Let the Gamers Do the Talking: A Comparative Study of Two Usability Testing Methods for Video Games

Efstratios Theodorou

Project report submitted in part fulfilment of the requirements for the degree of Master of Science (Human-Computer Interaction with Ergonomics) in the Faculty of Life Sciences, University College London, 2010.

Acknowledgements

I cannot thank enough my supervisor Duncan Brumby for his exceptional guidance throughout the project and for keeping me on track when I couldn't do it myself.

Many thanks to David Tisserand at SCEE for having the original idea for this project in the first place, for his always pertinent comments and for his professional insights.

Special thanks to the participants of my experiment, for without them this project wouldn't be possible. I certainly hope that they enjoyed their participation as I did. I am still laughing with some of their quotes.

To my friends, for all the times I said "no" this summer I want to say that I'm sorry. Starting from tomorrow, I will only say "yes"!

Of course, I cannot forget my family. For your moral and financial support this whole year of my studies (and all of my life actually), for your understanding when I was not calling for many days because I was lost in my project, for your love, thank you.

I could go on thanking people forever, but I'd better stop. Enjoy reading!

Stratos Theodorou, 07 September 2010

Abstract

Usability testing is an established practice in usability evaluation for video games and various techniques are used, including concurrent and retrospective verbal protocols. However, literature that compares these techniques is scarce or irrelevant to the domain of video games. The question is raised: Is it worth interfering with the gameplay by asking the player to think aloud when conducting usability testing on video games? A study is reported in which participants were asked to either play a game while thinking aloud, or play the game silently and participate in an interview afterwards. The results showed that participants who were thinking aloud reported significantly more problems than those who participated in interviews. Moreover, the cost for applying the methods was similar; therefore think-aloud protocol was more cost effective. It is suggested that usability practitioners adopt think-aloud protocol when conducting usability testing on video games, especially when only a small amount of users is available. The study also attempts to ignite the spark for a series of comparative usability testing studies on video games.

Contents

1 Introduction

"We needed help. We were 35-year-old hardcore gamers trying to see the game through the eyes of third-grade unicorn lovers. And then help came in the form of playtesting." (Andree-Anne Boisvert1, 2010)

User testing is an established practice in usability evaluation for video games (Pagulayan, Keeker, Wixon, Romero, & Fuller, 2007) and the above quote describes in the best possible manner why testing with real users is so important: the designers are not the users.

While many studies have compared user testing with analytical methods (Desurvire, Caplan, & Toth, 2004; Korhonen, 2010; and many more), there is a gap in HCI literature comparing the different techniques that are available for laboratory-based usability testing. The few studies that do exist, 1) are conducted in a context different than video games, thus not relevant, and more importantly, 2) they present conflicting results about the performance of the methods. It is no coincidence that usability research on video games is criticised by gaming professionals as having poor impact on design, both because of the inability to communicate research results to developers (Hopson, 2008) and because it helps identify problems but fails to propose solutions (Hopson, 2006; Wixon, 2003).

The aim of this study is to shed light on the strengths and weaknesses of two well-practiced usability testing methods, think-aloud protocol (during gameplay) and interviews (post-gameplay). The two methods are primarily assessed in terms of effectiveness (number of problems found). However, to ensure a strong impact of this study on the practice of video games usability testing, focus is also given on the industry-oriented measure of cost effectiveness. Two more criteria for the comparison of the methods are introduced, as an added value for this study: the severity of problems that the methods can find and the effect they may have on the gameplay experience.

This goal of the study can be summarized in one research question: *Is it worth interfering with the gameplay by asking the player to think aloud when conducting usability testing on video games?*

[1] Andree-Anne Boisvert is a usability specialist at Ubisoft Canada, a major video gaming company. This quote is from her talk at GDC Canada 2010 (Carless, 2010).

1.1 Document structure

Chapter 2 reviews relevant literature on usability evaluation, contextualized in the domain of video games where appropriate, with focus on two main themes: the status of user testing in usability evaluation and the comparative studies of empirical evaluation methods. A series of criteria that should be used when such studies are conducted is also presented.

Chapter 3 presents the methodology of the experiment that was conducted to compare think-aloud protocol and interviews. The four hypotheses that were made are presented first and the experimental design follows.

Chapter 4 outlines the data analysis process and presents the results of this experiment. Statistical analysis and testing of the significance of the results is presented where appropriate.

Chapter 5 discusses the findings of the experiment with focus on answering the research question, criticizes any limitations of the study and signals further research.

Chapter 6 summarizes the document and draws conclusions.

2 Literature Review

This chapter will provide an insight into the literature on usability research. The concepts of usability and usability evaluation are introduced and contextualized in the domain of video games. A short history of comparative usability studies is then presented, to highlight what the measures for usability are and should be and also position user testing in the vast field of usability evaluation. The two methods that are investigated in this project are then presented and comparative studies of their variations are discussed. The motivations for further research are explained and the set of criteria that this research should be based on are also discussed.

2.1 On usability

2.1.1 Usability and usability evaluation

According to the ISO 9241-11 (1998) definition, *usability* is an attribute that describes the effective, efficient and satisfactory use of a system. The latter is often underestimated in usability studies for the sake of performance, mainly because of the original development of usability evaluation methods for productivity software, which aims, well, productivity. To a certain degree this is to be expected, as the goal there is to enable faster, easier and less error-prone performance (Pagulayan, Keeker, Wixon, Romero, & Fuller, 2007).

Human-Computer Interaction research has contributed a wealth of knowledge on how to measure the usability of interactive systems and a great number of usability evaluation methods (UEMs) have been developed. A basic categorization of UEMs is that of analytical and empirical methods.

Analytical (or inspection) methods involve the evaluation of an interactive system by usability experts, using various techniques. They include Heuristic Evaluation (Nielsen & Molich, 1990), Cognitive Walkthrough (Polson, Lewis, Rieman, & Wharton, 1992), GOMS (Card, Moran, & Newell, 1983), and others.

Empirical methods involve the evaluation of an interactive system by real (or representative) users. They include a variety of techniques that allow the user to report the problems of the system, such as think-aloud protocol (Lewis, 1982), verbal protocol analysis techniques (Varela-Alvarez, 1995) adopted by psychology (lie interviews), the RITE method (Medlock,

Wixon, Terrano, Romero, & Fulton, 2002), guided interaction, questionnaires and others. These methods are commonly referred to as usability testing or user testing.

2.1.2 Usability vs. user experience vs. playability

It is critical to understand the differences of games and productivity software before attempting to measure usability for games. Pagulayan et al. (2007) explained the fundamental differences of video games and productivity software with the illustrative example of difficulty: productivity software has to be easy to use, while a game has to offer a certain level of challenge or the player will get bored. Lazzaro & Keeker (2004) presented a comprehensive list of the differences between the two types of software.

Taking a step beyond traditional usability and the performance goals of effectiveness and efficiency, Sharp, Rogers, & Preece (2007) suggested a wider set of *user experience* goals for the design of interactive systems, such as enjoyment, fun and emotional fulfilment. These terms have been discussed extensively in affective interaction literature. Hassenzahl, Platz, Burmester, & Lehner (2000) adopted the same perspective on usability that extends towards user experience and embrace the hedonic quality of a system as equally important and equally affecting the user's appeal as its ergonomic quality. Picard (1999) also stressed the importance of affect, as it can make up for the user's frustration caused by the lack of traditional usability, an opinion also shared by Norman (2004) who suggests that aesthetics and the emotion it evokes can enhance usability. Lazzaro (2008) aptly commented that HCI used to cover up usability problems with nice graphics, but today the danger lies in the misbelief that user experience may be improved just by fixing all usability problems, indicating that this step towards user experience is necessary.

In video games traditional usability and user experience converge to the concept of *playability*. A proper definition of playability does not exist and its use is not the same across studies. Some researchers imply both the interface usability and gameplay experience (Desurvire, Caplan, & Toth, 2004; Korhonen, 2010), while others separate playability from usability (Federoff, 2002; Järvinen, Heliö, & Mäyrä, 2002). HCI research has gone even further, identifying specific elements of a playable experience, like immersion (Brown & Cairns, 2004), flow (Chen, 2007) and presence (Jennett, Cox, & Cairns, 2008).

In this study the term *usability* will be used with its unified concept as an attribute of a problem-free, playable and enjoyable game.

2.1.3 Measuring usability in video games

Calvillo Gámez, Cairns, and Blandford (2008) attempted to identify the factors of game experience that were measured by researchers so far and what factors have to be measured. They acknowledged that user experience in games goes beyond usability and cognitive issues, and highlighted the importance of enjoyment. The limitations in literature and lack of an integrated measure of this experience are also noted. This lack of common language drives individual researchers to develop their own methods – custom heuristics, questionnaires, walkthroughs, etc – to evaluate the various aspects of this experience.

In video games, both inspection methods and user testing are widely used. Many sets of heuristics have been developed to measure (quantitatively) playability (Desurvire, Caplan, & Toth, 2004; Federoff, 2002; Pinelle, Wong, & Stach, 2008) and scales to measure (quantitatively) the overall gaming experience (Parnell, 2009) or specific elements of this experience, like immersion (Jennett et al., 2008) and engagement (Brockmyer et al., 2009).

User testing is inherently different than other methods, as it gives to the user the freedom to talk about anything that has a negative effect on the gaming experience, being an interface problem, a bad story or simply a boring game that impacts user experience. This renders user testing ideal for both finding interface problems and qualitatively evaluating user experience.

2.1.4 Ask the expert or ask the user?

In the long history of usability evaluation in HCI literature, both the analytical and the empirical paradigms have been extensively researched and compared with each other. The results are often contradictory. The discussion that is going on about the two paradigms is presented briefly, in usability in general and in usability for video games afterwards.

In two of the first comparative usability studies in HCI (Bailey, Allan, & Raiello, 1992; Jeffries, Miller, Wharton, & Uyeda, 1991) different analytical methods were compared with each other and with user testing. The results varied, with Jeffries et al. being in favour of heuristic evaluation as an efficient and low-cost method, and the Bailey et al. being in favour of usability testing as ideal for finding top-priority problems. These early studies, along with others in the following years, lead to the development of a consensus that experts report more low-severity and more false problems than users (Gray & Salzman, 1998). Thus, user testing became a gold standard, a benchmark for inspection methods.

Gray and Salzman (1998) heavily criticized the way past comparative studies were conducted. They identified flaws in experimental design that compromised their validity and prevented any generalization of their findings. Despite all the criticism, they also advocated the established status of user testing as the yardstick of evaluation. In a follow-up publication intended to comment on Gray and Salzman's work, Newman (1998) attempted to dethrone user testing, indicating that, as any other UEM, it is not flawless, with problems mostly emanating from the wrong application of the method. He used two case studies to depict how user testing, when it is badly designed and performed, can lead to catastrophic results. Later on, Blandford, Hyde, Green, and Connell (2008) again noted that analytical methods are complementary to user testing, which covers a broader range of possible issues and reveals issues that are not within the scope of analytical methods.

In video game testing, Desurvire et al. (2004) compared a set of heuristics for playability with think-aloud protocol. The results show that, as in productivity software, heuristics find more problems, but user testing still identified problems that can only be found by observing the user play. They concluded that user testing is still the benchmark for any evaluation of usability in video games.

Late developments in playability heuristics for mobile games by Korhonen (2010) showed that heuristics can detect playability issues equally well with user testing. Moreover, many arguments are presented as to why there are several advantages in using experts instead of users. Even then, the conclusion of this study is that analytical methods should be considered as complementary. This finding is consistent with past work in the domain of productivity software.

2.2 Let the users do the talking

User testing is evidently crucial for a thorough evaluation of an interactive system. *Think-Aloud Protocol* and *Interviews* are two well-practiced methods. In this section a short introduction to these methods is presented, comparative studies of variations of these methods are discussed and appropriate criteria for their assessment are suggested.

2.2.1 Think-aloud protocol and interviews

Think-aloud protocol was introduced in usability evaluation by Lewis (1982), being adopted by experimental psychology. In its typical form, concurrent think-aloud, as its name implies, it requires the user to think aloud while performing tasks on the system under assessment. It seems to be very popular in HCI practice, especially for laboratory-based usability testing.

Nielsen (2002) did a very thorough review of the method highlighting some of its advantages mentioned in past studies, such as its simplicity and the ability to detect cognitive activities of the user that may be invisible. They also note one disadvantage, that its results are dependent on the proper recording and analysis of the think-aloud session. Nørgaard and Hornbæk (2006) did a research on how the method is applied in practice and found that very often evaluators seek to confirm issue they are aware of, rather than finding new ones. Even if this is the case, it does not devalue the use of think-aloud protocol, especially in video game testing where it is a typical procedure (Pagulayan et al., 2007).

Interviews are more than a mere conversation; "a conversation with a purpose" as nicely described in Kahn and Cannell (1957). They usually involve the user performing task on the system under assessment silently and then verbalize retrospectively after the testing session, while the researcher is probing for feedback. Sharp, Rogers and Preece (2007, pp. 389-427) presented a detailed guide on how to conduct interviews of any form. It is noted that the goals of the interview have to be pre-determined, the interviewer must be consistent in the way the questions are asked and that it is typical to combine interviews with questionnaires. Semi-structured interviews are of particular interest in video game testing, as they allow a more exploratory approach which seems more appropriate when there are no specific tasks and the player is asked to play the game freely.

2.2.2 Retrospective vs. concurrent

Concurrent verbal protocols (like think-aloud protocol) and retrospective verbal protocols (like interviews) have been well-practiced in usability testing. However, despite the great number of comparative usability studies in HCI literature, the intrinsic comparison of empirical methods, especially verbal protocols, is scarce and with controversial results. In this section, four such comparative studies are presented.

