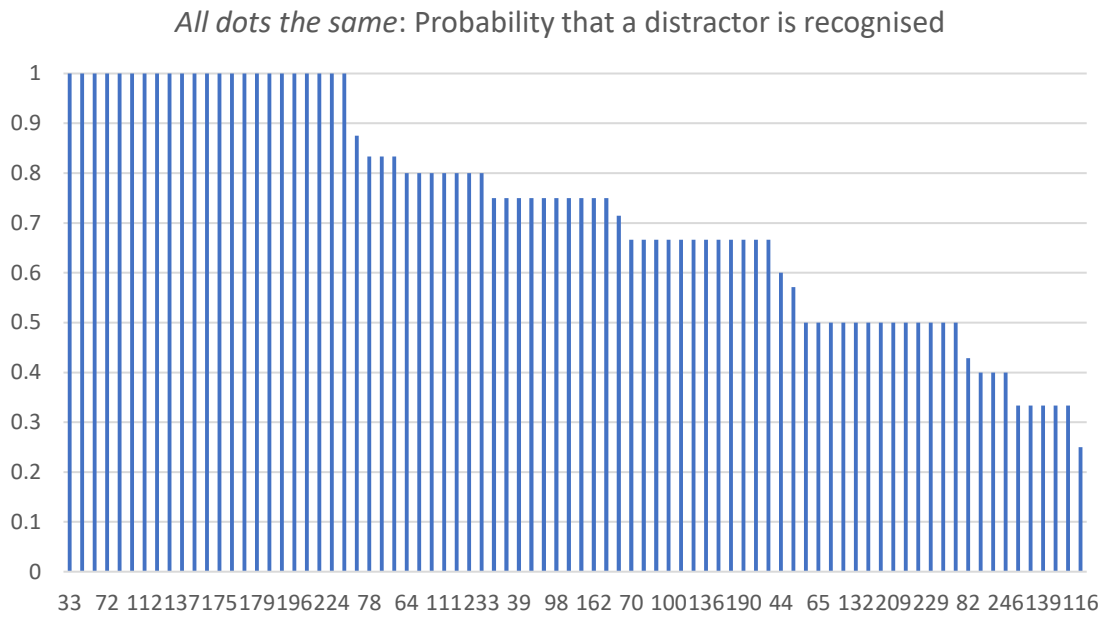

Figure 75 The probability that each different distractor shown in the *All dots the same game* will be recognised. The distractor times are ordered by probability – highest to lowest.

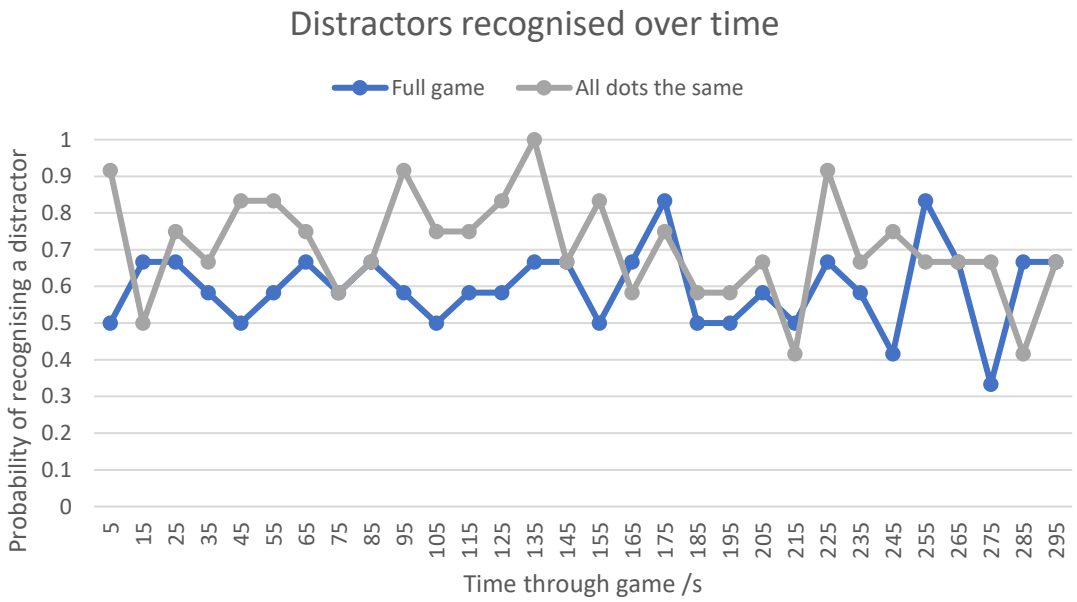

Figure 76 How the distractors were recognised over the time of each condition

Immersion Experience Questionnaire (IEQ)

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was a significant difference in the immersion scores between the *Full game* (M=113.67, SD=13.54) and the *All dots the same* game (M= 91.17, SD= 13.90) conditions; F(1,23)= 16.140,p=0.001, η_p^2= 0.423. There was not a significant Pearson's correlation between the number of distractors recognised and the immersion

score for either the *Full game* (r= -0.253, t(10)= -0.828, p= 0.427) or the *All dots the same game* (r= 0.289, t(10)= 0.955,p= 0.362). Plotting scatterplots (not shown here) did not show significant numbers of outliers which may have affected this result.

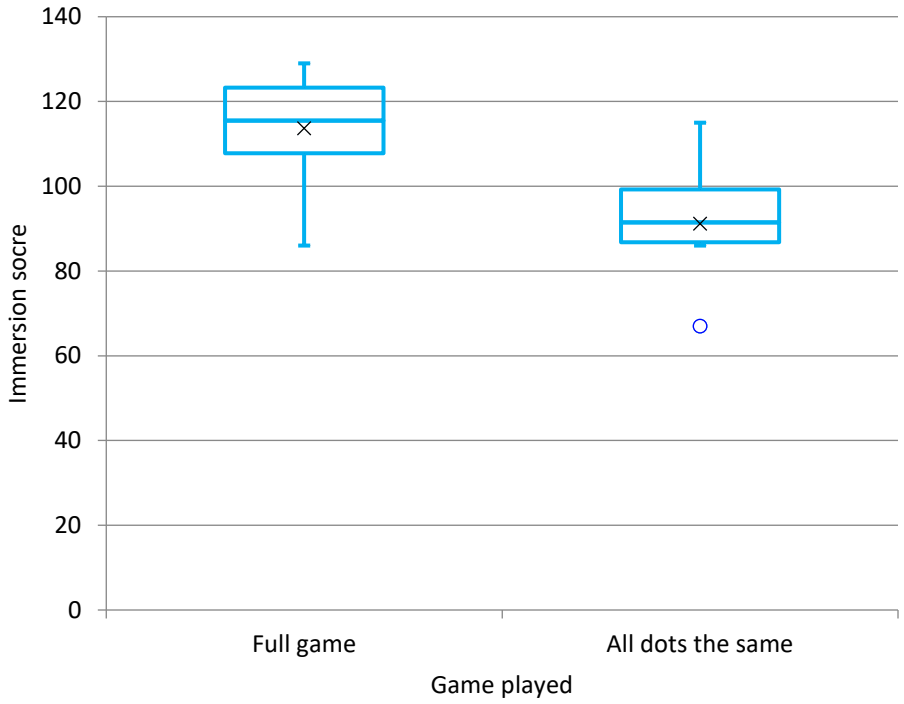

Figure 77. Boxplot showing IEQ scores for both conditions

IEQ scores can be broken down into five categories; *Cognitive involvement, Emotional involvement, Real world dissociation, Challenge and Control*. As can be seen from Table 58 there was a significant difference between the three conditions in the scores for *Cognitive involvement*, *Emotional involvement* and *Challenge*. Although there was a difference in the other scores, *Real world dissociation* and *Control* it was smaller and not significant.

	Full game Mean (SD)	**All dots the same game** Mean (SD)	**Effect size** η_p^2	**p value**	**F(1,23)**
Cognitive involvement	34.3 (5.5)	26.4 (4.6)	0.356	0.005	9.950
Emotional involvement	19.5 (6.2)	12.2 (3.6)	0.489	0.001	17.246
Real world dissociation	32.2 (7.6)	27.5 (6.7)	0.153	0.088	3.259
Challenge	14 (1.7)	9.4 (2.9)	0.444	0.001	14.384
Control	17.8 (4.2)	15.5 (2.0)	0.148	0.094	3.130
Immersion	110.2 (14.4)	84.6 (10.8)	0.530	<0.001	20.297

Table 58 Results from the IEQ

7.2.4. Discussion

The hypothesis that the number of distractors recognised by participants who played the *All dots the same game* would be higher than the number remembered by those playing the *Full game* was supported with a strong effect size (η_p^2= 0.245). As an initial exploration, this study was looking to see there would be any kind of effect between conditions. These results confirm that the number of in-game distractors that participants recognised was affected by how engaged they were in the game. However, as a small-scale and possibly underpowered study the effect size finding is unlikely to be robust.

I investigated whether that some distractors may be more memorable than others so plotted distractor recognition rates for both conditions in Figure 74 and Figure 4. If one distractor was particularly memorable or un-memorable then it would be expected to see the bar much higher or much lower than the others on the chart. This is not the case, so it is likely that each distractor image is around as memorable as the others. I also plotted the probability of distractors being recognised across the time of the experiment and found no strong pattern for either game.

The second hypothesis that the immersion score would be significantly higher in the *Full game* than the *All dots the same* game was supported with an extremely high effect size (η_p^2= 0.423). This shows that participants were significantly more immersed in the *Full game* than the *All dots the same* game. The very strong effect size suggests that participants had a very different experience when playing the two games and that their game experience was not significantly affected by the presence of the distractor images. Breaking down the IEQ score into its subfactors shows significant differences which are consistent with the differences between the two games. *Cognitive involvement*, *Emotional involvement* and *Challenge* are all significantly different with large effect sizes which is probably due to the significant differences in game elements between the two games. *Real World Dissociation* and *Control* do not have significant differences which is probably due to both games having the same control method and very similar graphics.

Participants recognised fewer distractor images during the more immersive game. This may be because they were paying less attention to the distractors rather than the particular elements of the game which are needed to play. Even though playing both games required them to look directly at the distractor images participants recognised fewer after playing the more immersive game. This supports the initial idea that participants would be subject to inattentional blindness while playing a game because their attention is focused on the elements necessary to play the game rather than elements that they deem irrelevant. The level of this attention seems linked to how engaged players are in the game which suggests that recognition of in-game irrelevant distractors may be an effective measure of game engagement.

Limitations

The main limitation of this study is that it was an initial exploration of the effect and had a small number of participants and so was likely to be underpowered. This lack of power means that these results may not be robust. However, the big effect size is encouraging. As this produced a significant effect the next step would be to rerun the experiment with a larger number of participants.
My main hypothesis is that recognition rates are different for the different games due to the different amounts of attention that participants paid to each game. However, it is also possible that recognition rates were affected by differences in cognitive load required for each game. Baddeley (2013) reports that memory of stimuli is negatively affected by high cognitive load so it could be that the *Full game* requires more cognitive load from participants than the *All dots the same* game which lowers the recognition rate of distractors. However, the results from chapter 5 suggest that there is no significant difference in cognitive load used between these games.

7.3. Experiment attention 8: In-game distractors replication

This experiment builds on the findings of the last experiment. That experiment showed a significant difference in the number of in-game distractors recognised between more and less engaging games. However, it was a small-scale experiment with a modest number of participants. This next experiment aimed to explore how robust this effect was. It also aimed to provide some degree of comparison with the experiments from the last chapter which used distractors around the game. To do this I increased both the number of participants and the number of different games which were compared. Apart from these two changes it was almost a complete replication of the previous experiment.

Aims

This experiment aimed to explore whether the difference in in-game distractor recognition between games with different levels of engagement was a robust effect. It also aimed to investigate this effect when playing an additional game; the *No goals* game (see chapter 3) and so provide some degree of comparison with the previous experiments with distractors around the game (see chapter 6).

Hypothesis

The main hypothesis of the experiment was:
The number of distractors that participants recognised will be lower for the *Full game* than for the *All dots the same* condition or the *No goals* condition. The null hypothesis is that there will be no difference between the number of distractors recognised.
There is also a secondary hypothesis that the immersion score will be higher for the *Full game* than for the other two games.

Design, materials and procedure

This was a between-participants experiment with three conditions. The independent variable was the game that participants played; either the *Full game,* the *No goals* game or the *All dots the same* game (see chapter 3). The rest of the design, materials and procedure were identical to the previous experiment and based on the distractor recognition paradigm outlined in the previous chapter (see chapter 6).

Participants

The first in-game distractors experiment had a larger effect size of η_p^2= 0.245. Based on this was thought it likely that this experiment would have the same or a larger effect size than the previous one. This experiment aimed to provide a degree of comparison with the very similar previous experiment using distractors around the game (see chapter 6). I performed a power calculation to estimate how many participants to have in the study. In the previous distractors around the game experiment the difference in the number of distractors recognised had an effect size of η_p^2= 0.162. This is equivalent to a Cohen's f of 0.4397. Using this effect size in a power calculation with a power of 0.8 (80%), an alpha of 0.05 and 3 conditions gives 17.65 participants per condition which I rounded up to 18 for each condition.

54 participants took part in this experiment (18 in each condition). Ages ranged from 18-57 (Mean 21.20). 28 were male and 49 were native speakers of English.

7.3.2. Results

Distractor symbols

There was a significant difference in the number of correct distractors remembered between the *Full game* (M=16.11, SD=3.01), the *No goals* game (M=18.33, SD=3.01) and the *All dots the same* game (M=18.56, SD=2.23) conditions; F(2,52)= 4.276, p=0.019, η_p^2= 0.144. I performed a Tukey's HSD post-hoc test to see which conditions were different from each other (See Table 59). I also plotted the number of distractors recognised on a boxplot (Figure 78).

Condition	*No goals*	*All dots the same*
Full game	p=0.029	p=0.051
No goals		p=0.969

Table 59 Tukey's HSD post-hoc test on the number of distractors recognised

For comparison with the previous experiment I also recalculated the ANOVA without the *No goals* condition. This also gave a significant difference between conditions with a similar effect size. F(1,35) = 4.910, p=.034, η_p^2= 0.126

I also plotted histograms of the probability of each different distractor being recognised for each different condition (Figure 79, Figure 80 and Figure 81). I also potted the probability of each distractor being recognise over the time of the game (Figure 82.

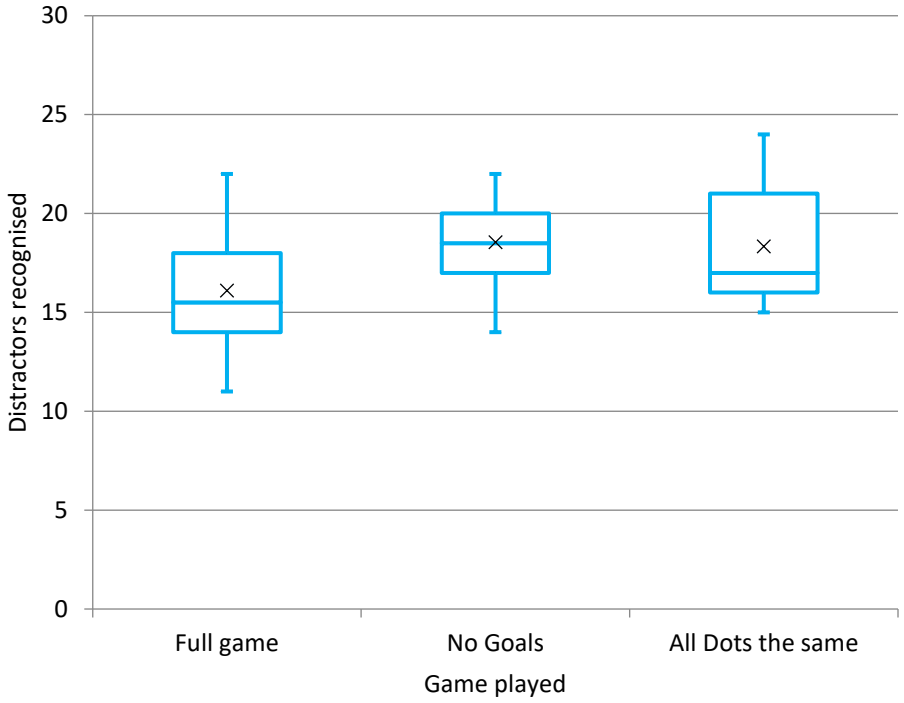

Figure 78 Boxplot of the number of distractors recognised for each game

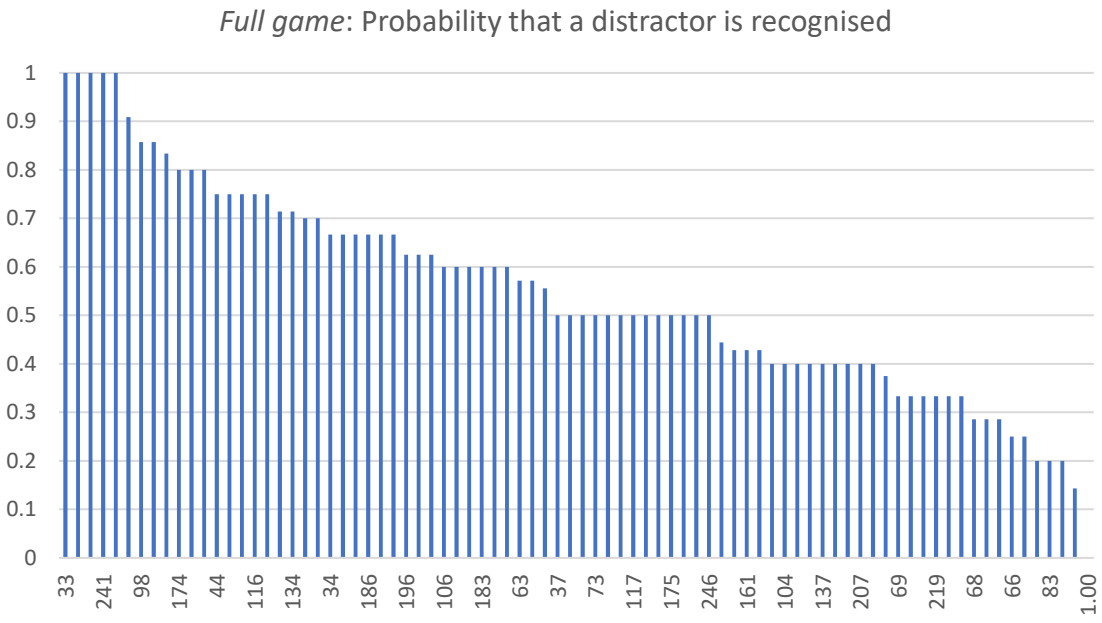

Figure 79 The probability that each different distractor shown in the *Full game* will be recognised. The distractor times are ordered by probability – highest to lowest.

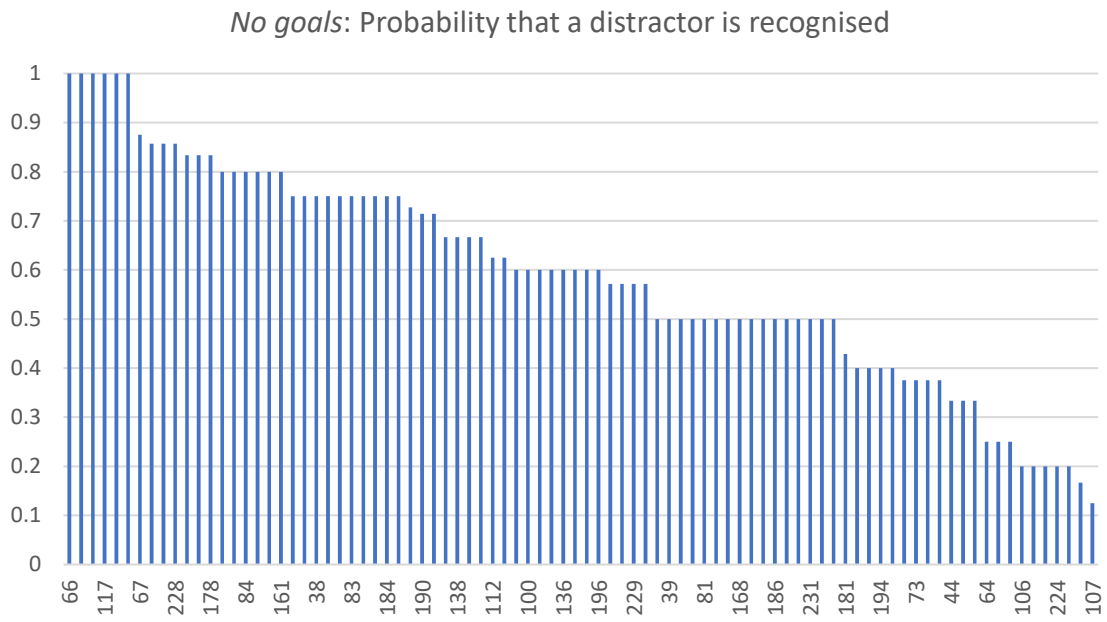

Figure 80 The probability that each different distractor shown in the *No goals* will be recognised. The distractor times are ordered by probability – highest to lowest.

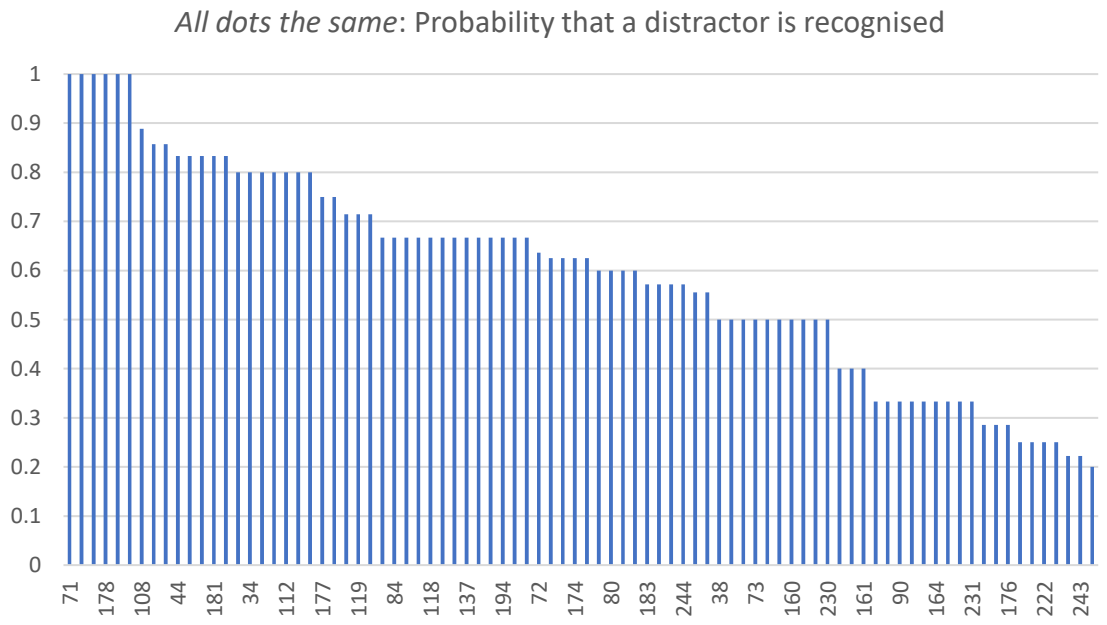

Figure 81 The probability that each different distractor shown in the *All dots the same* game will be recognised. The distractor times are ordered by probability – highest to lowest.

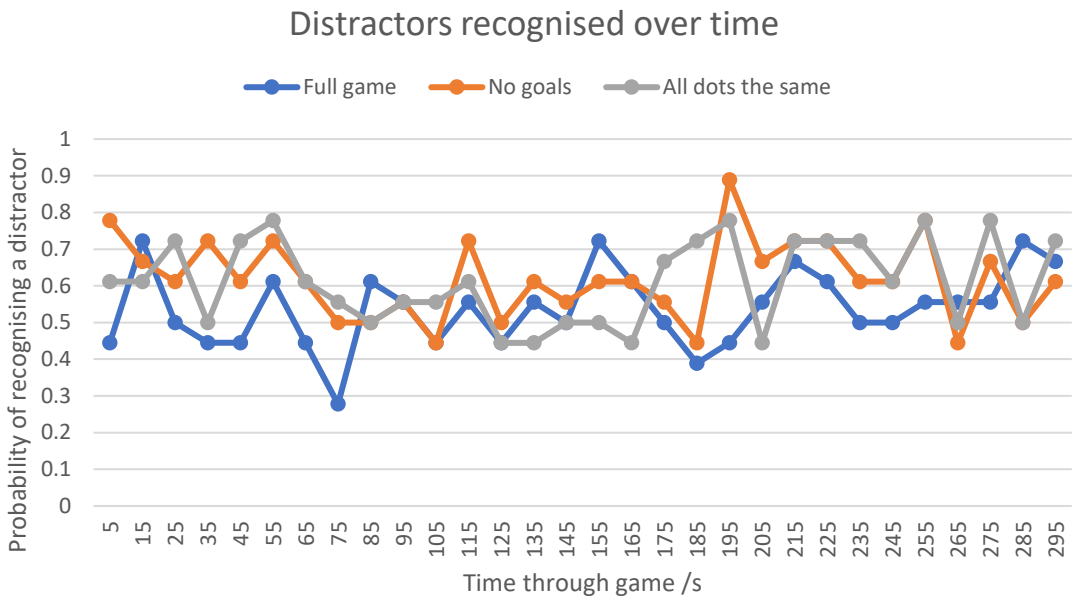

Figure 82 How the distractors were recognised over the time of each condition

Immersion Experience Questionnaire (IEQ)

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was a not a significant difference, although there was a trend towards significance (p=0.058) in the immersion scores between the *Full game* (M=102.50, SD=13.54), the *No goals* (M=102.94, SD=15.96) game and the *All dots the same* game (M= 92.94, SD=10.77) conditions; F(2,52)= 3.019,p=0.058 , η_p^2= 0.106.