Ohnemus and Biers (1993) compared three variations of think-aloud protocol: concurrent (users talking while using the system), immediate retrospective (conducted right after the testing session) and delayed retrospective (conducted the day after). In the retrospective conditions, the users were watching a video of their testing session while talking. The test system was database management package. Their results showed no significant difference among the three conditions in terms of performance, but in the retrospective conditions the users' comments were of higher value for the designers.

Henderson (1995) compared four user testing methods including two retrospective verbal protocols: semi-structured interviews and think-aloud protocol (while users were watching their video-recorded testing session). The test system was an office applications suite. The comparison did not include concurrent think-aloud protocol, but the study is of particular interest because of their results. Think-aloud protocol was found as more robust and efficient than interviews and the rest methods (questionnaires and logged data), even when conducted retrospectively. In all cases it found significantly more problems. The study also highlights that the methods are work-intensive and suggests low level coding as a counter-measure to this deficit.

Donker and Markopoulos (2002) compared concurrent think-aloud protocol, interviews and questionnaires on an educational game for children (8-14 years old). Their results showed that think-aloud protocols found overall more problems, but with no significant difference. They also reported that girls found more problems than boys, raising the question whether this would be the case with adult participants as well. In a follow-up study (Baauw & Markopoulos, 2004) a website and an educational game were tested. In both cases the previous results were confirmed. The study was extended by also counting the problems identified through observation (in the case of think-aloud protocol), and not only by means of the user's verbalization, following the natural inclination of the practitioner to use any data available. Again, the difference in problem detection was not significant.

Van den Haak and de Jong (2003) compared retrospective and concurrent think-aloud protocols on an online library catalogue. They didn't find any significant difference between the numbers of problems the two methods, even though concurrent think-aloud reported more problems. However, they raised two concerns for concurrent think-aloud protocols. They argue that more problems were found through observation rather than the actual verbalization of the user, and that thinking aloud has a strong negative impact on task performance, with the users who tested the system silently having better task success rates.

2.2.3 Critique on past studies

The aforementioned studies share some common problems. The results were conflicting: two studies found that think-aloud report more problems (Donker & Markopoulos, 2002; Henderson, 1995), but only one with significant difference. The rest found no difference at all. Also, the number of problems seems to be the only common measure of performance for the evaluation. Only one study reported that think-aloud protocol is more work-intensive (Henderson, 1995), thus cost criteria are essential, but the difference in resources required

to apply the methods was not analyzed. One other study reports that retrospective verbalizations lead to more useful comments (Ohnemus & Biers, 1993), thus better downstream utility, but again this has to be confirmed in video games. Also, only one study reports that thinking aloud while using the system has a negative effect on user's performance (van den Haak & de Jong, 2003). This may indeed be important in video games, where negative impact on the player's enjoyment because of thinking aloud may result in the game being rated by the gamer as worse than it really is, leading to false conclusions about the game's appeal.

To the extent it could be researched until the beginning of this project, comparative studies of concurrent and retrospective verbal protocols with different or more concrete results do not exist in HCI literature. More importantly, no such studies seem to exist in the domain of video games, other than educational games for children, much different than games that aim the player's enjoyment. Someone would expect otherwise for two so widely used methods. The reason for this is perhaps that both methods are so popular that their use in game testing is not considered worth discussing.

2.3 Criteria for comparing empirical methods in video games

Comparative usability studies stress the importance of predefining the comparison criteria. In this section, a set of criteria that have been collected from literature is presented. Most of the criteria were originally used to compare analytical methods, but here they are adjusted to the needs of the comparison of user testing methods for video games.

2.3.1 Performance criteria

As in every comparative usability study, what is of outmost importance is the performance of the methods in problem detection. An extensive review of performance criteria for UEMs was done by Hartson, Andre and Williges (2001). The three criteria they suggest are thoroughness, validity and reliability.

Thoroughness is defined as the proportion of real problems identified by a method to the real problems that exist in the system under assessment.

$$Thoroughness = \frac{number\ of\ real\ problems\ identified}{number\ of\ real\ problems}$$

Validity is defined as the proportion of the problems identified by the method that are real problems and not false positives.

$$Validity = \frac{number\ of\ real\ problems\ identified}{numbers\ of\ problems\ identified}$$

When comparing analytical and empirical methods, the problems that the users report are considered as real and the rest (the ones analytical methods find, but empirical don't) are considered as false positives. While this has received criticism (Gray & Salzman, 1998), no alternative has been suggested as for how to measure the thoroughness or validity of a method.

When only empirical methods are applied, all the reported problems are de facto considered as real and no false positives exist. Consequently, the validity of an empirical method is always 1 and so, thoroughness and validity are reduced to one performance measure. As Hartson et al. (2001) use the term *effectiveness* for the product of thoroughness and validity, it may be used in comparative usability studies for empirical methods as the *unique* measure of performance.

$$Effectiveness = \frac{number\ of\ problems\ identified}{total\ number\ of\ problems}$$

Reliability between evaluators as a measure of consistency across evaluators, is also important. It is of course relevant only when the methods are applied by more than one evaluator, thus not used in the present study.

2.3.2 Cost effectiveness

Game researchers argue that solely with performance criteria HCI research may have poor impact on practice (Hopson, 2006; Wixon, 2003). So, one realistic and practical suggested criterion of evaluation is cost effectiveness. Hartson et al. (2001) extended the definition of effectiveness to include cost factors, suggesting the effectiveness per cost unit. Thus, the cost has to be defined. When comparing two verbal empirical methods, like think-aloud protocol and interviews, a large volume of qualitative data is gatherer. The time needed to analyze the data is critical and highly-dependent on the method used to acquire it. The average time spent to collect and analyze this data can be used as a cost factor. Thus, cost effectiveness can be defined as the number of problems identified by the researcher per hour of work of the researcher.

2.3.3 Problem severity

The severity of the reported problems is another measure used in many comparative usability studies (Barendregt, Bekker, Bouwhuis, & Baauw, 2006; Jeffries et al., 1991) and in

the domain of video games in particular (Korhonen, 2010; Laitinen, 2005). Usually, a severity rating is assigned to each reported problem by a team of evaluators and the mean severity is used as the problem's severity score. In the case that a team of evaluators is not available, the severity rating may be assigned by the user. This is nevertheless a subjective measure and may differ from user to user, so averaging the result is crucial.

A widely used severity scale is Norman's severity ratings (Nielsen, n.d.). A scale from 1 to 4 is used, where 1 is a cosmetic problem and 4 is usability catastrophe. Another custom scale was used by Barendregt et al. (2006) to measure the so-called impact severity. It is a 3 point scale, from 1 to 3, where 1 is a minor problem, 2 is intermediate and 3 is severe. The term impact severity was used to separate it from the frequency severity, the number of participants that reported the problem, which is another measure that may approximate the actual severity of the problem (Pagulayan et al., 2007).

2.3.4 The effect on appeal

As discussed before, one study commented on the reduced performance of the users when thinking aloud (van den Haak & de Jong, 2003). While task performance may not be of great relevance in video games, a question is raised whether think-aloud would have the same effect on the game's appeal or the player's enjoyment. If it does, one might expect that the player would rate the system with a lower score, leading to false conclusions about the game's quality. This effect, if measured, could be another criterion for the evaluation.

Questionnaires and scales are typically used to measure quantitatively the elements of user experience, such as enjoyment and overall appeal of the game. An extensive review of all the existing scales that measure gaming experience is not deemed necessary. A simple appeal scale may be used, like the Attrakdiff by Hassenzahl et al. (2000), for productivity software or its adapted version for video games by Parnell (2009). The mean score of the game's appeal can be used to compare the two methods.

2.4 Summary

In this chapter a brief but hopefully insightful review of HCI literature on usability was presented. The concepts of usability and usability evaluation were contextualized in the domain of video games. It was illustrated how empirical evaluation methods are still regarded as the cornerstone of usability, withstanding criticism, and their importance for game testing was highlighted, both in productivity software and video games.

A few selected studies were presented, that compare retrospective and concurrent verbal protocols. While such studies generally favour slightly think-aloud protocol, as opposed to interview, no final decision can be made about which of the two methods is better at problem detection. Moreover, the two methods have never been compared for video games, making a future comparative study imperative.

The criteria for comparing two empirical methods in such a study were then suggested, as adopted from relevant literature, but also adjusted for the needs of empirical methods.

Having this background, the next step is to design a comparative usability study that will shed light onto the differences of the two methods in the field of video game testing.

3 Experiment

3.1 Rationale & Hypotheses

To investigate whether it is worth interfering with the gameplay by asking the player to think-aloud when testing the usability of video games, or it is preferable to have an interview afterwards, a controlled experiment was conducted.

In this chapter, the methodology followed in this experiment is presented. Participants evaluated the usability of two games using one of the two methods; think-aloud protocol or interview. Think-aloud protocol required the user to talk while playing the game and report anything that had a negative impact on the gameplay experience, while interviews required the player to only have a conversation about the game after playing the game uninterrupted. Below, the expected findings of the study are outlined.

Hypothesis #1: Effectiveness

> *It is hypothesised that Think-Aloud Protocol will be more effective at finding problems than Interviews.*

When thinking aloud, the players will be reporting the problems while they are experiencing them, thus they are likely to report more problems. On the contrary, in the case of interviews the interval between the gameplay session, where problems are experienced, and the interview session after the gameplay may cause some problems to be forgotten, thus not reported. This may be true particularly for problems experienced earlier rather than later in the gameplay session. Past studies showed that think-aloud protocol can predict more problems, but with no significant difference (Donker & Markopoulos, 2002; van den Haak & de Jong, 2003). Both studies were on productivity software and not video games.

Hypothesis #2: Cost effectiveness

> *It is hypothesised that, despite its potential advantage in problem detection, Think-Aloud Protocol may be less cost effective than Interviews.*

Think-aloud sessions typically produce large volumes of audio/video data that need to be systematically analyzed. Interviews are unlikely to last so long as a typical one-hour-long gameplay session. Given that, the time required for the analysis of think-aloud data is

expected to exceed that of interviews and the cost effectiveness of think-aloud protocol (in number of problems found per hour) lower than interviews. Literature suggests that all verbal protocol analysis techniques are resource-intensive, but no comment is made about the differences between the methods (Henderson, 1995). There is no stronger support for this hypothesis and the result will depend on how effective and resource-intensive the two methods prove to be.

Hypothesis #3: Problem severity

> *It is hypothesised that the average severity of the problems reported by Think-Aloud Protocol will be lower than that of Interviews.*

The aforementioned expected disadvantage (low effectiveness) of interviews may become an advantage when it comes to problem severity. The problems that may be forgotten, thus not reported during interviews, may only be minor problems that did not have a strong impact on the player's performance. So, the player may report more higher-severity problems during interviews, increasing the average severity of the problems reported by the method. There is no support in past literature for this hypothesis.

Hypothesis #4: Game's appeal score

> *It is hypothesised that the game's overall appeal score for Think-Aloud Protocol will be lower than that of Interviews.*

Requesting from the player to talk while playing may slightly deteriorate the gameplay experience and cause the player to enjoy the game less, thus assign a lower appeal score to the game. On the contrary, in the case of interviews the player is uninterrupted during gameplay and the gameplay experience is not affected. Past literature suggests that thinking aloud while using a system has a negative impact on task performance (van den Haak & de Jong, 2003). It will be interesting to see if this is the case for the game's appeal, too.

3.2 Method

A controlled experiment was conducted to evaluate the performance of the two methods and compare them against the hypotheses. Every attempt was made that the experiment resembles the conditions of a proper video game usability testing conducted by professionals. The methodology that was followed is described in this section.

3.2.1 Ethical considerations

Before the beginning of the experiment, this study was granted ethical clearance by the UCL Research Ethics Committee. Due to the qualitative nature of the two methods under question, data analysis relies on the collection of audio/video recordings of the participants' interactions, including the participants' faces (for the thinking-aloud sessions). Thus, the anonymization of the recordings and their secure storage were important ethical issues.

A risk assessment was also conducted before the experiment, with only one apparent risk, the probability of an epilepsy crisis of the participant while playing games, a rather unlikely but nevertheless serious risk. Therefore, the participant's informed consent was requested and participants were asked not to participate in case of history of epilepsy. The consent form can be found in the Appendix (A.1).

3.2.2 Participants

In total, twenty gamers participated in this study (*M* = 27.7 years old, *SD* = 4.9, 25% females). The participants were recruited through an online booking form that was sent to the UCLIC students' mailing list and all of them were HCI-E students. The participants had various levels of gaming background experience[1] (*M* = 4.3, *SD* = 1.0) and different frequencies[2] of game playing (*M* = 3.1, *SD* = 1.0). Participation was voluntary and no incentive was given to the participants.

The demographic details of the participants can be found in the Appendix (A.2).

3.2.3 Materials

The games

Each method was tested with two games of different genres. This ensured that the results of the study were not game-dependent or genre-dependent, increasing the ecological validity of this research.