I also recalculated the ANOVA without the *No goals* condition. This gave a significant difference between conditions with an increased effect size. F(1,35) = 5.198, p=0.029, η_p^2= 0.133

There was a significant Pearson's correlation between the number of distractors recognised and the immersion score for the *Full game* (r= -0.54, t(16)= -2.57, p = 0.02) but not the *No goals* (r= 0.234, t(16)= 0.963,p= 0.345) or the *All dots the same game* (r= -0.008, t(16)= -0.03,p= 0.973). Plotting scatterplots (not shown here) did not show significant numbers of outliers which may have affected this result.

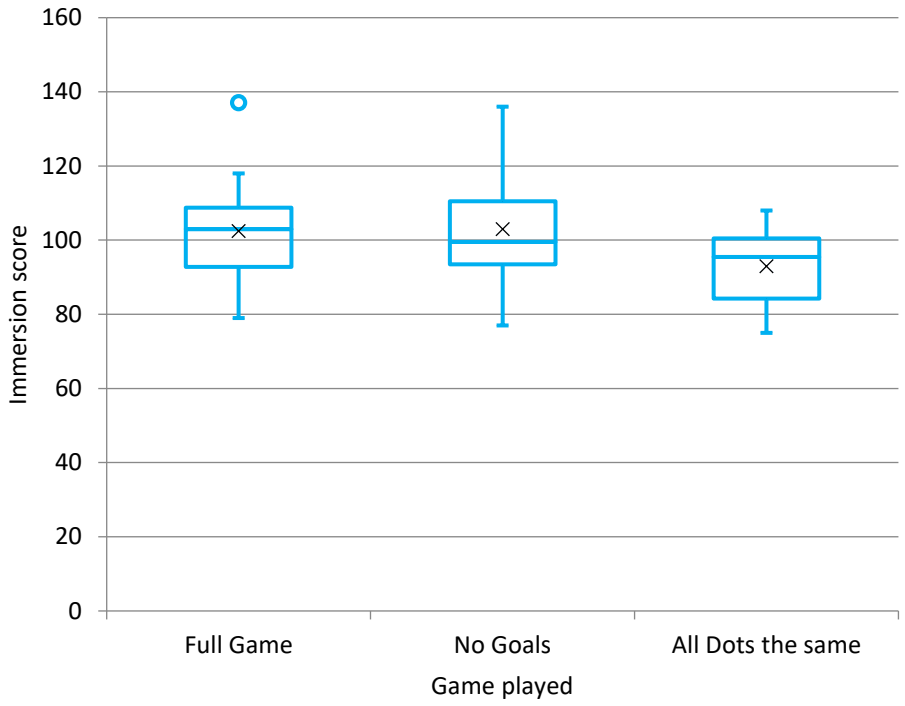

Figure 83 Boxplot showing IEQ scores for all three conditions

IEQ scores can be broken down into five categories; *Cognitive involvement, Emotional involvement, Real world dissociation, Challenge and Control*. As can be seen from Table 60 there was a significant difference between the three conditions in the scores for *Cognitive involvement*, *Challenge* and *Control*. Although there was a difference in the other scores, *Real world dissociation* and *Emotional Engagement* it was smaller and not significant.

	Full game Mean (SD)	**No goals** Mean (SD)	**All dots the same game** Mean (SD)	**Effect size** η_p^2	**p value**	**F(1,23)**
Cognitive involvement	33.44 (5.7)	32.83 (6.0)	28.83 (4.4)	0.131	0.028	3.845
Emotional involvement	16.22 (3.9)	15.67 (4.0)	14.89 (4.2)	0.019	0.614	0.492
Real world dissociation	31.89 (4.9)	34.06 (5.4)	32.17 (4.3)	0.039	0.363	1.035
Challenge	12.50 (2.9)	11.33 (2.8)	9.83 (2.6)	0.141	0.021	4.196
Control	16.00 (3.0)	17.22 (3.0)	14.61 (2.4)	0.131	0.028	3.827
Immersion	102.50, (13.54)	102.94, (15.96)	92.94, (10.77)	0.106	0.058	3.019

Table 60 Results from the IEQ

7.3.3. Discussion

The main hypothesis of the experiment, that there would be a significant difference in the number of distractors recognised was supported. A post-test showed a significant different between the *Full game* and the *No goals* conditions and a trend to significance between the *Full game* and the *All dots the same* conditions. There was no significant difference between the *No goals* and the *All dots the same* conditions. The effect size between conditions was moderate but smaller than that in the previous experiment. The main differences between this experiment and the last one was the additional of the *No goals* game condition and the experiment was performed on more participants. Removing the *No goals* condition from the analysis made little difference to the result, the difference between conditions was still significant with a similar effect size. As with the previous experiment I investigated whether some distractors may be more memorable than others so plotted distractor recognition rates for all three conditions in Figure 79, Figure 80 and Figure 81. If one distractor was particularly memorable or un-memorable then it would be expected to see the bar much higher or much lower than the others on the chart. This is not the case, so it is likely that each distractor image is around as memorable as the others. I also plotted the probability of distractors being recognised across the time of the experiment and found no strong pattern for either game.

The secondary hypothesis of the experiment, that there would be a significant difference in the immersion scores between conditions was not supported although there was a trend to significance (p=0.058). There was a moderate effect size which was smaller than that for both the previous experiment and also for a similar experiment with distractors around the game. (see previous chapter). Removing the *No goals* condition from the analysis showed a significant difference between the other two conditions with an increased effect size. In the experiments in the last chapter the mean immersion score for the *No goals* condition was lower than that for the *Full game.* In this experiment the mean immersion score for the *No goals* condition was slightly higher than that for the *Full game*. For this experiment the mean immersion score for the *Full game* was lower compared to previous experiments (such those described in the previous chapter) and the mean immersion score for *No* goals game was slightly higher. The lack of difference between these two conditions may explain the lack of a significant difference in immersion between the three conditions. Looking at the different immersion factors there is a particularly noticeable difference in *Emotional involvement.* In this experiment it has the smallest effect size (0.019) between conditions of all the factors. In the previous experiment this factor had the largest effect size (0.489) between conditions.

It is possible that this difference in immersion score is caused by the presence of the in-game distractors or it might just be chance variation. These distractors may distract from the *Full game* reducing the immersion score, they may also make the *No goals* game more interesting thus increasing the immersion. However, it is also possible that this difference is due to random variation in participants. None of these experiments have a

large number of participants and although the differences in significance and effect size are noticeable, they are not large.

In conclusion, this experiment replicates the findings of the previous experiment with regard to using in-game distractor recognition as a measure of game engagement. There was a significant difference between conditions which was related to how engaged players were in the game. This provides further support that in-game distractors can be used to measure the experience of playing self-paced games.

Limitations

This experiment aimed to measure different levels of engagement between three different games. To ensure that these games did indeed give different levels of engagement, participants filled in an immersion questionnaire. However, the difference in immersion between games was not significant, although there was a trend to significance. This may suggest that the differences in game experience were not large and may be a limitation of this experiment. However, the difference between conditions for the number of distractors recognised was significant which suggests that in this case the distractor recognition paradigm was a more reliable measure of the differences between the games.

The lack of a significant difference in immersions scores is probably down to either the *No goals* condition being more similar to the *Full game* condition or the *No goals* condition immersion score being more sensitive to changes in external factors such as the presence of distractors. Due to this uncertainty I removed the *No goals* condition from the next experiment.

Another limitation of this experiment is the lack of any kind of pattern in the distractor recognition rates over time. As with previous experiments this does not show any difference in engagement rates over the time of the game. It is also unclear whether this is due to the method of measurement or lack of variation in the game experience over time.

7.4. Experiment attention 9: In-game distractors, but told to remember the distractors

We have all had the experience of playing a game when we know that there is something else that we should be doing. But we tell ourselves we will just play "just one more game" and ignore the distraction from the other activity even though we are aware we need to do it.

None of the distraction experiments that I have performed so far have told participants that they will be tested on the distractors after the game. Participants were simply told "Play the game for five minutes and afterwards I will ask you some questions". I was interested to see whether this could be made more ecologically valid by telling participants that they will be tested on the distractors after playing the game. This may be a stronger test of how engaging the game is, for a really engaging game participants may be so engaged in playing the game that they ignore the distractors even though they know they should be looking at them.

Aims

This experiment aimed to test the effect on the in-game distractor recognition paradigm of a small change in the instructions given to participants. Instead of just being told to play the game, participants were told they would be tested on the in-game distractors at the end.

Hypothesis

The main hypothesis of the experiment was:
The number of distractors that participants recognised will be lower for the *Full game* than for the *All dots the same* condition. The null hypothesis is that there will be no difference between the number of distractors recognised.
There is also a secondary hypothesis that the immersion score will be higher for the *Full game* than for the other games.

Design, materials and procedure

This experiment was identical to the previous in-game distractor experiment apart from two differences. The first difference was in the instructions that participants were given. They were told "Play the computer game for five minutes. The game contains pictures. After playing the game you will be tested on the pictures and fill in a questionnaire. But try to do as well as you can at the game". They were told this both in the consent procedure information sheet and verbally before the experiment. The second difference was that this experiment only contained two conditions; the *Full game* and the *All dots the same* condition. The previous experiment showed that immersion in the *No goals* game may have larger variance and be more sensitive to changes in the experimental environment. Removing this condition reduces the number of participants needed and focuses the experiment on establishing an effect between the conditions which have previously been seen to be most different from each other. These changes meant that the experiment was a between-participants experiment with two conditions.

Participants

This experiment aimed to provide a degree of comparison with the previous in-game distractor experiment so had the same number of participants (18) per condition. 36 participants took part in the experiment, ages ranged from 18-34 (mean 21.2), 20 were male and 3 were not native speakers of English.

7.4.2. Results

Distractor symbols

There was no significant difference in the number of correct distractors remembered between the *Full game* (M=18.06, SD=2.38), and the *All dots the same* game (M=18.94, SD=3.64) conditions; F(1,35)= 0.751,p=0.392, η_p^2= 0.022. I plotted the number of distractors recognised on a boxplot (Figure 84).

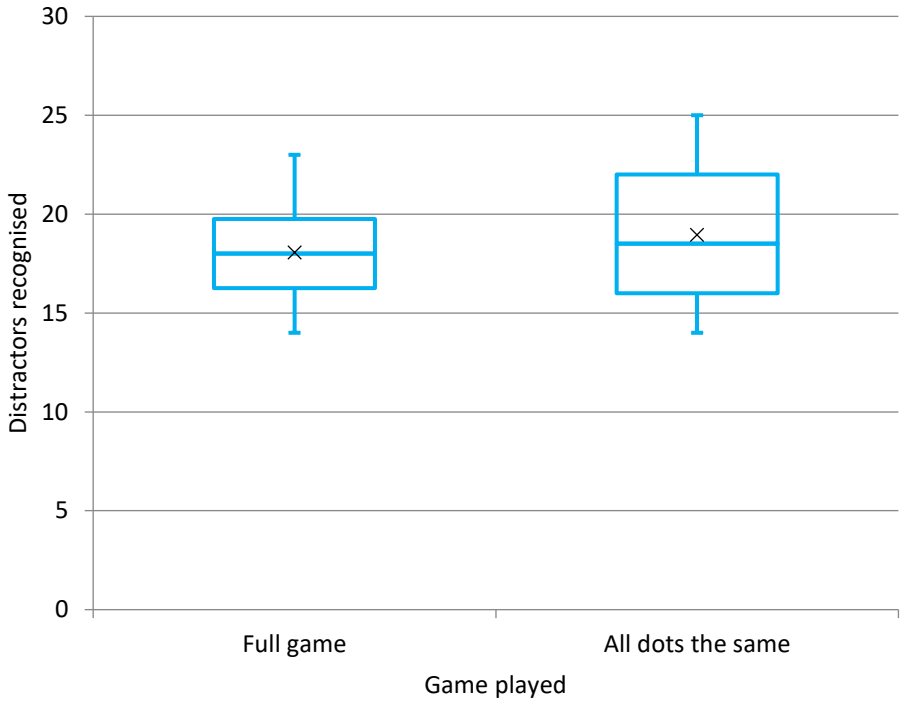

Figure 84 Boxplot of the number of distractors recognised for each game

Immersion Experience Questionnaire (IEQ)

The purpose of the IEQ was to confirm that participants in each condition had had a different game experience. There was no significant difference in the immersion scores between the *Full game* (M=100.28, SD=18.029) and the *All dots the same* game (M= 95.78, SD=18.17) conditions; F(1,35)= 0.556, p=0.461, η_p^2= 0.016.

There were no significant Pearson's correlations between the number of distractors recognised and the immersion score for the *Full game* (r= -0.152, t(16)= -0.615, p = 0.547) or the *All dots the same game* (r= -0.151, t(16)= -0.613,p= 0.549). Plotting scatterplots (not shown here) did not show significant numbers of outliers which may have affected this result.

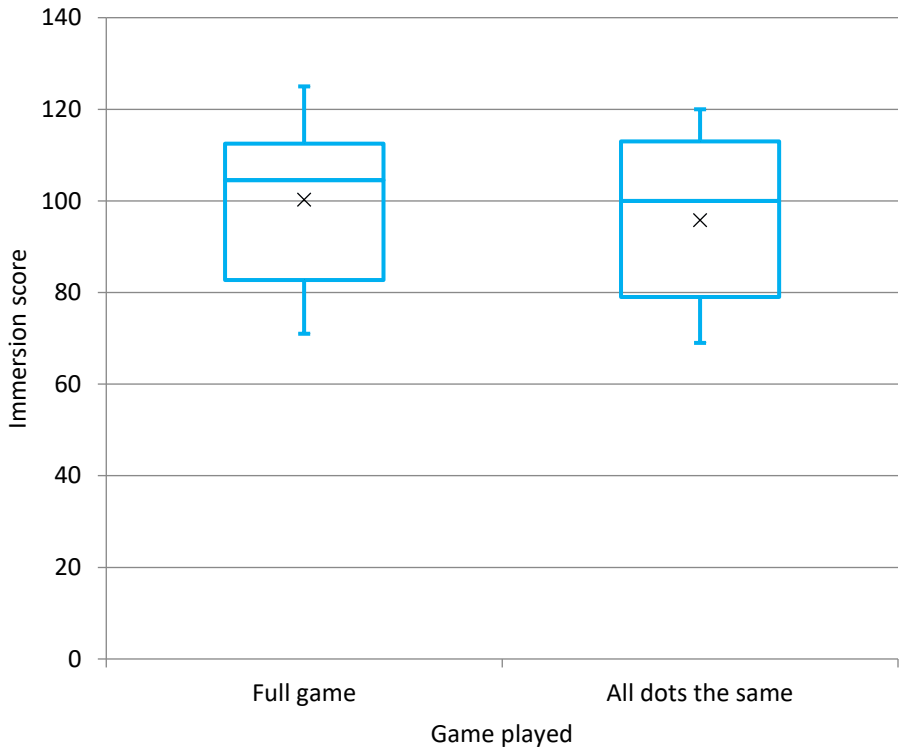

Figure 85 Boxplot showing IEQ scores for both conditions

IEQ scores can be broken down into five categories; *Cognitive involvement, Emotional involvement, Real world dissociation, Challenge and Control*. These are plotted in Table 61. There are no significant differences between any of the subfactors although *Challenge* is has a trend towards significance (p=0.059).

	Full game Mean (SD)	**All dots the same game** Mean (SD)	**Effect size** η_p^2	**Significance** (p value)	**F(1,23)**
Cognitive involvement	31.11 (6.6)	29.72 (6.1)	0.012	0.518	0.427
Emotional involvement	17.11 (4.9)	16.39 (5.7)	0.005	0.687	0.165
Real world dissociation	29.39 (5.4)	30.22 (6.8)	0.005	0.687	0.165
Challenge	13.39 (2.3)	11.72 (2.7)	0.101	0.059	3.831
Control	15.83 (3.3)	14.89 (3.1)	0.022	0.388	0.766
Immersion	100.28, (18.029)	95.78, (18.17)	0.016	0.461	0.556

Table 61 Results from the IEQ

7.4.3. Discussion

The main hypothesis of the experiment that there would be a significant difference in the number of distractors recognised for each condition was not supported. The effect size between conditions was extremely small.
The secondary hypothesis of the experiment that there would be a significant difference in the immersion scores between conditions was also not supported. The effect size between conditions was also extremely small. Looking at the sub-factors of the IEQ none of them showed a significant difference although of *Challenge* did show a trend to significance (p=0.059) and a moderate effect size. Not only did the distractor recognition paradigm not show a difference between the different games but the immersion experience questionnaire which previously has consistently shown a significant difference between these two games shows no significant difference.

It seems likely that telling participants that they will be tested on the distractors does make them pay a bit more attention to the distractors, but as they are also trying to play the game, they do not recognise that many more distractors. This does however, have the effect of reducing the immersion in the *Full game*. It has the converse effect on the *All dots the same* game which had a higher immersion score than in previously experiments. It may be that having to pay attention to the distractors makes the game more interesting and so increases the immersion score or it could be than the combination of game and distractors creates a new experience which is more immersive. The net effect is that the experience of the games becomes more similar which lessens any difference in immersion.

The conclusion of this experiment is that the instructions given to participants can make a large difference to both the distractor recognition paradigm and the immersion score for games they play. Future experiments using the distractor recognition paradigm should continue not to tell participants that they will be tested on the distractors after the game.

Limitations

This experiment set out to provide a more ecologically valid simulation of the experience of playing games in order to replicate the situation where players know they should be attending to some other task but choose to play the game instead. It could be that to see this effect, players would have to play the game for longer than five minutes or play a game of their own choosing rather than one they are given by an experimenter. It could also be argued that a lab experiment is too much of an artificial situation to see this effect and that participants are generally keen to obey all instructions in this kind of setting.

Previous experiments performed analysis of distraction recognition rates between game conditions. For this experiment participants in each condition experienced very similar game experiences so further analysis of the distractors recognised is unlikely to lead to meaningful results.

7.5. Chapter conclusions

The goal of the experiments in this chapter was to investigate whether a distractor recognition paradigm could be used to measure game experience when the distractors were part of the game and participants were looking at them directly. The first experiment in the chapter tested this with two games and a small number of participants. It found that the distractor recognition paradigm provided a measure of game engagement. The second experiment replicated these findings with three different games and a larger number of participants. The effect size in the experiment was reduced but this was probably due to random variation in the participants rather than the additional game condition. These two experiments show that in-game distractors could be used as an effective measure of engagement in self-paced games.

The third experiment investigated the effect of telling participants that they would be tested on the distractors at the end of the game. This had the effect of making participants' game experiences more similar which reduced the differences in both distractors recognised and the immersion scores. The conclusion of the experiment is that for the distractor recognition paradigm to provide an effective measure, participants must not be aware that they will be tested on the distractors after the game.

During these experiments, participants are looking directly at distractor images within the dots so the difference in recognition is not due to whether their gaze is directed at the image. It is more likely to be where their attention is focused. In the game with higher immersion they are focused on the game play, but in the less immersive game their attention wanders and they notice the distractors. This seems very similar to attentional blindness in which participants do not notice a stimulus which they are looking at because their attention is focused on another task.

8. Conclusions

This thesis explored novel ways of measuring the experience of playing self-paced games. To investigate new measures of game experience I took two different approaches which were based on different aspects of game experience. For each approach I performed an iterative sequence of laboratory-based experiments. The first approach looked at using pupil dilation to measure the cognitive load used when playing a game. The second approach looked at measuring how well games hold players' attention by seeing how many irrelevant distractor images they remember after finishing the game. The findings of each approach can be divided into three areas.

1. Findings which relate directly to the aim of the investigation which was to find a new measure of the experience of playing self-paced games.

2. Wider implications for game experience and other game related areas. All measures of experience are based on a model of how that experience is created and which factors are important in making up that experience. So, by creating new measures of game experience I was also creating hypotheses of what that experience involves for the player and then testing them in experiments. If those hypotheses were confirmed then it not only helps create a new measure of game experience it also gave new information about the experience of playing games. This new information had implications beyond the main goals of the investigation.

3. Experimental procedure findings. For each approach to creating a new measure I performed an iterative sequence of experiments, each building on the previous experiment. This allowed me to learn which design features were important in the design of experiments in this area and the difference made by particular design manipulations. Although this knowledge did not impact on the final conclusions of the thesis it may be useful for those trying to replicate these findings and future studies in the same area.

8.1. Findings

The overarching research question of this thesis is *Can new measures be developed to measure the experience of playing self-paced games?* This research question was then broken down into two sub-questions:

- *Can changes in pupil dilation be used to measure the experience of playing self-paced games?*
- *Can the experience of playing self-paced games be measured by seeing how likely players are to notice irrelevant stimuli while playing the game?*

Each of these research questions stimulated a series of iterative experiments. Each series of experiments contributed to answering the question but also generated other research contributions which were not necessarily connected to the related research question.

8.1.1. Can changes in pupil dilation be used to measure the experience of playing self-paced games?

I performed a sequence of experiments looking at using pupil dilation to measure the experience of playing self-paced games. These are described in chapters 4 and 5. The final conclusion of this experiment sequence was that changes in pupil dilation were unlikely to be useful in measuring game experience because participants were not using sustained cognitive load to play the game.

This conclusion was unexpected. The experiments were based on the game of *Two Dots* which is a puzzle game. Cognitive load approximates to the amount of mental effort that someone is putting into solving a problem and it seems intuitive that puzzle games which place a premium on mental effort will require sustained cognitive load. The sequence of experiments that I performed successfully measured cognitive load in non-game tasks and showed a clear difference between low and high cognitive load tasks with very large effect sizes. This technique was then successfully extended to an experiment using tasks taken from the game. These tasks were presented individually without giving participants the opportunity to choose their own tasks as they would when playing a real game. This sequence of experiments showed strong differences in pupil dilation and cognitive load between conditions so I was surprised when examination of pupil dilation during a game showed no significant changes due to cognitive load between different events in the game. The next experiment which compared different versions of the game, one of which had no cognitively challenging elements, showed no significant difference in pupil dilation due to cognitive load. Repeating the experiment did show a significant difference in pupil dilation due to additional motor actions in one of the games but confirmed that there was no significant difference due to cognitive load.

It is possible that additional noise factors in the game experiment prevented the detection of differences in cognitive load. However, the previous experiments had used a similar stimulus and shown a strong effect between conditions which was not noticeably affected by noise. If noise factors in the game experiment had created additional pupil dilation on top of the pupil dilation due to cognitive load, I would have expected to see higher pupil dilation in the game experiment than the previous game-task experiment. In fact, the opposite was the case and overall pupil dilation was much lower in the game experiments than the previous tasks. This suggests that participants were not using sustained cognitive load for any of the different game versions.

The essential difference between participants in the game-like task and those playing the actual game is that players of the game get to choose which moves they will look for and make, whereas those in the game-like task are presented with a set task which they have to solve. This suggests that players in the game could make more difficult high cognitive load moves if they wanted to, but as they had the choice they went with an easier option. Instead of high cognitive load moves they use other less cognitively intensive strategies such as pattern matching or trial and error. If sustained cognitive load is not a key feature of self-paced like *Two Dots*, then measuring cognitive load using pupil dilation is unlikely to be a useful measure of the experience of playing those games.

8.1.2. Wider implications of pupil dilation experiments

The finding that engaging self-paced games like *Two Dots* do not require sustained cognitive load could have wider implications than in measuring game experience. Games such as *Two Dots* are highly engaging so it may be that activities which need sustained cognitive load are less engaging. This could be a useful finding for game designers trying to create engaging activities and also those trying to make more traditional work activities engaging by adding game elements, for example Deterding et al. (2011). This finding is also consistent with load theory (Lavie et al., 2004, Lavie, 2005) which states that under high cognitive load participants are more likely to be distracted. Games like *Two Dots* are highly successful at keeping player's attention which would not be the case if they needed high cognitive load. Research on educational techniques (Sweller, 1988, Paas et al., 2003) has shown that learning is less effective under conditions of high cognitive load. If learning and engagement are both negatively affected by cognitive load then this suggests that those seeking to design engaging learning experiences should avoid situations and tasks which require learners to make sustained cognitive effort.

8.1.3. Pupil dilation experimental procedure findings

In chapter 4 I developed an experimental procedure which showed that pupil dilation can be used to measure cognitive load in a screen based task with a mouse input even though the screen display was not luminance balanced as in some other pupil dilation studies such as Cavanagh et al. (2014). A key factor in this procedure was that each condition showed participants exactly the same visual stimulus, however the

instructions they were given differed between conditions so that participants performed different tasks which required different amounts of cognitive load. Because both conditions showed the same visual stimulus, additional pupil dilation due to light from the stimulus was the same for each condition. The difference in pupil dilation between conditions was solely due to the different tasks that participants had performed. The other important part of the procedure was to take a baseline pupil dilation reading at the start of each trial. All pupil dilation readings during the trial are then measured against this baseline. This procedure minimises confounds due to changes in participants' emotional or arousal state during the study. For example, they may be more excited or anxious at the start of the study or more bored or tired towards the end. Normalising pupil dilation against this baseline reading ensures that only pupil dilation changes which happen within the trial are analysed.

8.1.4. Can the experience of playing self-paced games be measured by seeing how likely players are to notice irrelevant stimuli while playing the game?

The investigations in this thesis provide considerable evidence that the experience of playing self-paced games be measured by seeing how likely players are to notice irrelevant stimuli while playing the game. This achieves the main goal of the thesis by creating a new measure of the experience of playing self-paced games.