The first game was ***Eufloria*** (Kremers & May, 2009), an independent strategy game with a metascore of 64/100 (based on 9 critics) and a user score of 7.2/10 (based on 22 ratings). The metascore is an average ranking of the game based on several online game reviewers and the user score is an average based on users that voted the game online. Both can be

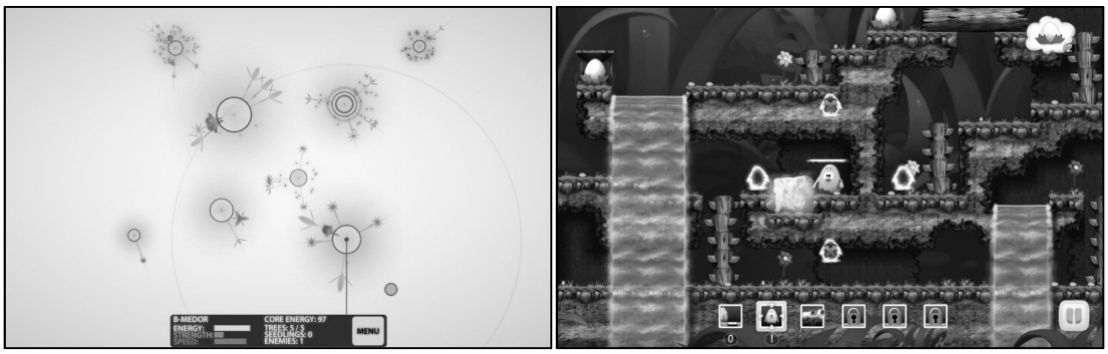
[1] Gaming background experience was measured on a Likert scale as follows:
5: 15+ years, *4*: 10-15 years, *3*: 5-10 years, *2*: 3-5 years, *1*: 0-2 years, *0*: Never
[2] Game playing frequency was measured on a Likert scale as follows:
5: Every day, *4*: A few times/week, *3*: A few times/month, *2*: A few times/year, *1*: Less often, *0*: Never

found on the website Metacritic (www.metacritic.com). The game requires the player to build and control an army of beings called “seedlings” and conquer planets by building trees on them and fighting enemy seedlings.

The second game was ***Toki Tori*** (Two Tribes, 2010), a casual puzzle game with a metascore of 80/100 (based on 10 critics) and a user score of 9.4/10 (based on 18 ratings).The game requires the player to navigate a little bird around each level, collecting all the scattered eggs. To do this, the player has to use various tools to bypass obstacles and find the unique combination of actions that allows the level to be completed.

These two games were specifically selected because they are appropriate for all ages and their metascores and genres vary enough. Also, none of the games has a very high metascore (>85), so it can be assumed that they haven’t undergone extensive usability testing and they have problems that would be uncovered in a usability study. Screenshots of the two games can be seen in Figure 1 below.

Figure 1. Screenshots of the two games, Eufloria (left) and Toki Tori (right).



Equipment

The experiments took place in a UCL lecture room. A 14.1” laptop (specifications: Core 2 Duo, 4GB RAM, nVidia 9300M GS, Microsoft Windows 7) was used as a platform to play the games. This computer was powerful enough to run seamlessly the game and the recording software at the same time. An external wireless mouse was connected, as both games required a mouse to be played and the laptop’s touchpad alone could impede the player’s performance. The computer was connected to the room’s audio equipment for better audio experience and the lights were slightly dimmed during gameplay to provide better visual experience, without hindering the recording of the player’s face while playing. The computer’s embedded microphone and web camera were used for the audio and video recording respectively. The trial version of the usability testing software Morae (TechSmith, 2010) was used for capturing gameplay video.

Guide for the researcher

A list of points that had to be covered by the participant – while thinking aloud or while conversing with the researcher during the interviews – was created. It included general points (e.g. graphics, sound effects, music, controls) and game-specific points (e.g. the toolbar icons in Toki Tori and the planet attributes in Eufloria). This list was compiled based on the author's collective experience as a video game player and hobbyist reviewer, the host's recommendations on which aspects of the game are typically evaluated and a great number of game reviews in respectable online gaming magazines. No reviews of the two games used in this study were read, to avoid biasing the researcher.

Pre-gameplay survey

A pre-gameplay survey was created to get basic demographic information and additional information about the participants' gaming habits (background experience, game playing frequency, favourite games & genres), and to ensure that the participants had never played the games before. The survey can be found in the Appendix (A.3).

Post-gameplay questionnaire

A post-gameplay questionnaire was created for the participants to evaluate their overall experience (see Appendix, A.4). It included Likert scale questions about the *overall rating* of the game (rated from 0 to 10), the participant's *enjoyment* (rated from 1 to 5) and questions that rated certain aspects of the game (graphics, music, controls and difficulty).

A modified version of the Appeal subscale of Hassenzahl et al's (2000) Attrakdiff was also included. This scale was modified to better fit video games (Parnell, 2009). The reason behind the decision to have both the two Likert scale questions (rating, enjoyment) and a separate appeal scale was the questionable appropriateness of that specific scale and the lack of other validated suitable scales. On the other hand, two questions alone ("How would you rate the game overall?" & "To what degree did you enjoy the game?") might not be enough to evaluate the gameplay experience. Thus, the appeal scale was included nevertheless.

The questionnaire also included open-ended questions that required the player to describe the top 3 problems they experienced and assign to them a severity rating according to the guide in Table 1 below. This was a revised and enhanced version of Nielsen's (n.d.) 4-point list used by professionals in video game testing (Laitinen, 2005). A fifth point was added, as it is typically advised that Likert scales have odd number of values, and a description was included to help the participant choose the appropriate rating.

Table 1. Problem severity rating guide.

RATING	***CATEGORY***	***DESCRIPTION***
1	Aesthetic issue	barely noticeable
2	Minor issue	slightly impairs gaming experience, if any at all
3	Intermediate issue	somewhat impairs gaming experience
4	Severe issue	seriously impairs gaming experience
5	Catastrophic issue	cannot continue playing if it persists

Both the pre-gameplay survey and the post-gameplay questionnaire were influenced, although not copied at any point, by the external host's own versions of such documents that are used in video games usability testing.

3.2.4 Design

Independent variables

A between-subject experimental design was used, with five participants per condition; each participant was assigned to one game and one method. There were two factors in this study; the method used to do the user testing, with two levels (Think-Aloud Protocol and Semi-Structured Interview), and the game played by the participant, with two levels (Game 1: Eufloria, Game 2: Toki Tori). Thus, there were four experimental conditions, summarized in Table 2 below.

Table 2. The four conditions of the experiment.

CONDITIONS	**GAME 1** *Eufloria*	**GAME 2** *Toki Tori*
METHOD 1 *Think-Aloud Protocol*	G1M1	G2M1
METHOD 2 *Semi-Structured Interview*	G1M2	G2M2

The effect of any confounding variables (age, gender, gaming background experience, game playing frequency) had to be minimized to insure the internal validity of the study. Such differences were equally scattered across the two conditions of the same game. Another confound was the time of the day of the gameplay session, as the accumulated fatigue could affect the participant. All four conditions were evenly allocated to different time slots in the ten days of the experiment. The allocation of the participants based on their individual characteristics can be found in the Appendix (A.5).

Dependent variables

The directly measured dependent variables were:

- The time that the gameplay session lasted, in minutes

- The overall rating of the game, on a Likert scale from 0 to 10
- The appeal score of the game, using the appeal scale discussed before
- The top 3 problems that the participant experienced, along with their severity ratings on a Likert scale from 1 to 5

The most important dependent variable was the number of problems (or percentage of problems) identified by each participant, indicating the effectiveness of the method. However, this could not be measured directly, and instead had to be calculated as a product of the data analysis process, as would the cost effectiveness of the methods, measured in number of problems found per hour.

3.2.5 Procedure

Pilot

A pilot testing preceded the experiment. Two friends of the author, both male around 27 years old, participated in a crossover pilot study, where each of them played both games with one method per game (two conditions per participant). The gameplay sessions lasted only 30 minutes per game. The session helped the author get a better idea of how to conduct the final experiment and refine many aspects of the study's design. The survey, the questionnaires, the interview questions and the equipment were all improved or recalibrated based on the participants' recommendations and the author's observations.

A major decision that had to be made was the duration of the gameplay session. It was clear after the pilot that 30 minutes are not enough for the participant to have enough progress in the game and evaluate all its aspects. Following the external host's advice and the relevant literature that suggests a minimum of one hour of gameplay (Pagulayan et al, 2007), the gameplay duration was doubled to one hour.

Experiment

When the participant came in the testing room on their allocated day and time, he/she was welcomed and asked to read an information sheet with details about the experiment. The same details were also explained orally by the researcher. Then the participant was asked to sign a consent form and complete the pre-gameplay survey while the researcher was setting up the recording software.

To counterbalance any social desirability bias, it was emphasized to the participant that it is the game that is being tested, and not his/her performance in the game, and that their

objective was to report anything that impedes the gameplay experience. This trick has being reported as successful in a past study (Donker & Markopoulos, 2002). No predefined task was given to the participants. No specific task was given to the participants. Instead, they were asked to play the game as they would play it at home. Such open-ended tasks are common in games' usability testing (Korhonen, 2010; Pagulayan et al., 2007).

In the case of think-aloud protocol, the participants were asked to think-aloud while playing, reporting anything that affects their experience. They were also told to feel free to comment on things that they like about the game, thus encouraging them to talk more. The researcher was sitting nearby, observing the participant and taking notes. If the participant was not talkative enough, the researcher was asking short questions, trying to be as unbiased as possible (e.g. by asking "How do you find the controls?" instead of "Do you find the controls hard?"). The gameplay session was followed by the completion of the post-gameplay questionnaire.

In the case of interviews the participant was just asked to play the game, keep in mind that the purpose of the session is to find the game's problems. The researcher was sitting at the back of the room, but no observations were made or notes taken, because the data gathering for this method had to be based only on the post-gameplay interviews and not the participant's gameplay session. The gameplay session was followed by the completion of the post-gameplay questionnaire. Then the Interview session started, always with the same question ("So, how was it?"), as suggested by the host, and proved to be a nice entry point for the conversation. A semi-structured format was adopted. The participant was encouraged to talk about the gameplay experience, while small questions were asked until all the points from the pre-made list were covered. Only then, the researcher took a short look at the participant's post-gameplay questionnaire and asked for clarification of the participant's answers when necessary. The duration of the interview was not pre-determined, as it depended on how talkative the participant was, and varied a lot (M = 15.45 minutes, SD = 5.08).

The intended duration of the gameplay session was 60 minutes. By the end of the 60 minutes some users requested to finish the level they were playing; they were allowed to do so. Others who finished a level just a bit before the end of the 60 minutes were asked to stop playing without proceeding to the next level. Thus, there was a small variation in the gameplay duration, with an average of 60.75 minutes (SD = 3.2).

The participant was then thanked for participating and escorted out of the room. The procedure that was followed in both cases is summarized in Figure 2 (Think-Aloud Protocol) and Figure 3 (Interview).

Figure 2. Graphical representation of the process followed for Think-Aloud Protocol.

Figure 3. Graphical representation of the process followed for Semi-Structured Interview.

In the next chapter the data analysis process and the results of this experiment will be presented.

4 Data Analysis & Results

In this chapter, the analysis process of the data acquired during the experiments is outlined and the results of this analysis are presented.

4.1 Data Analysis

The experiment presented in the previous chapter produced a large volume of qualitative and quantitative data, gathered from three sources: questionnaire answers, think-aloud video recordings and interview audio recordings.

4.1.1 Survey & Questionnaires

The questionnaires were transcribed and the participants' answers were quantified where appropriate to facilitate the upcoming statistical analysis. The answers to the qualitative open questions were only included in the questionnaire for completeness purposes and were not analyzed (they were irrelevant to the purpose of this study), with the exception of the question that asked the player to name the three most important issues and assign to them a severity rating. A summary of the participants' answers to the survey and the questionnaire can be found in the Appendix (A.6).

4.1.2 Audio/Video Recordings

The qualitative data gathered by the two methods (audio & video recordings) were transcribed and methodically coded using the *open coding* principles of Grounded Theory (Strauss & Corbin, 1998) to identify problems (codes), problem concepts and problem categories. This process was time and effort-intensive. The researcher had to go back and forth in the list of problems for each participant to identify common codes and group them. The participant's ID and the method the participant used were masked during the coding process, to avoid the researcher's bias towards one of the two methods.

Duplicate problems were eliminated to ensure the validity of the analysis. For example, participants 05 and 06 reported for Game 1 that *"Graphics are too abstract"* and participants P02, P17 and P19 reported that *"Graphics are too simple".* These two problems were grouped as one problem code. Also, false negatives were carefully eliminated. Issues that were reported and could easily be identified as non-problems were excluded from the problems list. Considering the inexperience of the researcher, only the most unambiguous

cases were dealt with, to preserve the integrity of the evaluation. For example, participant 18 reported “It’s not clear what the locks in the toolbar mean” at the very beginning of level 1 of Game 2. Of course it was so early in the game that nothing had been explained. One level later the participant commented about the same issue “Ah, now I get it”. This problem was excluded.

While analyzing think-aloud video data, it was clear in many occasions that the participant was experiencing problems without reporting them. Such problems were not counted, unlike what would happen in practice. This was indeed a strict application of the method, to ensure the objective comparison of the two methods.

No grounded theory was created after the coding process and no further coding was deemed necessary, as literature also suggests for such analyses (Diaper, 1989). After all, the goal of this study was to identify the games’ problems and not develop an understanding about what impedes the gamers’ performance, although some assumptions could be made.

The problem coding tables can be found in the Appendix (A.7).The results and statistical analyses follow.

4.2 Results

4.2.1 Problem detection

Definition: Anything that the participant explicitly reported as having a negative impact on the gameplay experience was regarded as a *problem*. This, of course, does not mean that all the reported problems were real and that the participant was not exaggerating about any minor detail perceived as a defect of the game. This has been criticized in literature (Gray & Salzman, 1998; Newman, 1998). However, a more reliable measure of what problems are real cannot be established. So, this definition is adopted throughout the study.