To develop this measure, I performed a series of experiments which looked at how well self-paced games hold players' attention. During these experiments I developed a measurement technique that I called the *distractor recognition paradigm.* In repeated experiments this successfully measured the difference in engagement between different variants of the self-paced game *Two Dots.* The *distractor recognition paradigm* is based on the idea that participants who are engaged in a game will focus their attention only on the game but those who are less engaged are more likely to be distracted and their attention will wander. In this paradigm participants play a game which is surrounded by irrelevant distractor images which change every 5 seconds. After the game has finished participants are tested on how many of the previously shown images they can correctly recognise. This was shown to be an effective measure of experience by experiments in chapter 6 which first tested this measure on two very different variants of the game *Two Dots* and then tested it on three much more similar game variants. Two experiments in chapter 7 extended this paradigm by putting the distractor images within the game rather than around the sides. The first experiment had a small number of participants and compared two games, the second experiment had three games and more participants. Both experiments showed that participants who played the more engaging game recognised fewer distractor images. This happens despite the fact that participants in all the game conditions needed to look directly at the distractor images to play the game. I concluded that the number of distractor images recognised is related to where participants' have focused their mental attention rather than simply where they are looking. Participants in the more engaging game focus their attention on the

game whereas those in the less engaging game are more likely to find their attention on the game slipping and so they look at the images and recognise them later.

8.1.5. Wider implications of distraction experiments

These experiments provide evidence that even simple self-paced games like *Two Dots* are highly effective at holding participants' attention. In chapter 6 I used the *distractor recognition paradigm* to measure how well the self-paced game *Two Dots* held participants' attention. In this paradigm the more distractor images that participants recognise the less the game held their attention. In the *Full game* version of *Two Dots* participants' recognition of distractors was no better than they would achieve by chance. This suggests that even a simple self-paced game like *Two Dots* can hold participants' attention completely. This is also consistent with the load theory of attention (Lavie et al., 2004, Lavie, 2005) which proposes that participants are less likely to be distracted when under high perceptual load. *Two Dots* requires players to find patterns in a complex visual stimulus (the grid of dots) so puts players under high perceptual load. Load theory would predict that under high perceptual load they are less likely to be distracted, which is indeed the case.

Two experiments in chapter 6 also used eye tracking to record how often participants looked at the distractor images rather than looking at the game. When compared with the *distractor recognition paradigm,* eye tracking proved to be a weaker measure of engagement than testing how many distractor images participants remembered. This implies that just because participants are looking at a distractor it does not mean that they will remember it if their attention is on the game. This looks very much like the psychological phenomena of *inattentional blindness* (Simons and Chabris, 1999, Most et al., 2001) in which participants are effectively blind to a stimulus if their attention is not on that stimulus. The experiments in chapter 7 investigated this further by putting the distraction images inside the game that participants were playing. The results showed that players who were engaged in the game did not recognise images which they had been shown even though they must have been looking directly at them in order to play the game. This provides further evidence that inattentional blindness is present in games and is more likely to happen if players are highly immersed in the game.

If inattentional blindness is a feature of game experience this has implications for other areas as well as measuring engagement. Serious games are those games which are designed for educational or persuasive purposes (Deterding, 2015, Anderson et al., 2010, Mortara et al., 2014). Designers of serious games often add the serious content they are trying to communicate alongside the gameplay. In games like these, inattentional blindness may mean that players are essentially blind to that additional content because they are focused on the game rather than the content. Inattentional blindness in games also has implications for studies on the effects of video games. These studies typically investigate the effects of violence in video games (Ferguson and Kilburn, 2010, Zendle et al., 2015) or use games to try to change players' behaviour in some other way (Thompson, 2012, Baranowski et al., 2008). Once again, if players are effectively blind to the graphical content of a game then it seems unlikely that they will

be affected either positively or negatively by that graphical content. This is supported by Zendle et al. (2015)'s finding that players of a video game were not primed by the graphic content of the game. This may be because players who were engaged in the game did not pay attention to that graphical content and so were effectively blind to it.

8.1.6. Experimental procedure findings from distraction experiments

I performed a sequence of experiments on attention and distraction in games. These experiments had many findings which were not directly related to the goal of creating a new measure of the experience of playing self-paced games. One of the findings was that using more interesting distractor images did not create a stronger measure of attention. Standing (1973) found that participants were very good at recognising images which they had previously be shown for a short time (5 seconds) with recognition rates of over 90%. This finding was replicated by two experiments in chapter 6. The second experiment found that Disney characters were equally as memorable as monochrome icons from the *Webdings* typeface. This was unexpected, because the Disney characters are more colourful and interesting so I expected them to be more memorable than the icons. However, when Disney characters were used in the *distractor recognition paradigm*, they were more distracting than *Webdings* icons. This had the effect of distracting participants from the *Full game* and making the otherwise less engaging versions of the game more interesting. Showing Disney characters around the game therefore changed the very game experience I was looking to measure. Because this reduced the difference between the experience of playing both games, using Disney distractors is a weaker measure of game experience than *Webdings* icons. I also used eye tracking to measure how often participants looked at the distractor images compared to the game. I found that although participants did appear to look at the game more for the more engaging game variant, this difference was not significant and had a weaker effect than the post-game distractor recognition test. This is likely to be because even though participants in the more engaging game may look at the distractor images their minds are focused on the game so they do not recognise any more distractors.

I also investigated two variants in the experiment setup to see what effect they had on the sensitivity of the distractor recognition paradigm. Both variants prevented the distractor recognition paradigm from being an effective measure. In the first variant the experimenter sat in a different room while the participant played the game and I also removed the chin rest and eye tracking equipment. It is likely that in this variant, participants who were less engaged by the game did look away from it but as they were not constrained by a chin rest, they looked around the room rather than at the distractors so did not remember any more distractors. In the second variant participants were told before the experiment that they would be tested on the distractors afterwards but should still do as well as they could at the game. In this variant it is likely that being told they needed to remember the distractor images impacts on participants' overall game experience. This means that participants are less engaged in the better games and more engaged in the worse ones and so have a more similar experience.

8.2. Limitations

There are a number of limitations to the studies and conclusions in this thesis. Some of these limitations apply to the whole thesis, others to particular studies or outcomes.

8.2.1. Limitations of the whole thesis

All of the game studies in this thesis are done on one game, *Two Dots*, or variants of that game. This had the advantage that it took less time to set up new studies on custom versions of the game because most of the software was already written. It also makes it possible to compare experiments and make links between them. What these studies cannot do is provide evidence that their findings can be generalised to apply to all self-paced games. However, the goal of this thesis was to explore new potential measures of game experience. *Two Dots* is representative of many commonly played self-paced games and if a measure works well on *Two Dots* then there is a good chance that it will work on other similar games.

In all of the game playing experiments participants played the game on a desktop computer in a laboratory for 5 minutes. Most self-paced games, even those in the casual genre, are played for much longer than 5 minutes. However, it should be appreciated that the first few minutes of gameplay are the most critical period for a game's design. If the game manages to engage players in this time then they are likely to continue, so being able to measure engagement from the first 5 minutes of a game can be very useful for designers.

Running experiments on a desktop computer made it easier to control participants' position relative to the screen and implement measures such as eye tracking. The laboratory setting also made it easy to control light levels and minimise external distractions. Many self-paced games are played on mobile devices and even those that are played on desktop computers tend to be played in the more relaxed surroundings of players' homes. This could be seen as an ecological limitation because although participants were playing a real game which has been played by millions of other players, they were not playing it in the usual situation. However, to be useful for designers or researchers it is better for a measure to be usable in a laboratory setting rather than requiring the measure to be used in player's homes. So, this may not actually be an ecological limitation as the measures are being tested in similar surroundings to where they would be used.

8.2.2. Limitations of pupil dilation studies

The final pupil dilation studies in chapter 5 found no significant difference in pupil dilation due to cognitive load between the different versions of the game. The conclusion of the chapter was that this was likely to be because playing the game does not require sustained cognitive load to play. However, this may not be the case. It is

possible that different versions of the game do use different patterns of cognitive load but this was not picked up by the analysis. The experiment which compared easy and hard game tasks in isolation did find a significant difference in cognitive load with a strong effect size. The real game variants had a number of features which make pupil dilation more difficult to measure and may have added confounds to the analysis. These are described below:

Motor actions

In the game-task experiment participants made a single click to indicate their response, but in a real game they have to move the mouse to a dot, click on the dot and then drag the mouse to indicate which dots they want to join. Participants in the *All dots the same* game made more complex mouse movements than the other games, probably because they were trying the make the game more interesting. In a reversal of my original hypothesis I found that participants in this game did have higher pupil dilation, probably because of the more complex mouse movements they made. It is possible that pupil dilation due to cognitive load is over shadowed by pupil dilation due to motor actions such as mouse movements. But as players have to control a game somehow this will always be an issue with interactive games.

Baseline measurements

The initial pupil dilation experiments in chapter 4 showed the importance of normalising against a baseline pupil dilation measure at the start of each trial. In the non-game pupil dilation studies there is a 7 second pause at the end of each trial to allow pupil dilation to relax back to a baseline and then a 2 second period where the baseline measurement is made. In the game pupil dilation studies there is no pause after each move because this would disrupt the game and the baseline measurement is taken as the first 0.5 second of each move. So, when one move begins it is possible the pupils are still dilated from the previous move. This would seem to be less optimal than the procedure in the non-game experiments but there is no evidence that this truncated baseline measurement has introduced confounds. In fact, repeated experiments in chapter 5 found a significant difference in pupil dilation between games which was probably down to differences in the motor actions made by participants. This was despite levels of pupil dilation in the game experiments being much lower than those in the non-game experiments. If the truncated baseline had introduced appreciable noise into the data then it is unlikely that I would have found a significant difference between such small pupil dilation readings.

Emotional reactions

It is also possible that during the game experiments participants have more emotional reactions than during the non-game experiments. In the game experiments participants can succeed or fail at the game and this may cause more of an emotional reaction. This emotional reaction could affect pupil dilation and add noise to the data. However, it is likely that participants in the initial audio stimulus experiment (described in section 4.2) did have a different emotional reaction between the hard and easy task. The evidence for this is the significance difference in baseline pupil dilation between those tasks.

Despite this difference in baseline, the analysis used successfully compensated for this emotional reaction and found a significant difference in pupil dilation due to cognitive load. This suggests that the data analysis used would not be unduly affected by noise due to participants' emotions.

Timing

In the game-task experiment participants viewed the stimulus, made their decision which answer to give and then responded with a single click. When playing the game making a move consists of dragging the mouse to select a group of dots and players may make their decision about which move to make before the move or during it. This makes it more difficult to fix on exactly when they made their decision. So, it is possible that variation in the point at which players make their decision means that the additional pupil dilation due to cognitive load is spread over a wide time period and so not detected by the analysis that I used. However, participants generally made their moves quickly (see section 5.9), for example in the *Full game* participants made their move in an average of 1143ms, which is around a second. Given that pupil dilation due to cognitive load does not change quickly, this uncertainty over time is unlikely to add significant confounds.

Peaks and troughs in the response

It is possible that different game experiences cause different patterns of cognitive load which would create different peaks and troughs in the pupil response. The main analysis comparing games looked at the mean pupil dilation in a particular 1 second time window. These mean value of that time windows across the whole game was then calculated for each participant. It may be that there are differences in cognitive load which cannot be seen by looking only at mean values for the whole game. To investigate this, I performed an exploratory analysis of differences in pupil dilation between different types of game event in the *Full game.* This found no significant differences between different game events. Also, although looking only at mean pupil dilation does lose some of the variation in data, if there are consistent differences between games then this should create a significant difference in means which was not the case.

8.2.3. Limitations of distractor recognition studies

In chapters 6 and 7 I performed repeated experiments to test a new measure of game experience known as the *distractor recognition paradigm.* These experiments found that it is an effective measure of game experience. However, this method has a number of limitations.

Participants cannot know that they will be tested on the distractor images

In the *distractor recognition paradigm* participants play a game which is either surrounded by or contains irrelevant distractor images. Once the game has finished, they are tested to see how many of these images they recognise. A key limitation of this as a method for measuring game engagement is that participants cannot know that they will be tested on the distractor images afterwards. In chapter 7 I tested this by telling the participants

before the game that they would be tested on the images. Participants responded by trying to pay attention to both the game and the distractors. This had the net effect of both changing the game experience and also changing the number of distractors they recognised so there was no significant difference between game conditions. This limitation may make this method unsuitable for real-world game development but may not be an obstacle to using it for focused studies by game designers or for academic studies on attention or game experience.

No pattern of engagement across time

One of the original reasons for creating an on-line measure of game engagement was so that it could show the peaks and troughs of game experience such as the feeling of achievement when a level is completed. I plotted graphs of how the recognition of distractors changed over the time of the trial but did not see any meaningful patterns for any of the games. It is possible that this is because *Two Dots* does not feature many particularly salient features which would lead to peaks and troughs in experience. It is also possible that there were peaks and troughs in the game experience but the time at which they occurred varied for each participant so that looking at mean values for all of the participants did not reveal any pattern. A third possibility is that the distractor recognition paradigm is not a sensitive enough measure to pick up these patterns. Small peaks of engagement caused by a success may be too small to make a difference to the number of distractors recognised. In the post-game distractor test participants are tested on images which were shown 10 seconds apart during the game. It is possible that peaks and troughs in engagement happen at much smaller time scales than this so make no difference to participants recognition of the images.