Number of problems - Effectiveness

The superiority of Think-Aloud Protocol at problem detection was evident. For Game 1, 55 problems were reported overall. The participants that used think-aloud protocol found a total of 54 of these problems (98.2% effectiveness), while participants during interviews only reported 32 (58.2% effectiveness). For Game 2, 57 problems were reported overall. Think-aloud participants found a total of 47 of the game’s problems (82.5% effectiveness), while interview participants found a total of 31 problems (54.4% effectiveness).

The absolute number of detected problems gives a clear idea about what to expect from the problem-detecting abilities of the methods, but it is not adequate. The mean ratio of reported problems, as a percentage out of the total number of the game's problems, was also calculated for every participant and then compared across the two methods for each game independently.

For Game 1, it was found that the participants reported a more problems using think-aloud protocol (M = 35.6 % of the game's problems, SD = 3.3) than using interviews (M = 21.8 % of the game's problems, SD = 7.7). Comparison of the two methods with Student's t-test found this difference in problem report rate to be significant ($t(5) = 3.68$, $p < .05$).

For Game 2, it was found that the participants reported more problems using think-aloud protocol (M = 30.9 % of the game's problems, SD = 6.6) than using interviews (M = 16.8 % of the game's problems, SD = 7.1). Comparison of the two methods with Student's t-test found this difference in problem report rate to be significant ($t(8) = 3.23$, $p < .05$).

Both the total number of problems and the mean per participant can be seen in Table 3 below.

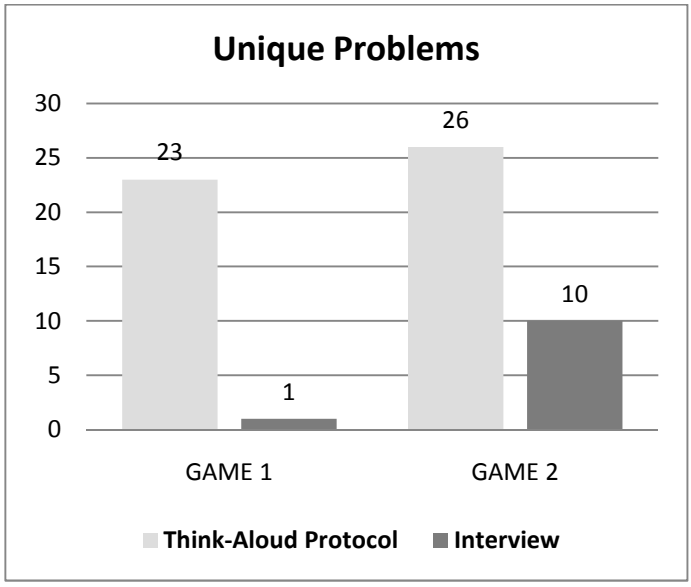
Table 3. Total number (effectiveness) & mean per participant (effectiveness per participant) of problems detected with each method. All numbers are percentages out of 55 for Game 1 and out of 57 for Game 2, which are the total numbers of problems identified in the two games respectively. Standard deviation in parenthesis.

	GAME 1		**GAME 2**	
	Think-Aloud	*Interview*	*Think-Aloud*	*Interview*
Total problems found by all participants	98.2%	58.2%	82.5%	54.4%
Mean problems found by each participant	35.6% (3.3)	21.8% (7.7)	30.9% (6.6)	16.8% (7.1)

Number of unique problems

An alternative measure of effectiveness is the number of unique problems reported by each method, i.e. the problems reported by one method and not by the other. In this case, as seen in Figure 4 below, Think-Aloud Protocol was again superior, finding significantly more unique problems than Interviews. For Game 1, Think-Aloud Protocol found 23 unique problems (42%) and Interviews only 1 unique problem (less than 2%). For Game 2, Think-Aloud Protocol found 26 unique problems (46%) and Interviews 10 unique problems (18%).

Figure 4. Total number of unique problems found by each method.



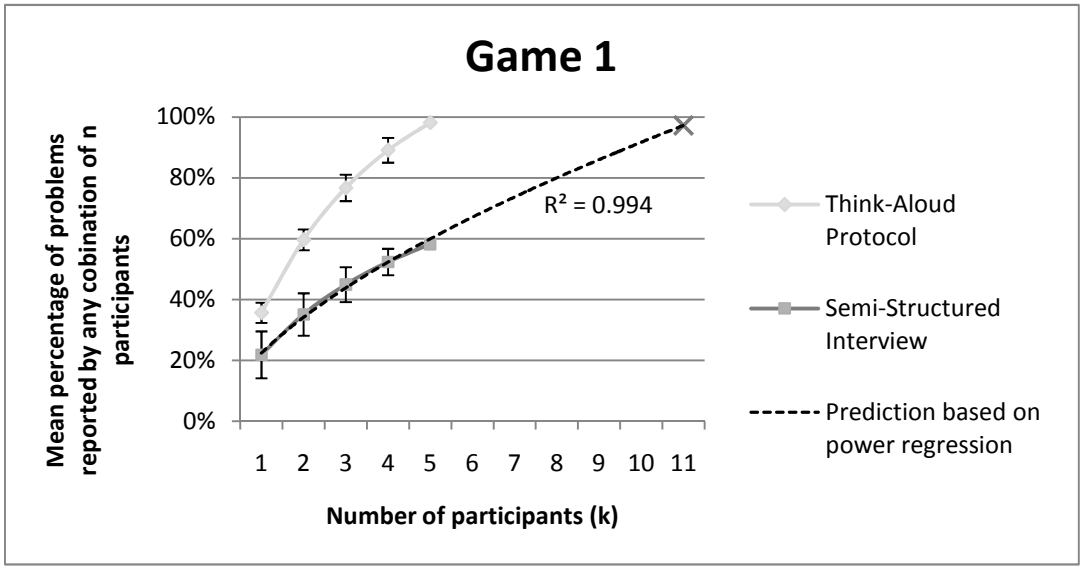
Estimated effectiveness of Interviews

Since Think-Aloud Protocol is so much better at detecting problems, the question was raised: *How many participants are needed for Interviews to have the same effectiveness as Think-Aloud Protocol*?

To answer this question, first all the possible combinations $\binom{5}{k}$ of the 5 participants in every condition were identified and the mean number of problems (as percentage of the total number of problems) was calculated for every combination. Then, these means were plotted for every game, so that regression analysis could predict, as accurately as possible, the percentage of problems that could be detected with more than five participants.

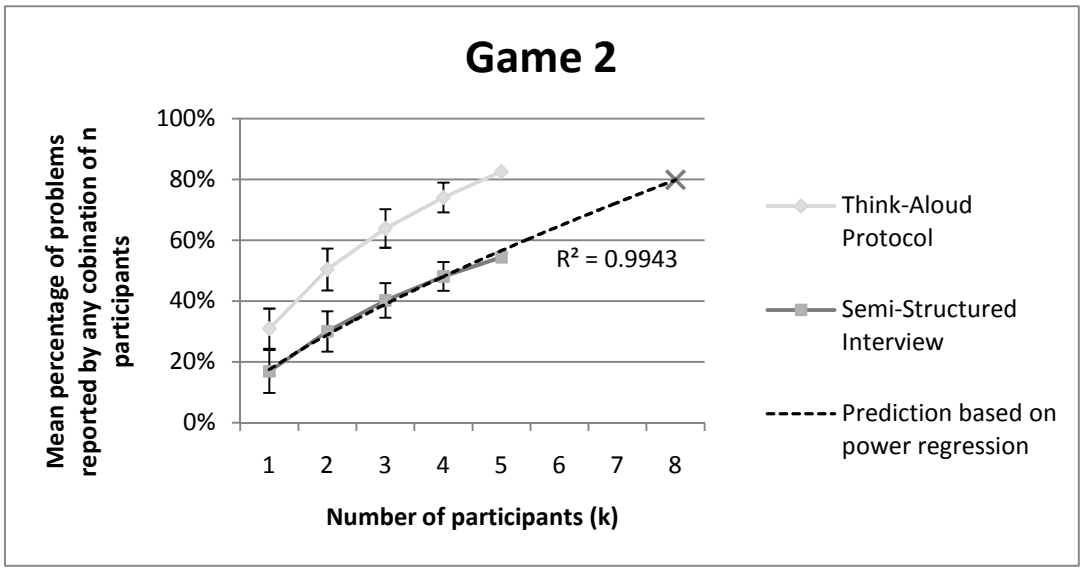
For Game 1, as seen in Figure 5 below, think-aloud participants found 98% of the game's problems. For Interviews to meet the 98% effectiveness, 11 participants are suggested, when a power regression model with excellent fit (R^2 = 0.994, $F(1,3) = 532.3$, $p < .001$) is used for prediction.

Figure 5. Prediction for the number of Interview participants needed to cross the 80% effectiveness boundary for Game 1. Each point represents the mean effectiveness (percentage of game's problems) reported by each combination $\binom{5}{k}$ of k participants (k=1,2,3,4,5). The "X" marks the predicted value.



For Game 2, as seen in Figure 6 below, think-aloud participants found approximately 80% of the game's problems. For Interviews to meet the 80% effectiveness, 8 participants are suggested, when a power regression model with excellent fit (R^2 = 0.993, $F(1,3)$ = 413.3, p < .001) is used for prediction.

Figure 6. Prediction for the number of Interview participants needed to cross the 80% effectiveness boundary for Game 2. Each point represents the mean effectiveness (percentage of game's problems) reported by each combination $\binom{5}{k}$ of k participants (k=1,2,3,4,5). The "X" marks the predicted value.



4.2.2 Cost effectiveness

The second part of the analysis investigates the cost effectiveness of the methods. The number of problems found per cost unit would be used as a measure of cost effectiveness. The cost was calculated in hours (of work), and included:

i. The duration each usability testing session lasted. For Think-Aloud Protocol the mean duration of the session was 76.5 minutes (SD = 4.6 minutes). Interviews had the additional overhead of the actual interview session, and the mean duration was 90.4 minutes (SD = 5.1 minutes).

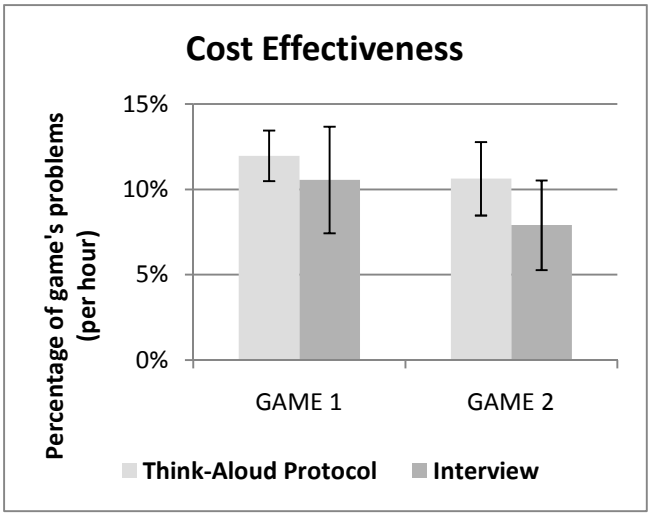
ii. The time spent by the researcher analyzing (coding and transcribing) the data. For Think-Aloud Protocol the mean duration of the analysis was 100.9 minutes (SD = 17 minutes) and for Interviews it was only 33.4 minutes (*SD* = 8.3 minutes).

Based on the above times, the cost effectiveness of the methods for each game was calculated. The complete time-problem analysis table can be found in the Appendix (A.8).

For Game 1, Think-Aloud Protocol was more cost effective, finding more problems per hour (*M* = 11.97 % of the game's problems per hour, *SD* = 1.48 %), in comparison with Interviews (*M* = 10.56 % of the game's problems per hour, *SD* = 3.12 %). However, a comparison with Student's t-test found that this difference was not significant ($t(5.7) = 0.91$, $p = .4$.

For Game 2, Think-Aloud Protocol was more cost effective, finding more problems per hour (*M* = 10.63 % of the game's problems per hour, *SD* = 2.15 %), in comparison with Interviews (*M* = 7.90 % of the game's problems per hour, *SD* = 2.63 %). However, a comparison with Student's t-test found that this difference was not significant ($t(7.7) = 1.80$, $p = .11$.