Weak relation to immersion with generally a lower effect size

Jennett (2010) considered game attention a form of selective attention. To investigate this claim I also measured immersion for all the games that I tested with the *distractor recognition paradigm.* There was no consistent correlation between the number of distractors recognised and the immersion score. However, games which had a higher mean immersion score also had a lower mean number of distractors recognised and vice versa. In general, the effect sizes between immersion scores for different games were higher than the effect sizes between the number of distractors recognised which suggests that the immersion questionnaire may be a more sensitive measure of engagement. Even if this is the case then the *distractor recognition paradigm* could still be useful for situations where participants' self-report is unreliable such as less engaging experiences which do not create full immersion. It could also be used as a measure of attention or to confirm the results from an immersion questionnaire.

Uncertainty about the mechanism for distractor recognition

Although repeated experiments in chapter 6 and chapter 7 show that the number of distractors recognised is inversely related to how engaging the game is, there is still uncertainty about the mechanisms involved. The most likely explanation seems to be that in the engaging games participants' focus of attention is held by the game so they do not look at the distractor images. In the less engaging games participants lose

attentional focus on the game and their attention drifts to look at the distractor images. This additional attention on the images means that they recognise more distractors afterwards. It is unlikely that participants are actually distracted by their inherent interest in the *Webdings* icons. Firstly, the *Webdings* icons are not that interesting and secondly participants almost certainly did find the Disney distractors more interesting and were more distracted from the game which had the effect of lowering their immersion in that game. There was no evidence that the *Webdings* icons had any effect on game immersion. I investigated whether some *Webdings* icons were more distracting than others and found no evidence that this is the case which would further support the idea the participants are not distracted by the icons and the additional recognition is caused by participants becoming bored of the game. Baddeley and Hitch (1974) found that additional cognitive load reduced the effectiveness of memory so it is possible that the more engaging version of *Two Dots* also required more cognitive load which reduced memory and meant that participants recognised fewer distractors. However, the previous pupil dilation experiments found no significant cognitive load was used while playing the game so this seems unlikely.

8.3. Further work

Games are complex systems which are full of uncertainty. Accordingly, the experience of playing games is a complex phenomenon and seeking ways to measure it opens up a wide variety of issues and approaches. Potentially there is wide variety of future work to be done on measuring the experience of playing self-paced games. However, the results from these studies did suggest some future investigations which would be particularly rewarding. These are described below.

8.3.1. Further work on pupil dilation studies

Different games

The experiments in chapter 5 suggested that playing the game of *Two Dots* does not require sustained cognitive load. This was an unexpected finding because *Two Dots* is a puzzle game and I expected it to place a premium on mental effort. It would be interesting to repeat this experiment using different games. These could be other digital self-paced casual games but it would also be particularly interesting to measure cognitive load on more traditional games which have more of an "intellectual" reputation such as chess. Games like chess are also self-paced with fairly simple "graphics" so it would be possible to get participants to play a version on screen and use a similar method to measure pupil dilation. One issue would be that chess is a two-player game and much more complex than *Two Dots* which would make it more difficult to ensure that participants had a similar experience. It would be also be interesting to measure cognitive load in fast-paced action games such as *Dota 2* because they are known for being extremely challenging. Measuring cognitive load would give insight into the nature of that challenge. However, the fast speed, changing graphics

and high arousal of these games would make measuring pupil dilation even more of a challenge than with self-paced games.

Redesign *Two Dots* to remove pupil dilation confounds

The experiments in chapter 5 showed a significant difference in pupil dilation due to cognitive load in a game-like task but not the actual game. This may be because the game-like task was carefully designed to remove possible sources of noise which could add confounds to the pupil dilation data. This included having a 7 second pause after each trial for pupil dilation to relax back to a baseline and also having participants respond by making single click with the mouse. The game experiment procedure was also designed to remove some confounds. However, to keep the gameplay flowing each move was directly after the next one and participants made complex mouse movements to choose their move. It may be possible to design a version of the game which is halfway between these two experiments. This would still feel like a game and give participants a choice of which moves to make but would have long pauses after each move to allow pupil dilation to relax and also some other way of choosing a move which is short and unlikely to add to pupil dilation. This additional experiment would add further evidence to help decide whether playing *Two Dots* does not require sustained cognitive load or whether it just appears that way due to confounds added to the pupil dilation data by the other elements of the game.

Different data analysis

The experiments in chapter 5 did not find a significant difference in pupil dilation due to cognitive load between different games. This may be because playing the game did not use significant cognitive load, however it may also be that the pupil dilation data contained noise from other sources so the analysis used did not pick up differences due to cognitive load. It may be possible to use different types of analysis which would remove this noise. Chapter 5 does contain some exploratory analysis which compared different stages of the game with each other. This did not find any significant differences, however there may be a game events or combinations of game events that I did not consider which relate to differences in cognitive load. For example, it may be possible that moves require higher cognitive load if they have both a long thinking time but only join a small number of dots. Similarly, there may be other ways of analysing the pupil dilation data as a whole which are able to extract the variation due to cognitive load from the rest of the data. For my analysis I looked mainly at mean values across the game, an alternative would be to have looked at the peak pupil dilation values for each condition. More sophisticated mathematical techniques such as Fourier analysis or wavelet transforms may be useful for extracting variation due to cognitive load. Finally, it may be that only some of the participants in the study used sustained cognitive load to play the game, so it may be possible to do more individual analysis of the data to investigate this.

Investigate pupil dilation as a measure of emotion

This thesis investigated using pupil dilation as a measure of cognitive load in games and did not find evidence that it would be effective. However, another aspect of game

experience are the emotional changes that players experience as they succeed or fail at the game's challenges. Pupil dilation has previously been found to change due to participants' emotional state (Beatty and Lucero-Wagoner, 2000). In chapter 5 the exploratory analysis found evidence that pupil dilation was higher just before a success than just before a failure. I expected that participants would try harder (and thus have higher pupil dilation) just before they were about to fail, so this result was surprising. It may be that this pupil dilation was due to players experiencing positive emotion due to succeeding rather than being due to changes in cognitive load. It may be possible to use a different set of experiments to investigate whether pupil dilation can be used as an effective measure of emotion in self-paced games.

Compare game pupil dilation with other sources of cognitive load

It would be possible to perform an experiment to compare the cognitive load used during a game with other sources of cognitive load such as the task to remember 4 numbers used by Kahneman and Beatty (1966). This would involve a 2-factor experiment. 1 factor would be the game played – either the *Full game* with puzzle elements or the *All dots the same* game without puzzle elements. Another factor would be whether participants were also asked to remember 4 numbers at the same time as playing the game. Analysing the results would allow us to find what proportion of variance in the pupil dilation was due to differences in the game and what proportion was due to the additional number task. If there was no significant variation due to the number task, then this would indicate that the additional game elements had added sufficient confounds to the pupil data to prevent an accurate test of pupil dilation due to cognitive load. On the other hand, if the number task was responsible for significantly more variation than the type of game then this would indicate that the difference in pupil dilation due to cognitive load between those games was less than that required by a simple number task and so not a key part of the experience of playing the game.

8.3.2. Further work on distractor recognition studies

Different games

All of the experiments with the *distractor recognition paradigm* in chapter 6 and chapter 7 were done on the game *Two Dots,* which participants played for 5 minutes. To be an effective measure of game experience the paradigm needs to be tested with other games. Real games are designed to be played for longer than 5 minutes so the paradigm should also be tested for longer than this. Standing (1973) found that participants could recognise thousands of images that they had been shown over a long period of time which suggests that the paradigm should be able to measure longer experiences. The measure could also be tested with games which are not self-paced. However, it may be that even less engaging action games hold attention so well that participants do not recognise any of the distractor images which would prevent the measure being effective.

Test game mechanics

The *distractor recognition paradigm* could be used to measure how well individual game mechanics hold player's attention. For example, games commonly use design patterns such as creating uncertainty and then resolving it (Costikyan, 2013) or giving players meaningful choices (Cardona-Rivera et al., 2014). The *distractor recognition paradigm* could be used to measure how well these mechanics hold players' attention or to tune them to provide the optimum design. This would involve creating game-like activities which consisted of just the particular game mechanics which were being investigated and then testing how well they hold players' attention using the *distractor recognition paradigm.*

Investigate the mechanism behind the distractor recognition paradigm

In the *Limitations* section of this chapter I discussed how there is uncertainty about how the *distractor recognition paradigm* works. Although it is likely that players do not recognise the distractors because their attention is focused on the game it is possible that there is some other mechanism at work such as increased cognitive load which prevents participants from remembering the distractors even though they were looking at them. It would be possible to perform a new experiment to investigate this. This experiment would create two versions of the *Full game* of *Two Dots*. Both games would have the same gameplay mechanism so they would have similar cognitive load demands. One version would contain in-game distractor images but ask players to join dots of the same colour and ignore the distractors. The other version would have the same gameplay mechanism but ask players to join dots containing the same distractor image and ignore the colours. So, one game would focus players' attention on the distractor images and the other would focus the attention on the colours. I would expect participants who were focusing on the images to recognise more of them afterwards. If the results show that this is the case the difference is likely to be due to a focus of attention. However, if both games show the same distractor recognition rate then the distractor recognition rates may be related to a factor that both games have in common such as the amount of cognitive load used.

Investigate how focussing attention in games can be used for other purposes

The *distractor recognition paradigm* relies on the finding that participants who are more engaged in the game are less likely to be distracted and look at irrelevant images within or around the game. However, it may be possible to design the game in such a way that more engaged players focus their attention on particular aspects or elements of the game. The distraction experiments in this thesis suggest that if players focus their attention on a particular game element, they are more likely to remember that element after the game. If shown to be correct this could be useful in the design of serious games which aim to teach players particular facts or concepts. Games could be designed to focus players' attention on the information that needed to be learnt which would increase learning and the effectiveness of the serious game.

8.4. Final thoughts

This thesis set out to find new ways of measuring the experience of playing self-paced games. It considered two possible methods; measuring pupil dilation and measuring attention using distractor images. The evidence from these experiments suggests that pupil dilation is unlikely to be a useful measure of game experience but that distractor recognition be can used, albeit with some caveats. However, what these experiments have also provided is new insights into the nature of games and why we play them.

These experiments show that playing games probably does not require sustained cognitive load. This provides an explanation why we prefer to play games rather than other tasks such as filling in our tax return. If they do not require sustained cognitive load it may be that games are simply easier than other more demanding activities. However, these experiments also show that games are really good at holding our attention, even if they are self-paced games like *Two Dots*. This ability to hold our attention is another reason why we play games. This holding of attention helps us to think about things we do want to think about such as saving the world, being a hero or just winning the game. It may also help us stop thinking about things in our lives which we do not like and do not want to think about. Game writers such as Salen and Zimmerman (2004) and Huizinga (1938/2014) have referred to a "Magic circle" which is the space where play happens. These studies suggest that this circle is created by attention, when we play a game we choose to focus our attention within that circle and ignore the world outside of it. By choosing to play a game we are choosing where we want to put our attention and games are machines that help us focus that attention on our chosen subject. The games that we chose and the people we chose to play them with determine where our attention will be focused. Let us hope we choose well.

9. Appendix: Consent form

This is an example of the consent form that all experimental participants signed before starting the experiment. The form used was almost identical for each experiment. The only differences were in sections 2 and 3 which describe the purpose and content of the experiment.

Informed Consent – Study IGD3

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

1 Who is running this?

The study is being run by Joe Cutting, who is a PhD student in the department of Computer Science at the University of York, as part of the research for his PhD.

2 What is the purpose of this study?

The study aims to investigate your experience of playing a simple computer game.

3 What will I have to do?

You'll play a simple computer game for 5 minutes. The game contains pictures. After playing the game you'll be tested on the pictures and fill in a questionnaire.

4 Who will see this data?

The experimenter with you (Joe Cutting) will see this data. Joe's supervisor Dr Paul Cairns may also see the data. Joe will process experiment data for further analysis. However, once it has been processed, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way. Your data will be stored in a secure location which only Joe will have access to.

5 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

6 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to the experimenter and save any questions you may have until the end of the study.

7 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood these conditions and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

Email:

10. References

ABUHAMDEH, S., CSIKSZENTMIHALYI, M. & JALAL, B. 2015. Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation. *Motivation and Emotion,* 39**,** 1-10.

ADAMS, E. 2014. *Fundamentals of game design*, Pearson Education.

AGARWAL, R. & KARAHANNA, E. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly***,** 665-694.

AMBINDER, M. Biofeedback in gameplay: How valve measures physiology to enhance gaming experience. 2011.

AMUSEMENT VISION 2001. Super Monkey Ball. Tokyo, Japan.

ANDERSON, E. F., MCLOUGHLIN, L., LIAROKAPIS, F., PETERS, C., PETRIDIS, P. & DE FREITAS, S. 2010. Developing serious games for cultural heritage: a state-of-the-art review. *Virtual reality,* 14**,** 255-275.

APP ANNIE 2018. The Data Behind 10 Years of the iOS App Store.

ASTON-JONES, G. & COHEN, J. D. 2005. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.,* 28**,** 403-450.

BADDELEY, A. 2013. *Essentials of human memory (classic edition)*, Psychology Press.

BADDELEY, A. D. & HITCH, G. 1974. Working memory. *Psychology of learning and motivation.* Elsevier.