Figure 7. Cost effectiveness in percentage of the game's problems found per hour of work



4.2.3 Problem severity

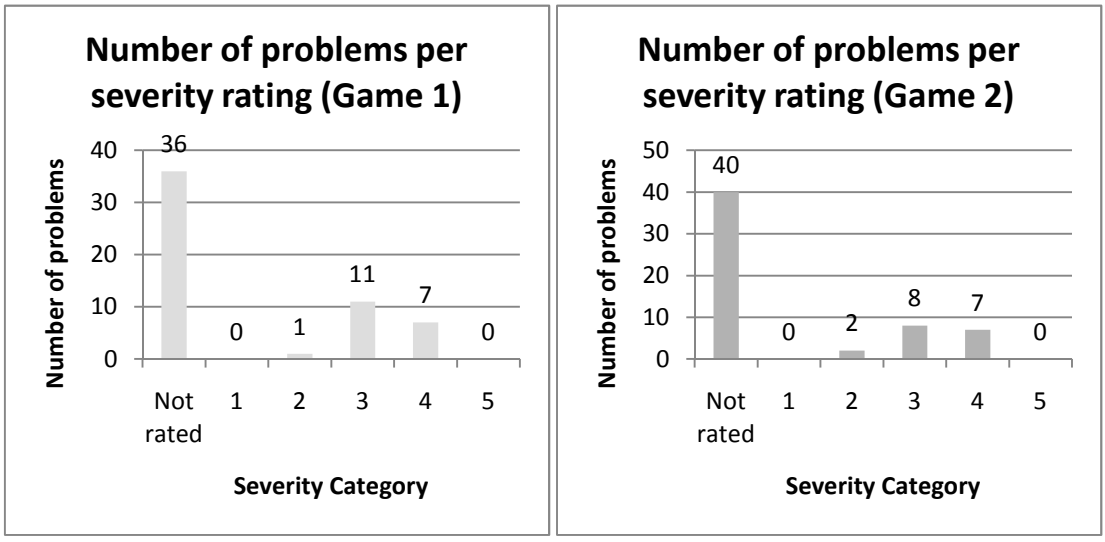
During the analysis of the results to investigate the severity of the problems detected by each method, there was an issue: the relatively big number of problems reported for each game (55 for Game 1 and 57 for Game 2, as mentioned before) in comparison to the small number of problems – only three – that the participant was asked to rate with a severity rating (from 1 to 5) in the post-gameplay questionnaire. As a result, there was large variation of the problems each participant chose to rate and small overlap of problems rated by more than one participant.

For Game 1, only 19 of the 55 problems were assigned a severity rating by at least one participant. However, for any comparisons to be made, participants from both conditions must have rated the problems. Of these 19 problems, only 5 were rated by participants of both methods. Thus, any statistic analysis is futile, especially when even these 5 problems were rated by only one or two participants per method.

The situation was the same for Game 2, where 17 of the 57 problems were assigned a severity rating by at least one participant. Again, of the 17 problems, only 5 were rated by participants of both methods, with only one or two participants per method.

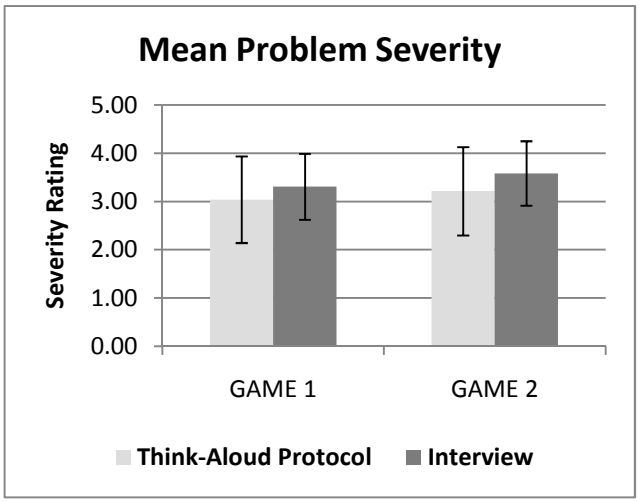
The distribution of problems per severity category for both games can be seen in the graphs below.

Figure 8. Distribution of problems in severity categories for Game 1 (left) and Game 2 (right)."Not rated" indicates the problems that were not given a severity rating by any participant.



In any case, the average severity of these five problems for each game was calculated. The results are summarized in Figure 9 below. There is apparently no difference, as the standard deviation is too large. Moreover, these averages do not involve the same problems, as different participants rated different problems. No clear conclusion can be drawn.

Figure 9. Mean problem severity (error bars indicate the standard deviation). Each bar is based only on the 2-3 problems per condition that were rated by participants of both methods.



4.2.4 The effect of talking on game appeal

To investigate the effect of thinking aloud while playing the game, as compared to playing silently in the case of Interviews, three measures were calculated: the overall rating of the game, the game's appeal score and the overall enjoyment.

Overall game rating

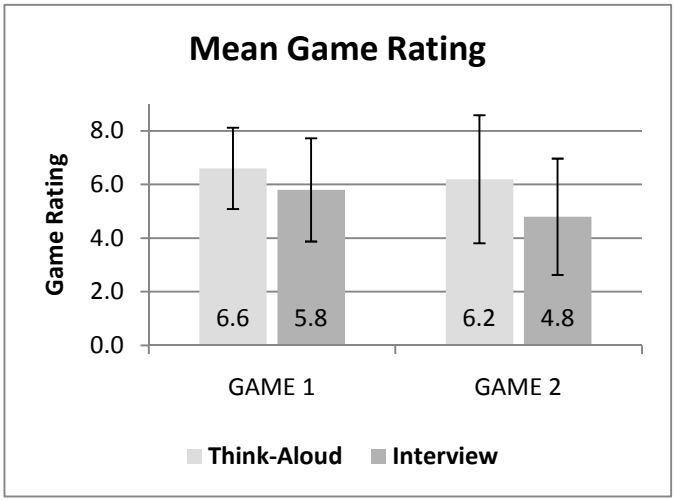
Only a small difference was found between the two methods for both games, concerning the overall game rating, which was based on the participants' responses on a 0-10 Likert scale.

For Game 1, it was found that participants gave higher overall ratings using think-aloud protocol (M = 6.6, SD = 1.52) than with interviews (M = 5.8, SD = 1.92). However, statistical analysis using a Mann-Whitney U test (suitable for ordinal data) found that this difference in rating score was non-significant, U = 15, p = .67.

For Game 2, it was found that participants gave higher overall ratings using think-aloud protocol (M = 6.2, SD = 2.39) than with interviews (M = 4.8, SD = 2.17). However, statistical analysis using a Mann-Whitney U test found that this difference in rating score was non-significant, U = 19, p = .19.

To get a clearer idea of the size of the differences, a graphical representation of the results can be seen in Figure 10 below.

Figure 10. Mean overall game rating (error bars indicate the standard deviation).



Game appeal score

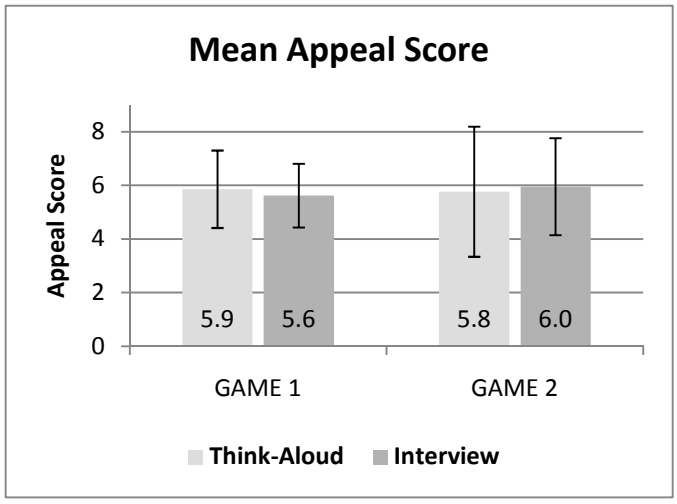
No actual difference between the two methods was found for either of the two games concerning the appeal score, which was calculated as an average of the 8 Likert point values of the Appeal Scale (see section 3.2.3). Each value was from 1 to 7, thus the overall score varied originally from 8 to 56. The scores were then normalized in the 0-10 space for more efficient comparison with the overall game rating.

For Game 1, it was found that participants gave higher overall ratings using think-aloud protocol (M = 5.9, SD = 1.44) than with interviews (M = 5.6, SD = 1.18). However, statistical analysis using a Mann-Whitney U test found that this difference in rating score was non-significant, U = 14.5, p = .75.

For Game 2, on the contrary, it was found that participants gave lower overall ratings using think-aloud protocol (M = 6.2, SD = 2.39) than with interviews (M = 4.8, SD = 2.17). However, statistical analysis using a Mann-Whitney U test not only found that this difference in rating score was non-significant, but strongly supported the null hypothesis (that there is no difference at all), U = 13, p = 1.

A graphical representation of the results can be found in Figure 11 below.

Figure 11. Mean appeal score (error bars indicate the standard deviation).

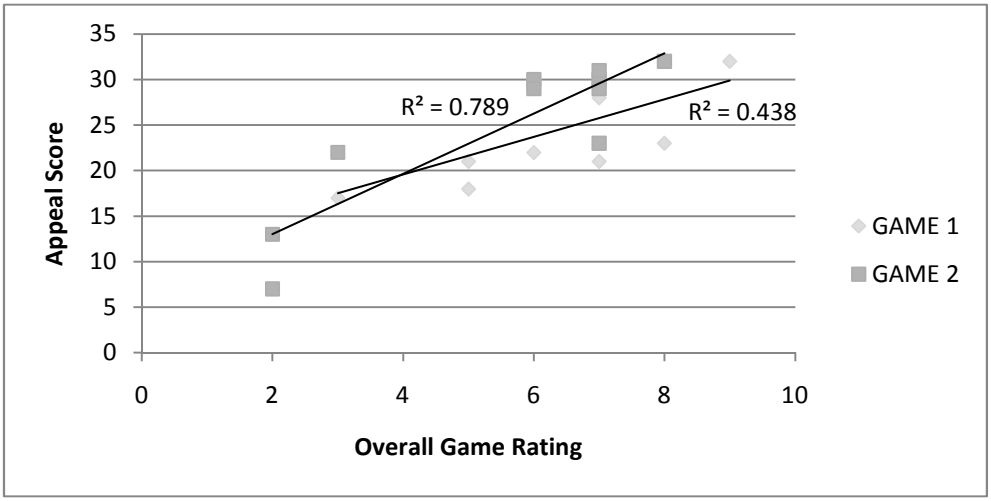


The third measure, overall enjoyment, is not included in further analysis, as it was found to be almost identical to the overall rating. A correlation test between the two showed that they are very strongly correlated (R^2 = 0.933, $t(8)$ = 10.59, p < .001).

Appeal score & overall rating correlation

To investigate the potential relationship between the appeal score and the overall rating, a correlation test was run. For Game 1, reasonable relationship was found between the two measures (R^2 = 0.439, $t(8)$ = 2.50, p < .05). For Game 2, the measures were found to have stronger relationship (R^2 = 0.789, $t(8)$ = 5.47, p < 0.001). This relationship can be seen in Figure 12 below. Apparently, the players' rating of the game correlates with their appeal of the gameplay experience.

Figure 12. Correlation between Appeal Score & Overall Game Rating.



In the next chapter, these results are discussed.

5 General Discussion

5.1 Summary of Findings

The aim of this study was to answer the research question: *Is it worth interfering with the gameplay by asking the player to think aloud when conducting usability testing on video games?* It would be wise to start this section by giving the answer to this question.

In short, yes it is worth it. The results clearly showed that think-aloud protocol can find significantly more problems than interviews. Moreover, there was a clear trend for think-aloud protocol to be more cost effective, even though no irrecusable evidence can be provided for this. These findings alone are enough to strongly support that, given the limitations of the study that will be discussed further down, think-aloud protocol is indeed a better technique to adopt for usability testing in video games.

The results of the experiment exceeded the expectation that think-aloud protocol would be more effective due to the prompt response of the participant when a problem is experienced. Not only was it more effective, but also found the most unique problems. Moreover, regression analysis showed that many more participants are needed for interviews to have the same effectiveness (6 more for Game 1 and 3 more for Game 2). The most relevant past study, in which these two methods were compared, showed that think-aloud was better at finding problems, but with no significant difference (Donker & Markopoulos, 2002).

Regarding the cost effectiveness of the methods, the results were unexpectedly in favour of think-aloud protocol. Because both methods are resource-intensive, requiring lengthy processes of transcribing and coding (Henderson, 1995), think-aloud was expected to be inferior as it produces larger volume of data. However, because of the great difference in problem detection (greater than the difference in cost), think-aloud protocol was still more cost effective.

5.2 Critique

Issues that came up during this study and limitations that may compromise the quality of this work and the generalization of the results are discussed in this section.

The "how many participants do I need?" debate

As Wixon (2003) states, the question of how many participants are needed to find an adequate number of usability problems is a classic debate in Human-Computer Interaction. Nielsen (2000) famously argues that five participants are usually enough to find about 85% of the problems, based on past work (Nielsen & Landauer, 1993). This is also supported by 3 studies presented in Dumas and Fox (2008). Sharp et al. (2007) suggest a more practical rule of 6-12 participants per testing session.

In the present study, the five-participants rule was confirmed only for think-aloud protocol in Game 2, as it found 82.5% of the problems. The rule proved to be rather pessimistic for Game 1, where 98.2% of the problems were found. However, in the case of Interviews the rule did not apply at all. As regression analysis showed, the number of participants adequate to find enough problems is within the 6-12 range.

Problem number as a measure of performance

The number of problems is one of the most widely used measures of effectiveness (Jeffries et al., 1991; Nielsen & Molich, 1990), but also the most criticised (Molich & Dumas, 2008; Sears, 1997), as it may be subject to limitations. To counterbalance any social desirability bias, the participants were told that they should report *anything* that impedes their gameplay experience. There is a concern that this approach may have been too encouraging for the participants, who tried to please the researcher by reporting as many problems as possible (Donker & Markopoulos, 2002), leading to false increased effectiveness of think-aloud protocol caused by a reverse bias. Interviews may not have been so influenced, because of the 1 hour interval between this statement of the researcher (before the gameplay session) and the interview session.

Defining "real problems"

The traditional approach of "whatever the user reports as a problem, is a problem" as it is typically done in comparative usability studies (Gray & Salzman, 1998). This raises the question of how real these problems were. A follow-up analysis of the study's results investigated the case where only problems reported by at least two participants are counted as real problems, considering that such problems can more reliably be considered as real. Again, think-aloud protocol found significantly more problems than interviews (52.7% for think-aloud and 29.1% for interview in Game 1; 40.4% for think-aloud and 22.8% for interview in Game 2).