BARANOWSKI, T., BUDAY, R., THOMPSON, D. I. & BARANOWSKI, J. 2008. Playing for real: video games and stories for health-related behavior change. *American journal of preventive medicine,* 34**,** 74-82. e10.

BEATTY, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin,* 91**,** 276.

BEATTY, J. & LUCERO-WAGONER, B. 2000. The pupillary system. *Handbook of psychophysiology,* 2**,** 142-162.

BERENBAUM, H., BREDEMEIER, K. & THOMPSON, R. J. 2008. Intolerance of uncertainty: Exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. *Journal of Anxiety Disorders,* 22**,** 117-125.

BINDA, P. & MURRAY, S. O. 2014. Keeping a large-pupilled eye on high-level visual processing. *Trends in cognitive sciences*.

BLYTHE, M. A., OVERBEEKE, K. & MONK, A. F. 2004. *Funology: from usability to enjoyment*, Springer Science & Business Media.

BOLLS, P. D., LANG, A. & POTTER, R. F. 2001. The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research,* 28**,** 627-651.

BRADY, T. F., KONKLE, T., ALVAREZ, G. A. & OLIVA, A. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences,* 105**,** 14325-14329.

BROCKMYER, J. H., FOX, C. M., CURTISS, K. A., MCBROOM, E., BURKHART, K. M. & PIDRUZNY, J. N. 2009. The development of the Game Engagement Questionnaire:

A measure of engagement in video game-playing. *Journal of Experimental Social Psychology,* 45**,** 624-634.
BROWN, E. & CAIRNS, P. A grounded investigation of game immersion. CHI'04 extended abstracts on Human factors in computing systems, 2004. ACM, 1297-1300.
BURNS, C. G. & FAIRCLOUGH, S. H. 2015. Use of auditory event-related potentials to measure immersion during a computer game. *International Journal of Human-Computer Studies,* 73**,** 107-114.
CAILLOIS, R. 1961. *Man, play, and games*, University of Illinois Press.
CAIRNS, P., COX, A. & NORDIN, A. I. 2014a. Immersion in digital games: a review of gaming experience research. *Handbook of digital games, MC Angelides and H. Agius, Eds. Wiley-Blackwell***,** 339-361.
CAIRNS, P., COX, A. L., DAY, M., MARTIN, H. & PERRYMAN, T. 2013. Who but not where: The effect of social play on immersion in digital games. *International Journal of Human-Computer Studies,* 71**,** 1069-1077.
CAIRNS, P., LI, J., WANG, W. & NORDIN, A. I. The influence of controllers on immersion in mobile games. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014b. ACM, 371-380.
CALLEJA, G. 2007. Digital game involvement a conceptual model. *Games and culture,* 2**,** 236-260.
CALLEJA, G. 2011. *In-game: From immersion to incorporation*, MIT Press.
CALVILLO-GÁMEZ, E. H. & CAIRNS, P. Pulling the strings: A theory of puppetry for the gaming experience. Conference Proceedings of the Philosophy of Computer Games, 2008. 308-323.
CALVILLO-GÁMEZ, E. H., CAIRNS, P. & COX, A. L. 2010. Assessing the core elements of the gaming experience. *Evaluating user experience in games.* Springer.
CAMPBELL, J. 1987. The hero's journey.
CARDONA-RIVERA, R. E., ROBERTSON, J., WARE, S. G., HARRISON, B. E., ROBERTS, D. L. & YOUNG, R. M. Foreseeing Meaningful Choices. AIIDE, 2014.
CARTWRIGHT-FINCH, U. & LAVIE, N. 2007. The role of perceptual load in inattentional blindness. *Cognition,* 102**,** 321-340.
CAVANAGH, J. F., WIECKI, T. V., KOCHAR, A. & FRANK, M. J. 2014. Eye Tracking and Pupillometry Are Indicators of Dissociable Latent Decision Processes.
CHARMAZ, K. & BELGRAVE, L. L. 2007. Grounded theory. *The Blackwell encyclopedia of sociology*.
CHATFIELD, T. 2009. Videogames now outperform Hollywood movies. *The Observer*, Sunday 27th Sept
COOK, D. 2012. *Loops and Arcs* [Online]. Available: http://www.lostgarden.com/2012/04/loops-and-arcs.html [Accessed].
COSTA, VINCENT D. & RUDEBECK, PETER H. 2016. More than Meets the Eye: the Relationship between Pupil Size and Locus Coeruleus Activity. *Neuron,* 89**,** 8-10.
COSTIKYAN, G. 2013. *Uncertainty in games*, Mit Press.
CROOK, J. 2014. Two Dots, The Sequel To Betaworks' Dots, Is A Beautiful Monster. Available: http://techcrunch.com/2014/05/31/two-dots-the-sequel-to-betaworks-dots-is-a-beautiful-monster/.
CSIKSZENTMIHALYI, M. 1991. *Flow: The psychology of optimal experience*, HarperPerennial New York.
CSIKSZENTMIHALYI, M. 2013. *Flow: The psychology of happiness*, Random House.
CSIKSZENTMIHALYI, M. & LARSON, R. 1987. Validity and reliability of the Experience-Sampling Method. *The Journal of nervous and mental disease,* 175**,** 526-536.
DAVIES, N. 2011. *Flat Earth news: an award-winning reporter exposes falsehood, distortion and propaganda in the global media*, Random House.

DAVIS, I. 1994. Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation, space, and environmental medicine*.

DE FOCKERT, J. W., REES, G., FRITH, C. D. & LAVIE, N. 2001. The role of working memory in visual selective attention. *Science,* 291**,** 1803-1806.

DE GEE, J. W., KNAPEN, T. & DONNER, T. H. 2014. Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences,* 111**,** E618-E625.

DECI, E. L. & RYAN, R. M. 2008. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian psychology/Psychologie canadienne,* 49**,** 182.

DENISOVA, A. & CAIRNS, P. First Person vs. Third Person Perspective in Digital Games: Do Player Preferences Affect Immersion? Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015. ACM, 145-148.

DENISOVA, A., GUCKELSBERGER, C. & ZENDLE, D. Challenge in digital games: Towards developing a measurement tool. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017. ACM, 2511-2519.

DENISOVA, A., NORDIN, A. I. & CAIRNS, P. The convergence of player experience questionnaires. Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play, 2016. ACM, 33-37.

DETERDING, S. 2015. The lens of intrinsic skill atoms: A method for gameful design. *Human–Computer Interaction,* 30**,** 294-335.

DETERDING, S., SICART, M., NACKE, L., O'HARA, K. & DIXON, D. Gamification. using game-design elements in non-gaming contexts. CHI'11 extended abstracts on human factors in computing systems, 2011. ACM, 2425-2428.

DEUBEL, H. & SCHNEIDER, W. X. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research,* 36**,** 1827-1837.

DREDGE, S. 2014. Why is Candy Crush Saga so popular? *The Guardian*, 26 March 2014.

DREW, T., VÕ, M. L.-H. & WOLFE, J. M. 2013. The invisible gorilla strikes again sustained inattentional blindness in expert observers. *Psychological science,* 24**,** 1848-1853.

DUNCAN, J. & HUMPHREYS, G. W. 1989. Visual search and stimulus similarity. *Psychological review,* 96**,** 433.

EA SPORTS 2002. NHL 2003. Burnaby, British Columbia, Canada.

EINHÄUSER, W., STOUT, J., KOCH, C. & CARTER, O. 2008. Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences,* 105**,** 1704-1709.

FANG, X., CHAN, S., BRZEZINSKI, J. & NAIR, C. 2010. Development of an instrument to measure enjoyment of computer game play. *Intl. Journal of Human–Computer Interaction,* 26**,** 868-886.

FERGUSON, C. J. & KILBURN, J. 2010. Much ado about nothing: The misestimation and overinterpretation of violent video game effects in Eastern and Western nations: Comment on Anderson et al.(2010).

FIRAXIS 2012. X-COM: Enemy Unknown. Sparks, Maryland, USA.

GAGL, B., HAWELKA, S. & HUTZLER, F. 2011. Systematic influence of gaze position on pupil size measurement: analysis and correction. *Behavior research methods,* 43**,** 1171-1181.

GILZENRAT, M. S., NIEUWENHUIS, S., JEPMA, M. & COHEN, J. D. 2010. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience,* 10**,** 252-269.

GOW, J., CAIRNS, P., COLTON, S., MILLER, P. & BAUMGARTEN, R. Capturing player experience with post-game commentaries. Proc. 3rd Int. Conf. on Computer Games, Multimedia & Allied Technologies, 2010. Citeseer.

GRIP, T. 2010. Where is your self in a game? *In the games of madness* [Online]. Available from: http://frictionalgames.blogspot.co.uk/2010/09/where-is-your-self-in-game.html.
GSC GAME WORLD 2007. S.T.A.L.K.E.R.: Shadow of Chernobyl Kiev, Ukraine.
HARRIGAN, K. A., COLLINS, K., DIXON, M. J. & FUGELSANG, J. Addictive gameplay: What casual game designers can learn from slot machine research. Proceedings of the International Academic Conference on the Future of Game Design and Technology, 2010. ACM, 127-133.
HOLLAND, S. & MORSE, D. R. 2001. Audiogps: Spatial audio in a minimal attention interface.
HOSSAIN, G. & ELKINS, J. 2016. When does an easy task become hard? A systematic review of human task-evoked pupillary dynamics versus cognitive efforts. *Neural Computing and Applications*, 1-15.
HOWSON, G. 2008. The Sims is biggest selling PC game 'franchise' ever. *The Guardian*.
HUDSON, M. & CAIRNS, P. 2014. Interrogating social presence in games with experiential vignettes. *Entertainment Computing,* 5**,** 101-114.
HUHTALA, J., ISOKOSKI, P. & OVASKA, S. The usefulness of an immersion questionnaire in game development. CHI'12 Extended Abstracts on Human Factors in Computing Systems, 2012. ACM, 1859-1864.
HUIZINGA, J. 2014. *Homo Ludens* Routledge.
HUNICKE, R., LEBLANC, M. & ZUBEK, R. MDA: A formal approach to game design and game research. Proceedings of the AAAI Workshop on Challenges in Game AI, 2004.
IGN STAFF. 2001. *You got game, but can you write?* [Online]. Available: http://uk.ign.com/articles/2001/03/22/you-got-game-but-can-you-write [Accessed].
IJSSELSTEIJN, W., DE KORT, Y., POELS, K., JURGELIONIS, A. & BELLOTTI, F. Characterising and measuring user experiences in digital games. 2007.
IQBAL, S. T., ZHENG, X. S. & BAILEY, B. P. Task-evoked pupillary response to mental workload in human-computer interaction. CHI'04 extended abstracts on Human factors in computing systems, 2004. ACM, 1477-1480.
JAINTA, S. & BACCINO, T. 2010. Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology,* 77**,** 1-7.
JENNETT, C., COX, A. L. & CAIRNS, P. Investigating computer game immersion and the component real world dissociation. CHI'09 Extended Abstracts on Human Factors in Computing Systems, 2009. ACM, 3407-3412.
JENNETT, C., COX, A. L., CAIRNS, P., DHOPAREE, S., EPPS, A., TIJS, T. & WALTON, A. 2008. Measuring and defining the experience of immersion in games. *International journal of human-computer studies,* 66**,** 641-661.
JENNETT, C. I. 2010. *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion.* UCL (University College London).
JOSHI, S., LI, Y., KALWANI, R. M. & GOLD, J. I. 2015. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*.
JUST, M. A. & CARPENTER, P. A. 1993. The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale,* 47**,** 310.
JUUL, J. 2013. *The art of failure: An essay on the pain of playing video games*, Mit Press.
KAHNEMAN, D. & BEATTY, J. 1966. Pupil diameter and load on memory. *Science,* 154**,** 1583-1585.

KAHNEMAN, D., FREDRICKSON, B. L., SCHREIBER, C. A. & REDELMEIER, D. A. 1993. When more pain is preferred to less: Adding a better end. *Psychological science,* 4**,** 401-405.
KAMEHAN STUDIOS 2002. Tactical Ops: Assault on Terror. Paris, France.
KATIDIOTI, I., BORST, J. P. & TAATGEN, N. A. 2014. What happens when we switch tasks: Pupil dilation in multitasking. *Journal of experimental psychology: applied,* 20**,** 380.
KAYE, L. K., MONK, R. L., WALL, H. J., HAMLIN, I. & QURESHI, A. W. 2018. The effect of flow and context on in-vivo positive mood in digital gaming. *International Journal of Human-Computer Studies,* 110**,** 45-52.
KINOSHITA, S. 1995. The word frequency effect in recognition memory versus repetition priming. *Memory & Cognition,* 23**,** 569-580.
KIVIKANGAS, J. M., CHANEL, G., COWLEY, B., EKMAN, I., SALMINEN, M., JÄRVELÄ, S. & RAVAJA, N. 2011. A review of the use of psychophysiological methods in game research. *journal of gaming & virtual worlds,* 3**,** 181-199.
KLINE, P. 2013. *Handbook of psychological testing*, Routledge.
KOSTER, R. 2013. *Theory of fun for game design*, " O'Reilly Media, Inc.".
KULTIMA, A. Casual game design values. Proceedings of the 13th international MindTrek conference: Everyday life in the ubiquitous era, 2009. ACM, 58-65.
KULTIMA, A. & STENROS, J. Designing games for everyone: the expanded game experience model. Proceedings of the International Academic Conference on the Future of Game Design and Technology, 2010. ACM, 66-73.
LANG, P. J., GREENWALD, M. K., BRADLEY, M. M. & HAMM, A. O. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology,* 30**,** 261-261.
LAVIE, N. 2005. Distracted and confused?: Selective attention under load. *Trends in cognitive sciences,* 9**,** 75-82.
LAVIE, N., HIRST, A., DE FOCKERT, J. W. & VIDING, E. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General,* 133**,** 339.
LAW, E. L.-C., BRÜHLMANN, F. & MEKLER, E. D. Systematic Review and Validation of the Game Experience Questionnaire (GEQ)-Implications for Citation and Reporting Practice. Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, 2018. ACM, 257-270.
LAZZARO, N. 2004. Why we play games: Four keys to more emotion without story.
LEVINSON, D. B., SMALLWOOD, J. & DAVIDSON, R. J. 2012. The persistence of thought evidence for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science,* 23**,** 375-380.
LINDEN LAB 2003. Second Life. San Francisco, USA.
LLERAS, A., CHU, H. & BUETTI, S. 2017. Can we "apply" the findings of Forster and Lavie (2008)? On the generalizability of attentional capture effects under varying levels of perceptual load. *Journal of Experimental Psychology: Applied,* 23**,** 158.
LOEWENFELD, I. E. & LOWENSTEIN, O. 1993. *The pupil: Anatomy, physiology, and clinical applications*, Iowa State University Press Ames.
LOEWENSTEIN, G. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin,* 116**,** 75.
LOMBARD, M. & DITTON, T. 1997. At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication,* 3**,** 0-0.
MALLE, B. F. & KNOBE, J. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology,* 33**,** 101-121.
MANDRYK, R. L. & ATKINS, M. S. 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International journal of human-computer studies,* 65**,** 329-347.

MARSHALL, S. P. The index of cognitive activity: Measuring cognitive workload. Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on, 2002. IEEE, 7-5-7-9.
MARSHALL, S. P. 2007. Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine,* 78**,** B165-B175.
MAXIS 2004. The Sims 2. Redwood Shores, California, U.S.
MÄYRÄ, F. & ERMI, L. 2005. Fundamental components of the gameplay experience: analysing immersion.
MCALLISTER, G., MIRZA-BABAEI, P. & AVENT, J. 2013. Improving gameplay with game metrics and player metrics. *Game Analytics.* Springer.
MCCARTHY, J. & WRIGHT, P. 2004. Technology as experience. *interactions,* 11**,** 42-43.
MCMILLAN, L. 2013. *The Rational Design Handbook: An Intro to RLD* [Online]. Gamasutra. Available: http://gamasutra.com/blogs/LukeMcMillan/20130806/197147/The_Rational_Design_Handbook_An_Intro_to_RLD.php [Accessed 4th September 2015].
MEHRABIAN, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology,* 14**,** 261-292.
MEKLER, E. D., BOPP, J. A., TUCH, A. N. & OPWIS, K. A systematic review of quantitative studies on the enjoyment of digital entertainment games. Proceedings of the 32nd annual ACM conference on Human factors in computing systems, 2014. ACM, 927-936.
MICROPROSE 1991. Civilization. Maryland, US.
MILLER, J. D. & TANIS, D. C. 1971. Recognition memory for common sounds. *Psychonomic Science,* 23**,** 307-308.
MORESI, S., ADAM, J. J., RIJCKEN, J., VAN GERVEN, P. W., KUIPERS, H. & JOLLES, J. 2008. Pupil dilation in response preparation. *International Journal of Psychophysiology,* 67**,** 124-130.
MORTARA, M., CATALANO, C. E., BELLOTTI, F., FIUCCI, G., HOURY-PANCHETTI, M. & PETRIDIS, P. 2014. Learning cultural heritage by serious games. *Journal of Cultural Heritage,* 15**,** 318-325.
MOST, S. B., SIMONS, D. J., SCHOLL, B. J., JIMENEZ, R., CLIFFORD, E. & CHABRIS, C. F. 2001. How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science,* 12**,** 9-17.
MURNANE, K. 2016. 'Civilization VI': Three Innovations That Make Sid Meier's Masterpiece Even More Compelling. *Forbes.*
NABI, R. L. & KRCMAR, M. 2004. Conceptualizing media enjoyment as attitude: Implications for mass media effects research. *Communication Theory,* 14**,** 288-310.
NACKE, L. 2009. Affective ludology: Scientific measurement of user experience in interactive entertainment.
NACKE, L. E., STELLMACH, S. & LINDLEY, C. A. 2010. Electroencephalographic assessment of player experience: A pilot study in affective ludology. *Simulation & Gaming*.
NINTENDO 1996. Super Mario 64. Tokyo, Japan.
NINTENDO 1998. The Legend of Zelda: Ocarina of Time. Tokyo, Japan.
NORDIN, A. I., ALI, J., ANIMASHAUN, A., ASCH, J., ADAMS, J. & CAIRNS, P. Attention, time perception and immersion in games. CHI'13 Extended Abstracts on Human Factors in Computing Systems, 2013. ACM, 1089-1094.
NORDIN, A. I., CAIRNS, P., HUDSON, M., ALONSO, A. & CALVILLO, E. H. The Effect Of Surroundings On Gaming Experience. Proceedings of the 9th International

Conference on the Foundations of Digital Games (FDG 2014). Society for the Advancement of the Science of Digital Games, 2014. Citeseer.
O'BRIEN, H. L. & TOMS, E. G. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology,* 59**,** 938-955.
PAAS, F., RENKL, A. & SWELLER, J. 2003. Cognitive load theory and instructional design: Recent developments. *Educational psychologist,* 38**,** 1-4.
PAIVIO, A. & CSAPO, K. 1971. Short-term sequential memory for pictures and words. *Psychonomic Science,* 24**,** 50-51.
PHILLIPS, B. 2006. Talking about games experiences: A view from the trenches. *interactions,* 13**,** 22-23.
PINKER, S. 2014. *The sense of style: The thinking person's guide to writing in the 21st century*, Penguin.
PLAYDOTS, I. 2014. Two Dots. New York, USA.
POELS, K., DE KORT, Y. & IJSSELSTEIJN, W. It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology. Proceedings of the 2007 conference on Future Play, 2007. ACM, 83-89.
POWER, C., CAIRNS, P., DENISOVA, A., PAPAIOANNOU, T. & GULTROM, R. 2018. Lost at the Edge of Uncertainty: Measuring Player Uncertainty in Digital Games. *International Journal of Human–Computer Interaction***,** 1-13.
QIN, H., PATRICK RAU, P.-L. & SALVENDY, G. 2009. Measuring player immersion in the computer game narrative. *Intl. Journal of Human–Computer Interaction,* 25**,** 107-133.
RAVAJA, N. 2004. Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology,* 6**,** 193-235.
RAVAJA, N. 2009. The psychophysiology of digital gaming: The effect of a non co-located opponent. *Media Psychology,* 12**,** 268-294.
RAVAJA, N., SAARI, T., SALMINEN, M., LAARNI, J. & KALLINEN, K. 2006. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology,* 8**,** 343-367.
RAVEH, D. & LAVIE, N. 2015. Load-induced inattentional deafness. *Attention, Perception, & Psychophysics,* 77**,** 483-492.
REES, G., RUSSELL, C., FRITH, C. D. & DRIVER, J. 1999. Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science,* 286**,** 2504-2507.
REISENZEIN, R. 1994. Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology,* 67**,** 525.
RENSINK, R. A. 2015. A function-centered taxonomy of visual attention. *Phenomenal Qualities: Sense, Perception, and Consciousness***,** 31.
RICHER, F. & BEATTY, J. 1985. Pupillary dilations in movement preparation and execution. *Psychophysiology,* 22**,** 204-207.
ROVIO 2009. Angry Birds. Espoo, Finland: Rovio.
RUSSELL, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology,* 39**,** 1161.
RUSSELL, J. A., WEISS, A. & MENDELSOHN, G. A. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology,* 57**,** 493.
RYAN, R. M., RIGBY, C. S. & PRZYBYLSKI, A. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion,* 30**,** 344-360.
SALEN, K. & ZIMMERMAN, E. 2004. *Rules of play: Game design fundamentals*, MIT press.
SANDERS, T. & CAIRNS, P. Time perception, immersion and music in videogames. Proceedings of the 24th BCS Interaction Specialist Group Conference, 2010. British Computer Society, 160-167.
SCHELL, J. 2008. *The Art of Game Design: A book of lenses*, CRC Press.

SEAH, M.-L. & CAIRNS, P. From immersion to addiction in videogames. Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1, 2008. British Computer Society, 55-63.
SHEPHERD, M., FINDLAY, J. M. & HOCKEY, R. J. 1986. The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology,* 38**,** 475-491.
SHINE, J. M., BISSETT, P. G., BELL, P. T., KOYEJO, O., BALSTERS, J. H., GORGOLEWSKI, K. J., MOODIE, C. A. & POLDRACK, R. A. 2016. The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron,* 92**,** 544-554.
SILVIA, P. J. 2006. *Exploring the psychology of interest*, Oxford University Press.
SIMONS, D. J. & CHABRIS, C. F. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London,* 28**,** 1059-1074.
SINCLAIR, J., HINGSTON, P. & MASEK, M. Considerations for the design of exergames. Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia, 2007. ACM, 289-295.
SLATER, M., USOH, M. & STEED, A. 1994. Depth of presence in virtual environments. *Presence,* 3**,** 130-144.
SLEEPING BEAST GAMES 2012. Spaceteam. Montréal, Canada.
SMALLWOOD, J., MCSPADDEN, M. & SCHOOLER, J. W. 2008. When attention matters: The curious incident of the wandering mind. *Memory & Cognition,* 36**,** 1144-1150.
SONY STUDIO LIVERPOOL 2009. Wipeout HD Fury. Liverpool, UK.
STANDING, L. 1973. Learning 10000 pictures. *The Quarterly journal of experimental psychology,* 25**,** 207-222.
SWEETSER, P., JOHNSON, D., WYETH, P. & OZDOWSKA, A. GameFlow heuristics for designing and evaluating real-time strategy games. Proceedings of the 8th Australasian Conference on Interactive Entertainment: Playing the System, 2012. ACM, 1.
SWEETSER, P. & WYETH, P. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE),* 3**,** 3-3.
SWELLER, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science,* 12**,** 257-285.
TATUM, W. O. 2014. Ellen R. Grass Lecture: Extraordinary EEG. *The Neurodiagnostic Journal,* 54**,** 3-21.
THOMPSON, D. 2012. Designing serious video games for health behavior change: current status and future directions. SAGE Publications.
THOMPSON, M., NORDIN, A. I. & CAIRNS, P. Effect of touch-screen size on game immersion. Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers, 2012. British Computer Society, 280-285.
TRAVELLERS TALES 1998. A Bug's Life. Knutsford, England.
UNSWORTH, N. & ROBISON, M. K. 2018. Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective, & Behavioral Neuroscience,* 18**,** 638-664.
VALVE 1998. Half-Life. Bellevue, Washington, U.S.
VALVE 2004. Half-Life 2. Bellevue, Washington, USA: 2004.
VALVE SOUTH 2008. Left 4 Dead. Lake Forest, California, USA.
VAN DEN HOOGEN, W. M., IJSSELSTEIJN, W. A. & DE KORT, Y. A. Effects of sensory immersion on behavioural indicators of player experience: movement synchrony and controller pressure. Breaking new ground: innovation in games, play, practice and theory. Proceedings of the 2009 Digital Games Research Association Conference. London: Brunel University, 2009. 1-6.

WEBER, R., BEHR, K. M., TAMBORINI, R., RITTERFELD, U. & MATHIAK, K. 2009. What do we really know about first-person-shooter games? An event-related, high-resolution content analysis. *Journal of Computer-Mediated Communication,* 14**,** 1016-1037.

WEBSTER, D. M. & KRUGLANSKI, A. W. 1997. Cognitive and social consequences of the need for cognitive closure. *European review of social psychology,* 8**,** 133-173.

WEINBERG, G. L., LANGER, P. L., LYNCH, T., CHEN, V. S., KÖNIG, L., JAROSZ, S. P. & KNOWLES, A. 2013. Low-attention gestural user interface. Google Patents.

WOLFE, J. M. 2014. Approaches to visual search: Feature integration theory and guided search. *The Oxford handbook of attention,* 11**,** 35-44.

YANNAKAKIS, G. N. & HALLAM, J. 2008. Entertainment modeling through physiology in physical play. *International Journal of Human-Computer Studies,* 66**,** 741-755.

ZENDLE, D., CAIRNS, P. & KUDENKO, D. Higher Graphical Fidelity Decreases Players' Access to Aggressive Concepts in Violent Video Games. Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play, 2015. ACM, 241-251.

ZENDLE, D., CAIRNS, P. & KUDENKO, D. 2018. No priming in video games. *Computers in Human Behavior,* 78**,** 113-125.