Calculating cost effectiveness

Two limitations have to be acknowledged about the way cost effectiveness was calculated. First, the time spent on the data coding process depends heavily on the researcher's competence and experience. Thus, cost-effectiveness, a product of this process, is also subject to the same limitations. Were this experiment to be conducted by a different researcher, a kind of evaluator effect might appear (Hertzum & Jacobsen, 2001). However, in defence of the methodology followed in this study, it should be highlighted that the researcher did everything within his power to facilitate the impartial treatment of both methods.

Second, in the case of interviews, in practice it is often to have 5 or more users play a game simultaneously in a big testing room and after a certain amount of gameplay pick each user one by one for a 10 minute interview. This is much more efficient than the way this study was conducted, as the feedback of 5 users can fit in a session less than two hours long, rendering interviews very cost effective. On the other hand, analyzing every word of the participants' think-aloud recordings is death by detail. A study of several think-aloud sessions found that researchers mainly seek to confirm problems they already know and that systematic analysis of collected data is scarce (Nørgaard & Hornbæk, 2006). In that case, think-aloud may still be very cost effective (if no data analysis follows) however the goal of the test is changed from finding problems to confirming problems.

In any case, the results may not be generalizable across other studies, but they are still valid when comparing internally the two methods.

Calculating problem severity

Identifying any strong impact of either method on the severity of the reported problems was pursuit as a potential added value for this project. However, due to unforeseen limitations, no noteworthy results were produced. Perhaps it was too optimistic to ask the players to rate the problems. Severity ratings are common in usability studies for video games, but all paradigms in past studies indicate that this is done by experienced professionals, not the users themselves (Korhonen, 2010; Pinelle et al., 2008). Pagulayan et al. (2007) suggest that attitudinal measures are used to provide information about the severity of a player's specific erroneous attitude, but how this can be quantified is not presented. They do, however, propose the alternative of using the number of participants that reported a problem as an estimate of its severity. This was also suggested in Barendregt et al. (2006) and may indeed be a better measure of severity in future studies.

On the appeal and ratings

Concerning the appeal of the game, the breakthrough was not the fact that no difference was found across the methods, but three incidental lateral findings. First of all, there was strong indication that one simple question, "How would you rate the game overall?", can convey the same message as an 8-point long appeal scale. Parnell (2009) when introducing the appeal scale, in attempt to explain a series of correlations, suggested that there may indeed be a construct of "overall quality of the game" that the appeal scale can measure. Perhaps this is exactly what the *overall rating* measures, too. As an implication, there is no apparent reason to use a whole scale in a post-gameplay questionnaire, especially when the precious time of the user (and the extra space in the questionnaire) can be used to answer more game-specific questions.

A second finding questions the quality of online game reviews. Both the online metascore and online user score were better for Toki Tori (8 and 9.4 respectively) than Eufloria (7.2 and 6.4 respectively). The results of this study rated both games with lower scores than their metascores, but also gave Toki Tori a lower rating (5.5) than Eufloria (6.2).

A third finding was that, in contrast to the results of van den Haak and de Jong (2003), thinking aloud had no apparent impact on the players' performance, as the maximum level they reached in the game was not related to the method they used to verbalize.

Design limitations to be corrected in future studies

There are a few more issues worth discussing about the way this study was conducted, especially about the experimental design. It is acknowledged that the 60 minutes of gameplay is the absolute minimum in game testing (Pagulayan et al., 2007). If incentive would be given to the participants they may have been willing to participate in a longer session, providing a better view on the usability of the games. Also, ideally, representative participants would be recruited, that would fit better the profile of the target user of the games. Another limitation is the games themselves. Future studies should involve testing games of other genres, especially with 3D graphics and more complicated controls than the ones chosen here. Also, experts should be involved in future processes to assign severity ratings and add their expertise to the proper coding of the data. Nevertheless, given the resources available for the study, every measure was taken to ensure that its internal validity is not compromised.

5.3 Implications

It is strongly believed that this study can have a significant contribution to the HCI research on usability testing methods for video games. It can serve as a starting point for discussion and for future studies that will extend our knowledge of how the results of game testing (number of problems, severity) can be affected by the techniques used. Moreover, the almost complete lack of comparative user-based studies for video games, along with the equivocal results of such studies for productivity software, renders this study an important addition in the field.

As for the implications for the practice of usability testing on video games, it cannot be emphasized enough that the strong results regarding the effectiveness of think-aloud protocol may set a paradigm for usability professionals. It is strongly recommended that they *let the user do the talking* while playing, incorporating think-aloud protocol in their testing routine.

5.4 A look into the future

The author's honest motivation behind this study was the development of a usability evaluation framework for video games. Something like that would require deep understanding of all usability evaluation methods and the way they are used in the context of video games. It would also require the precise identification of all aspects of the gaming experience that need to be evaluated. Last but not least, it would require insights into the requirements of the gaming industry. Such a framework would not be intended for researchers, but for application within the game development process.

6 Conclusions

If there is one lesson to be learnt from this work, let it be that asking the gamers to think aloud while playing a game can reveal more problems than asking them to play silently and verbalize afterwards. There was strong evidence that the use of think-aloud protocol in usability testing is both more effective and cost effective, especially when a small amount of participants is available.

In the introduction of this dissertation it was noted that *the designer is not the user*. The role of the user as evaluator of the design was made clear throughout this study. Just as the designer cannot fully understand the user's behaviour and identify all the problems of the game, the user cannot be allowed to take over the role of the designer. "The design should be left to designers" (Pagulayan & Steur, 2004). The balance has to be kept so that overzealous play testers do not impact negatively the integrity of the design and HCI must help with adequate methodologies. Future studies should endeavour to develop methods that not only find real problems, but also suggest plausible solutions.

This study does not "reinvent the wheel" of usability testing. It is, however, a fresh addition to the research on usability and a tool for the practitioner's inventory.

References

Baauw, E., & Markopoulos, P. (2004). A comparison of think-aloud and post-task interview for usability testing with children. In *Proceeding of the 2004 conference on Interaction design and children building a community - IDC '04* (pp. 115-116). New York, NY, USA: ACM Press. doi: 10.1145/1017833.1017848.

Bailey, R., Allan, R., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 409-413).

Barendregt, W., Bekker, M. M., Bouwhuis, D. G., & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and after some practice. *International Journal of Human-Computer Studies*, *64*(9), 830-846. doi: 10.1016/j.ijhcs.2006.03.004.

Blandford, A., Hyde, J., Green, T., & Connell, I. (2008). Scoping Analytical Usability Evaluation Methods: A Case Study. *Human-Computer Interaction*, *23*(3), 278-327. doi: 10.1080/07370020802278254.

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., Pidruzny, J. N., et al. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624-634.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*, 1297. New York, New York, USA: ACM Press. doi: 10.1145/985921.986048.

Calvillo Gámez, E. H., Cairns, P., & Blandford, A. (2008). Assessing the Gaming Experience using Puppetry. In *Evaluating User Experience in Games workshop at ACM CHI 2008*. New York, NY, USA: ACM Press.

Card, S., Moran, T. P., & Newell, A. (1983). *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates.

Carless, S. (2010). GDC Canada: Ubisoft's Boisvert On Efficient Game PlayTesting. *Gamasutra*. Retrieved from http://www.gamasutra.com/view/news/28460/GDC_Canada_Ubisofts_Boisvert_On_Efficient_Game_PlayTesting.php.

Chen, J. (2007). Flow in Games. *Communications of the ACM*. Retrieved from http://www.jenovachen.com/flowingames/thesis.htm.

Desurvire, H., Caplan, M., & Toth, J. a. (2004). Using heuristics to evaluate the playability of games. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*, 1509. New York, New York, USA: ACM Press. doi: 10.1145/985921.986102.

Diaper, D. (1989). Task observation for human-computer interaction. In D. Diaper, *Task Analysis for Human Computer Interaction* (pp. 210-237). Ellis Horwood.

Donker, A., & Markopoulos, P. (2002). A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *Proceedings HCI 2002* (pp. 305-316). Springer.

Dumas, J. S., & Fox, J. E. (2008). Usability Testing: Current Practice and Future Directions. In A. Sears & J. A. Jacko, *The Human-Computer Interaction Handbook* (2nd., pp. 1129-1149). Lawrence Erlbaum Associates. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:USABILITY+TESTING:+CURRENT+PRACTICE+AND+FUTURE+DIRECTIONS#0.

Federoff, M. A. (2002). Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.8294&rep=rep1&type=pdf.

Gray, W., & Salzman, M. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction*, *13*(3), 203-261. doi: 10.1207/s15327051hci1303_2.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *13*(4), 373-410. doi: 10.1207/S15327590IJHC1304_03.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and Ergonomic Quality Aspects Determine a Software's Appeal. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00*, *2*(1), 201-208. New York, New York, USA: ACM Press. doi: 10.1145/332040.332432.

Henderson, R. (1995). An examination of four user-based software evaluation methods. *Interacting with Computers*, *7*(4), 412-432. doi: 10.1016/0953-5438(96)87701-0.

Hertzum, M., & Jacobsen, N. E. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *13*(4), 421-443. doi: 10.1207/S15327590IJHC1304_05.

Hopson, J. (2006). We're Not Listening: An Open Letter to Academic Game Researchers. *Gamasutra*. Retrieved from http://www.gamasutra.com/features/20061110/hopson_01.shtml.

Hopson, J. (2008). HCI impact and uncitedness. *Interactions*, *15*(3), 45. doi: 10.1145/1353782.1353794.

ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. *International Organization for Standardization*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Ergonomic+Requirements+for+Office+Work+With+Visual+Display+Terminals#1.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world. In *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91* (pp. 119-124). New York, NY, USA: ACM Press. doi: 10.1145/108844.108862.

Jennett, C., Cox, A. L., & Cairns, P. (2008). Being "In the Game." In *Conference Proceedings of the Philosophy of Computer Games* (pp. 308-323). Retrieved from http://opus.kobv.de/ubp/volltexte/2008/2468/.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and Defining the Experience of Immersion in Games. *International Journal of Human-Computer Studies*, *66*(9), 641-661. doi: 10.1016/j.ijhcs.2008.04.004.

Järvinen, A., Heliö, S., & Mäyrä, F. (2002). Communication and Community in Digital Entertainment Services. Prestudy Research Report. *Hypermedia Laboratory, University of Tampere*. doi: 10.1177/0196859904267230.

Kahn, R. L., & Cannell, C. F. (1957). *The dynamics of interviewing* (p. 368). New York: Wiley.

Korhonen, H. (2010). Comparison of Playtesting and Expert Review Methods in Mobile Game Evaluation. In *Proceedings of the 3rd International Conference on Fun and Games* (pp. 18-27). ACM Press.

Kremers, R., & May, A. (2009). Eufloria. Omni Systems. Retrieved from http://www.eufloria-game.com/index.html.

Laitinen, S. (2005). Better Games Through Usability Evaluation and Testing. Retrieved from http://www.gamasutra.com/features/20050623/laitinen_01.shtml.

Lazzaro, N. (2008). Why We Play: Affect and the Fun of Games. In A. Sears & J. A. Jacko, *The Human-Computer Interaction Handbook* (2nd., pp. 679-700). Lawrence Erlbaum Associates.

Lazzaro, N., & Keeker, K. (2004). What's my method?: a game show on games. In *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (pp. 1093-1094). New York, NY, USA: ACM Press. doi: 10.1145/985921.985922.

Lewis, C. H. (1982). Using the "Thinking Aloud" Method In Cognitive Interface Design. *IBM Research Report RC9265*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Using+the+'thinking-aloud'+method+in+cognitive+interface+design#0.

Medlock, M. C., Wixon, D., Terrano, M., Romero, R. L., & Fulton, B. (2002). Using the RITE Method to improve products: a definition and a case study. In *Usability Professionals Association*. Orlando, Florida.

Molich, R., & Dumas, J. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*(3), 263-281. doi: 10.1080/01449290600959062.

Newman, W. M. (1998). On Simulation, Measurement, and Piecewise Usability Evaluation. In G. M. Olson & T. P. Moran, *Commentary on "Damaged Merchandise?". Human-Computer Interaction.* (Vol. 13 (3), pp. 263-323).

Nielsen, J. (n.d.). Severity Ratings for Usability Problems. *Useit.com*. Retrieved from http://www.useit.com/papers/heuristic/severityrating.html.

Nielsen, J. (2000). Why You Only Need to Test with 5 Users. *Useit.com*. Retrieved from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2002). Getting access to what goes on in people's heads? - Reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 101-110). ACM Press.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (pp. 206-213). doi: http://doi.acm.org/10.1145/169059.169166.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (pp. 249-256). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=97243.97281.

Norman, D. (2004). *Emotional Design: Why we love (or hate) everyday things*. Basic Books. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Emotional+Design:+Why+we+love+(or+hate)+everyday+things#0.

Nørgaard, M., & Hornbæk, K. (2006). What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. In *Proceedings of the 6th conference on Designing Interactive systems* (pp. 209-218). ACM Press. doi: http://doi.acm.org/10.1145/1142405.1142439.

Ohnemus, K. R., & Biers, D. W. (1993). Retrospective Versus Concurrent Thinking-Out-Loud in Usability Testing. In *Human Factors and Ergonomics Society Annual Meeting Proceedings* (pp. 1127-1131). Human Factors and Ergonomics Society. Retrieved from http://www.ingentaconnect.com/content/hfes/hfproc/1993/00000037/00000017/art00001.

Pagulayan, R. J., & Steury, K. (2004). Beyond usability in games. *Interactions*, *11*(5), 70. doi: 10.1145/1015530.1015566.

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., & Fuller, T. (2007). User-Centered Design in Games. In A. Sears & J. A. Jacko, *The Human-Computer Interaction Handbook* (2nd., Vol. 14, pp. 741-759). CRC Press. doi: 10.1145/1273961.1273973.

Parnell, M. J. (2009). Playing with Scales: Creating a Measurement Scale to Assess the Experience of Video Games. Retrieved from http://www.uclic.ucl.ac.uk/distinction-projects/.

Picard, R. W. (1999). Affective Computing for HCI. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces - Volume 1* (pp. 829-833). Lawrence Erlbaum Associates. Retrieved from http://portal.acm.org/citation.cfm?id=742338.

Pinelle, D., Wong, N., & Stach, T. (2008). Heuristic Evaluation for Games: Usability Principles for Video Game Design. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (pp. 1453-1462). ACM Press.

Polson, P. G., Lewis, C. H., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, *36*(5), 741-773. Citeseer. doi: 10.1016/0020-7373(92)90039-N.

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction*, *9*(3), 213–234. Taylor \& Francis. doi: 10.1207/s15327590ijhc0903.

Sharp, H., Rogers, Y., & Preece, J. J. (2007). *Interaction Design: Beyond Human-Computer Interaction* (2nd.). John Wiley & Sons.

Strauss, A. L., & Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (2nd.). Sage Publications. Retrieved from http://www.google.com/books?hl=el&lr=&id=wTwYUnHYsmMC&oi=fnd&pg=PR9&dq=strauss+corbin+%22basics+of+qualitative+research%22&ots=VfU9eYpVRx&sig=K_EZbHOyfdqoPs_QTPYa1odyKBE#v=onepage&q&f=false.

TechSmith. (2010). Morae. Retrieved from http://www.techsmith.com/morae.asp.

Two Tribes. (2010). Toki Tori. Two Tribes. Retrieved from http://www.tokitori.com.

Varela-Alvarez, R. D. (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics*, *38*(10), 2030-2044.

Wixon, D. (2003). Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. *Interactions*, *10*(4), 28-34. ACM New York, NY, USA. doi: 10.1145/838830.838870.

van den Haak, M. J., & de Jong, M. D. (2003). Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. In *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.* (pp. 285-287). IEEE. doi: 10.1109/IPCC.2003.1245501.

www.Metacritic.com. (n.d.). Metacritic Game Rankings. *CBS Interactive Inc.*. Retrieved from http://www.metacritic.com/games.

Appendices

A.1 Consent form

(This form is to be completed independently by the participant after reading the Information Sheet and/or having listened to an explanation about the research.)

Title of Project: **Let the gamers do the talking**

This study has been approved by the UCL Research Ethics Committee [Project ID Number]: **MSc/0910/018**

Participant's Statement

I ……..

agree that I have

- ***read the information sheet and/or the project has been explained to me orally;***
- ***had the opportunity to ask questions and discuss the study;***
- ***understood that my participation will be audio/video recorded; I am aware of and consent to the use of these recordings only for purposes directly connected to this study (e.g. video data analysis), and that I will not be identified by name in these recordings, and that after the research is finished these recordings will be deleted or stored securely, in line with data protection requirements;***
- ***understood that my participation will be audio/video recorded and I am aware of and consent to, any use you intend to make of the recordings after the end of the project;***
- ***understood that I must not take part if I have had any signs of epilepsy or motion sickness when playing games in the past.***

I understand that I am free to withdraw from the study without penalty if I so wish and I consent to the processing of my personal information for the purposes of this study only and that it will not be used for any other purpose. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.

Signed: Date:

Investigator's Statement

I, **Efstratios Theodorou**, confirm that I have carefully explained the purpose of the study to the participant and outlined any reasonably foreseeable risks or benefits (where applicable).

Signed: Date:

A.2 Participant demographics

Table 4. *Participant demographics for Game 1 (Eufloria)*

Participant ID	**P05**	**P07**	**P08**	**P17**	**P19**	**P02**	**P06**	**P10**	**P13**	**P16**
Method	**Think-Aloud Protocol**					**Semi-Structured Interview**				
Age	29	22	26	22	28	22	22	26	27	22
Gender	M	F	M	M	M	M	M	M	M	F
Language	English	Other	English	English	Other	English	English	Other	Other	English
Frequency	4	3	3	4	2	5	4	4	3	3
Experience	5	5	5	5	4	5	4	5	3	4
Favourite Platform	Xbox 360	PC	PS3	PS3	PC	Xbox 360	PC	PC	PC	PSP
Favourite Genre	Sports	Strategy	Strategy	Shooter	Strategy	Shooter	Sports	RPG	Shooter	Fighting
How interested are you in strategy games?	4	5	5	4	4	2	4	5	4	4
Have you played Eufloria?	No	No	No	No	No	No	No	No	No	No

Table 5. *Participant demographics for Game 2 (Toki Tori)*

Participant ID	**P01**	**P04**	**P11**	**P15**	**P18**	**P03**	**P09**	**P12**	**P14**	**P20**
Method	**Think-Aloud Protocol**					**Semi-Structured Interview**				
Age	25	38	25	34	33	25	28	32	31	36
Gender	M	M	M	F	M	M	M	M	F	F
Language	Other	English	Other	English	English	Other	Other	Other	Other	Other
Frequency	3	4	4	1	3	3	2	3	3	1
Experience	4	5	4	5	5	4	5	5	3	1
Favourite Platform	PC	Wii	Web	Mobile Phone	Mobile Phone	PC	PC	Wii	Mobile Phone	PC
Favourite Genre	Shooter	Sports	Adventure	Platform	Puzzle	Strategy	Puzzle	Sports	RPG	Puzzle
How interested are you in puzzle games?	4	5	5	3	4	3	5	4	3	5
Have you played Toki Tori?	No	No	No	No	No	No	No	No	No	No

A.3 Pre-Gameplay Survey

This survey is to be completed before participating in this study. *[5 minutes]*

Age: _____

Gender: ☐ Female ☐ Male

First language: ☐ English ☐ Other

Please, answer a few simple questions about your gaming experience:

1. **How often do you play video games?**

   ☐ Every day
   ☐ A few times per week
   ☐ A few times per month
   ☐ A few times per year
   ☐ Less often
   ☐ Never

2. **For how long now have you been playing video games?**

   ☐ 0-2 years
   ☐ 3-5 years
   ☐ 5-10 years
   ☐ 10-15 years
   ☐ 15+ years
   ☐ Never

3. **Which of the following platforms do you use to play games?**

   ☐ Sony PlayStation 2
   ☐ Nintendo Wii
   ☐ Sony PSP
   ☐ Personal Computer *
   ☐ Web Browser *
   ☐ Sony PlayStation 3
   ☐ Microsoft Xbox 360
   ☐ Nintendo DS
   ☐ Mobile Phone *
   ☐ Other: _____

   ** Choose only if you use them for playing games*

4. **Which gaming platform do you use more often?**

   _____

5. **Which game types do you play?**

   a. Action
      e.g. Assassin's Creed, Prince of Persia
   b. Puzzle
      e.g. Tetris, Bejeweled
   c. Adventure
      e.g. Myst, Syberia
   d. Simulation
      e.g. Flight Simulator
   e. Shooter
      e.g. Doom, Call of Duty
   f. Sports
      e.g. Fifa, Virtua Tennis

g. Platform
e.g. Mario, Sonic

h. Fighting
e.g. Street Fighter

i. Role-Playing
e.g. Oblivion, Mass Effect

j. Strategy
e.g. Civilization, Age of Empires

k. Racing
e.g. Need for Speed, Moto GP

l. Casual
e.g. Facebook games, flash games

m. Party games
e.g. Guitar Hero, SingStar

n. Online multiplayer games
e.g. World of Warcraft, Everquest

o. Exergames
e.g. Wii Fit

p. Educational games
e.g. Democracy

6. Which of the above game types is your favorite?

_____

7. Which are your three favorite games (in order)?

#1 _____

#2 _____

#3 _____

8. How interested are you in strategy/tactics games?

1	2	3	4	5
Not at all		Moderately interested		Very interested

9. How interested are you in puzzle/platform games?

1	2	3	4	5
Not at all		Moderately interested		Very interested

10. Have you ever played *Eufloria*?

☐ Yes ☐ No

11. Have you ever played *Toki Tori*?

☐ Yes ☐ No

Thank you
Enjoy the gameplay session

A.4 Post-Gameplay Questionnaire

How would you **RATE** the game **OVERALL**?

0	1	2	3	4	5	6	7	8	9	10

Terrible Average Awesome!

Comments:

To what degree did you **ENJOY** the game?

1	2	3	4	5

Not at all Very much

Comments:

In your opinion, the game was:

Unpleasant								**Pleasant**
Bad								**Good**
Unaesthetic								**Aesthetic**
Rejecting								**Inviting**
Unattractive								**Attractive**
Discouraging								**Motivating**
Undesirable								**Desirable**
Boring								**Fun**

How would you **RATE** the following aspects of the game?

GRAPHICS

1	2	3	4	5

Very poor Average Very good

SOUND

1	2	3	4	5

Very poor Average Very good

DIFFICULTY

1	2	3	4	5

Very difficult Average Very easy

CONTROLS

1	2	3	4	5

Very difficult Average Very easy

What were the 3 things that you **LIKED** most about this game (in order)? Why?

#1	
#2	
#3	

What were the 3 **WORST** things that you experienced while playing this game (in order)? Why? Please, assign a **SEVERITY RATING** for each issue, according to the guide below.

Order	Problem description	Severity
#1		
#2		
#3		

Severity rating guide:

(1)	***Aesthetic issue***	*barely noticeable*
(2)	***Minor issue***	*slightly impairs gaming experience, if any at all*
(3)	***Intermediate issue***	*somewhat impairs gaming experience*
(4)	***Severe issue***	*seriously impairs gaming experience*
(5)	***Catastrophic issue***	*cannot continue playing if it persists*

If you could **CHANGE** only **ONE** thing in the game, what would it be?

End of questionnaire. Thank you for participating.

A.5 Allocation of participants

Table 6. *Means of participants' individual characteristics (quantified), based on their allocation across conditions. The standard deviation is shown in parenthesis.*

	GAME 1		**GAME 2**	
	Think-Aloud	*Interview*	*Think-Aloud*	*Interview*
Age	25.4 (3.3)	24.8 (2.5)	31.0 (4.8)	30.4 (4.2)
Game Playing Frequency[1]	3.2 (0.8)	3.6 (0.8)	3.0 (1.2)	2.8 (0.9)
Background Gaming Experience[2]	4.8 (0.4)	4.4 (0.8)	4.6 (0.5)	4.6 (1.7)

[1]*Frequency of video game playing was measured on a Likert scale as follows:*
5: Every day, 4: A few times/week, 3: A few times/month, 2: A few times/year, 1: Less often, 0: Never

[2]*Experience in video game playing was measured on a Likert scale as follows:*
5: 15+ years, 4: 10-15 years, 3: 5-10 years, 2: 3-5 years, 1: 0-2 years, 0: Never

A.6 Survey and questionnaire summaries

Table 7. *Summary of participants' answers to the pre-gameplay survey & post-gameplay questionnaire. Quantitative questions only.*

PARTICIPANT			**SURVEY**					**GAMEPLAY**		**QUESTIONNAIRE**		
ID	*GAME*	*METHOD*	*AGE*	*GEN*	*LANG*	*FREQ*	*EXP*	*TIM*	*LVL*	*RAT*	*ENJ*	*APP*
P01	G2	M1	25	M	O	3	4	58	16	2	2	1.7
P02	G1	M2	22	M	E	5	5	60	6	3	2	4.0
P03	G2	M2	25	M	O	3	4	60	16	3	2	5.2
P04	G2	M1	38	M	E	4	5	64	17	8	5	7.6
P05	G1	M1	29	M	E	4	5	64	4	6	3	7.1
P06	G1	M2	22	M	E	4	4	60	6	6	3	6.9
P07	G1	M1	22	F	O	3	5	60	4	7	4	5.0
P08	G1	M1	26	M	E	3	5	67	5	9	5	7.6
P09	G2	M2	28	M	O	2	5	60	6	7	4	7.4
P10	G1	M2	26	M	O	4	5	60	6	8	4	5.5
P11	G2	M1	25	M	O	4	4	58	19	7	4	5.5
P12	G2	M2	32	M	O	3	5	60	17	6	4	7.1
P13	G1	M2	27	M	O	3	3	60	5	7	4	6.7
P14	G2	M2	31	F	O	3	3	60	7	2	1	3.1
P15	G2	M1	34	F	E	1	5	66	9	7	4	7.1
P16	G1	M2	22	F	E	3	4	60	4	5	3	5.0
P17	G1	M1	22	M	E	4	5	52	4	5	2	4.3
P18	G2	M1	33	M	E	3	5	64	13	7	4	6.9
P19	G1	M1	28	M	O	2	4	62	4	6	3	5.2
P20	G2	M2	36	F	O	1	1	60	7	6	4	6.9

LEGEND
GAME: G1 = Eufloria, G2 = Toki Tori
METHOD: M1 = Think-Aloud Protocol, M2 = Interview
GEN: Gender, M = Male, F = Female
LANG: Language, E = English, O = Other
FREQ: Game playing frequency (1-5)
EXP: Gaming background experience (1-5)
TIM: Gameplay time (minutes)
LVL: Level reached during gameplay
RAT: Overall rating of the game (0-10)
ENJ: Overall enjoyment of gameplay experience (1-5)
APP: Game's appeal score (normalized, 0-10)

A.7 Coding Tables

Table 8. *Problem coding for Game 1. The numbers under each participant id indicate that the participant has reported the problem (i.e. cell is not blank).If the number is >0, the participant has also assigned a severity rating to the problem. For the severity rating scale, see bottom of table.*

	GAME 1 PROBLEM CODING			**THINK-ALOUD**					**INTERVIEW**				
id	*Description*	*Concept*	*Category*	*Participant Id 05*	*07*	*08*	*17*	*19*	*Participant id 02*	*06*	*10*	*13*	*16*
1	Cannot scroll right/left (only up/down)	Scrolling	Controls	0	4								
2	Not clear how to scroll / holding LMB or moving to the screen's edge don't work	Scrolling	Controls	0	0	0				3		3	
3	Player cannot assign planets to automatically send seedlings to another planet	Seedlings	Controls			3							
4	Player has to click-drag-click to control seedlings / too many clicks	Seedlings	Controls					0			0		
5	Not clear how to control a certain type of seedlings	Seedlings	Controls			0	0		0				
6	Not clear how to select multiple seedlings	Seedlings	Controls	0	0	0	0		0	0	0	0	0
7	Not clear how to control (send) seedling	Seedlings	Controls	3	4	0	0	0	0	0	4	0	0
8	No way to fully zoom out / view all planets at once	Zoom	Controls		0			3					0
9	Levels are too long	Levels	Design		0					0		0	
10	Text not explanatory enough - lack of pictures	Intro	Dialogues	0		0			0				
11	Instructions popup too long / too much text	Intro	Dialogues	0		0		0			0		
12	Level 3 outro is not clear; if level ends or there are more objectives	Language	Dialogues		0								
13	Level 2 intro not explaining what GROWERS or MOTHER TREE are	Language	Dialogues			0							
14	Level 3 intro not clear that it just progresses the story	Language	Dialogues			0							
15	Game is too easy / not challenging enough	General	Difficulty			3		0	0		0		
16	Unclear why player can't reach far planets	Controls	Gameplay	0			0						
17	Unclear why seedlings don’t follow a straight line	Controls	Gameplay				0	0					
18	Not clear how to conquer enemy planet / if more seedlings conquer it faster	Enemies	Gameplay	0	0				4				
19	No info about how to produce seedlings faster (e.g. core power)	Pace	Gameplay		0								
20	Confusing zoomed in aspect in the beginning of game	Zoom	Gameplay					0	0		0		0
21	Not clear what you have to do in the beginning	Begin.	Goals	0	0	0	0		0	0	0		0
22	Not clear how close you are to a level's end (no progress indicator or objectives)	End	Goals	0	4								0
23	Not clear what you have to do to finish a level	End	Goals	0	0						0		
24	Change of player's colour at each new level not explained	Colours	Graphics					0	0				
25	New empire (lvl) 4 has similar colour with player's empire	Colours	Graphics	0		0		0					
26	Circle that surrounds asteroids not explained	Cues	Graphics	0									
27	Planting a tree has no significant effect-small change visually	Cues	Graphics	0	0			0					

28	Graphics too abstract / colourless / simple / don't tie in with story	General	Graphics	3			3	3	0	0			
29	The line connecting the asteroid with its menu is confusing	Glitches	Graphics					0					
30	After discovering a planet, not clear that more appear, unless zoom out	Zoom	Graphics		0								
31	Zoomed out, no detailed visuals (e.g. number of trees, seedlings unclear)	Zoom	Graphics				3	0			0		
32	Zoomed-in aspect is visually appealing, but useless	Zoom	Graphics	0		2			3	0			
33	No info about the controls in PAUSE menu / have to restart lvl 1 to read	Hints	Instructions			0		0		0	0		4
34	SPEED attribute not clear or explained	Attributes	Labels		0								
35	Planet attributes generally not clear	Attributes	Labels	0			0				0		3
36	ENERGY attribute not clear or explained	Attributes	Labels		0	0		0	0	0			
37	Number on arrow when sending seedlings not clear what it indicates	Controls	Labels				0		0				
38	Tree limit not clear if it is a max limit or the desired amount to finish level	Planet	Labels		0								
39	Asteroid's label gets smaller when you zoom out	Planet	Labels	0									
40	In a new level, not clear why numbers of seedlings of all types are zero	Planet	Labels			0							
41	Planet menu has too much info - certain stats may be missed	Planet	Labels				0						3
42	OPTIONS menu items not clear (e.g. what Delay does)	Clarity	Menus						0				
43	Bad/Confusing use of English in menus	Language	Menus		0								
44	After pressing level's number, game doesn’t start	Level start	Menus		0								
45	Finishing level, next level not selected (If you press play, same level starts)	Level start	Menus					0					
46	After pressing level's number, player clicks the right arrow, not play	Level start	Menus				0	0					
47	Music is too calming - makes you feel tired/sleepy	General	Music				1	0	2		3	2	
48	Tutorial (first few levels) too slow / not interesting	General	Pace	0			0				3		
49	Gameplay is slow / boring / repetitive	General	Pace	0			4	0	4	4		4	
50	No incentive for player to finish game faster (e.g. time limit, more points)	General	Pace			0							
51	Slow rate of introducing new elements	Rate	Pace	4			4		0			0	
52	Seedlings take too much time to move	Waiting	Pace				0			0			
53	Unclear if player should be waiting after planting 1st tree	Waiting	Pace	0		0		0					
54	Trees take too much time to produce seedlings	Waiting	Pace					2	4	0			
55	Story is boring / not interesting	General	Story				0			0			

Table 9. *Problem coding for Game 2. The numbers under each participant id indicate that the participant has reported the problem (i.e. cell is not blank).If the number is >0, the participant has also assigned a severity rating to the problem. For the severity rating scale, see bottom of table.*

GAME 2 PROBLEM CODING				**THINK-ALOUD**					**INTERVIEW**				
s/n	*Description*	*Concept*	*Category*	*Participant Id*					*Participant id*				
				05	*07*	*08*	*17*	*19*	*02*	*06*	*10*	*13*	*16*
1	Character can't jump	Abilities	Controls	0	0	0		0			0		0
2	Ghost trap not clear how to use, unless you read instructions	Ghost trap	Controls	0	0	0			3				
3	Cannot fire the gun in the beginning of the level when facing forward	Gun	Controls				0						
4	No instructions about how to use freeze gun / clicking on enemy to kill it	Gun	Controls	0		0	0	0					
5	Use of gun requires that player moves the mouse/attention away from action	Gun	Controls	0			4						
6	Trying to select the gun, it fires	Gun	Controls		0					0			
7	Inconvenient placement of keys / you cannot change them	Keys	Controls						0	3			
8	Switching tools with keyboard is slow; important when you have to use the gun	Keys	Controls						0				
9	Clicking far away, not clear which path the character will follow	Path	Controls					0			0		3
10	Character must stand at very precise places to use teleport	Teleport	Controls	0	0		0						
11	Teleport gets stuck once selected / unclear how to unselect it	Teleport	Controls										0
12	Tool use is confusing at first / not clear that you have to click the tool's icon to use	Tools	Controls	4	3		0	0				0	
13	Change of area (from forest to castle) is just a change in graphics	Areas	Design			0							
14	Player cannot identify with the chicken character	Character	Design	0									
15	Transitional level of two first areas only text, no video	Levels	Design	0									
16	Intro dialogues do not appear with level restart, only with restart through main menu	Accessibility	Dialogues				0						
17	Level 3 intro: bad English	Language	Dialogues					0					
18	READY? indication unnecessary	Waiting	Dialogues			0		0	4				
19	Not allowed to make mistakes / having to restart after every even small mistake	Actions	Gameplay	0	0	0					0	3	0
20	"Stupid" action elements (e.g. quickly avoiding enemies), despite being a puzzle game	Actions	Gameplay							0			
21	Solution of levels is determined / no freedom to explore alternative solution	Freedom	Gameplay	0								0	
22	You are not allowed to explore next levels if you are stuck in a previous one	Freedom	Gameplay								0	0	
23	Game not fun while playing / only fun when finishing a level	Motivation	Gameplay	0									
24	Game not engaging / motivating enough to make you want to solve the puzzles	Motivation	Gameplay									3	
25	No replayability once you have cracked the level	Motivation	Gameplay										0
26	When you unlock a tool, it's not always available in your inventory (like RPG games)	Tools	Gameplay									0	

27	Mixed message: game looks childish, but it's too challenging	Appeal	General	0	0		2					0		
28	Not clear what you have to do in level 1 (go through level and collect the eggs)	Clarity	Goals				0	0			0	0		
29	When introducing the bridge (lvl 2), not clear that the goal is still to collect the eggs	Clarity	Goals				0							0
30	Yellow cue for changing direction is confusing	Cues	Graphics	0		0		0						
31	Average / poor graphics	General	Graphics	0		2					4			0
32	Height of steps is confusing (you can walk on short ones, but not on a bit higher ones)	Glitches	Graphics							0		0		
33	Lava pits unnecessary; you can walk over them, although they appear like gaps	Glitches	Graphics	0	0			0						
34	Waterfalls confusing / too opaque; do not add to the gameplay	Glitches	Graphics					0						
35	No in-game hints / you have to restart to read instruction dialogues	Accessibility	Instructions	0		0		0	0					
36	No hint when you get stuck at the same point many times	Accessibility	Instructions					0						0
37	Unclear instructions / wrong level of information	Clarity	Instructions				5			0		3		0
38	Egg-counter invisible / not clear what it shows	Accessibility	Labels	0										0
39	Toolbar not visible enough / bad placement	Accessibility	Labels		0									0
40	Number of uses for each tool not visible/clear enough	Accessibility	Labels				0	4				0		4
41	Top-left white arrow in menus, not visible enough	Accessibility	Menus				0							
42	At the level-selection menu, level icon not clear that is a button / can be pressed	Level sel.	Menus				0							
43	Bonus levels not accessible, even when unlocked	Level sel.	Menus					0						
44	CONTROLS option not properly placed in Pause menu (wrong position & cues)	Options	Menus		2									4
45	RESTART option ambiguous (restarts game or level?)	Options	Menus		0		0	0						
46	TO MENU option not clear	Options	Menus					0						
47	Creating profile, no button to save the name	Profile	Menus			0	0							
48	No confirmation/explanation of Wildcard when selected	Wildcard	Menus	0		0								
49	You have to exit the gameplay to use the Wildcard	Wildcard	Menus					0						
50	Music gets annoying soon	General	Music	4	4	3	0	3	5			0	0	
51	When music is off, no rewarding sound at the end of the level	Sound FX	Music		0			0						
52	No music / sound effects in the intro	Sound FX	Music					0						
53	Game is slow paced / boring (generally)	General	Pace	0	0	0								
54	Tutorial levels are slow paced	General	Pace				0	3						
55	Gameplay is repetitive after a few levels / levels are similar	Repetiton	Pace			2		0	0	0		0		
56	Chicken walks too slowly	Waiting	Pace	0										
57	Story is invisible / not interesting	General	Story							4		0		

A.8 Time-problem analysis (for cost effectiveness)

Table 10. *Time-problem analysis table, per participant. Cost-effectivenss (in percentage of the game's problems per hour) is the quotient of PROBLEM RATIO / TOTAL TIME. The ratio refers to the percentage of the game's problems found by the participant. Total time is the sum of all times (in minutes).*

PARTICIPANT		SESSION DURATIONS			ANALYSES TIMES		TOTAL	PROBLEMS	COST
PID	COND	GAMEPLAY	FIX	INTERVIEW	TRANSCRIBING	CODING	TIME	RATIO	EFFECTIVENES
P01	G2M1	58	15	0	85	33	191.0	36.8%	11.57%
P02	G1M2	60	15	16	25	20	136.3	32.7%	14.41%
P03	G2M2	60	15	14	21	9	118.5	12.3%	6.22%
P04	G2M1	64	15	0	75	15	169.0	24.6%	8.72%
P05	G1M1	64	15	0	75	39	193.0	40.0%	12.44%
P06	G1M2	60	15	11	16	9	110.7	23.6%	12.81%
P07	G1M1	60	15	0	65	36	176.0	34.5%	11.78%
P08	G1M1	67	15	0	80	30	192.0	32.7%	10.23%
P09	G2M2	60	15	13	19	10	117.1	12.3%	6.29%
P10	G1M2	60	15	19	31	13	138.0	23.6%	10.28%
P11	G2M1	58	15	0	80	18	171.0	24.6%	8.62%
P12	G2M2	60	15	14	18	7	113.5	10.5%	5.56%
P13	G1M2	60	15	13	20	7	115.4	12.7%	6.62%
P14	G2M2	60	15	16	21	18	129.8	24.6%	11.36%
P15	G2M1	66	15	0	75	15	171.0	29.8%	10.46%
P16	G1M2	60	15	11	17	10	113.1	16.4%	8.68%
P17	G1M1	52	15	0	55	15	137.0	32.7%	14.33%
P18	G2M1	64	15	0	75	14	168.0	38.6%	13.78%
P19	G1M1	62	15	0	78	40	195.0	38.2%	11.75%
P20	G2M2	60	15	28	35	8	146.1	24.6%	10.08%

NOTES:
PID: Participant's ID, COND: Condition of experiment
FIX: An average of 15 minutes was added for the time spent for the survey and the questionnaire
All the times are in minutes.