Comparison of Playtesting and Expert Review Methods in Mobile Game Evaluation

Hannu Korhonen
Nokia Research
P.O. Box 1000
33721 Tampere, Finland

hannu.j.korhonen@nokia.com

ABSTRACT

Selecting an evaluation method for product evaluations depends on many issues, such as the development stage of the product, time schedule, resources, and money that can be invested on the evaluation. The user testing and expert review methods are probably the most common ones when productivity software is being evaluated. Conducting a playtesting with players is commonly used by game designers, but the expert review method has not received that much attention, although it has proven to be an efficient and useful method. In this paper, we present a comparison study of the playtesting and expert review methods in mobile game evaluation. Our objective is to compare the effectiveness of the expert review method with playtesting. Results indicate that the expert review method is able to identify playability problems as accurately as playtesting, but in addition, it identifies problems that are crucial for the playability of the game.

Categories and Subject Descriptors

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces, *Evaluation/methodology*.

General Terms

Measurement, Experimentation, Human Factors

Keywords

Evaluation method, Expert Review, Playtesting, Playability Heuristics, Comparison Study.

1. INTRODUCTION

HCI researchers have developed multiple evaluation methods for testing the usability of productivity software. Selecting a method is not only about the method itself, but it also depends on other attributes such as resources, time, and money that can be invested on the evaluation. There are basically two mainstreams of evaluation methods: the user testing and inspection methods. In user testing, persons from a target user group interact with a product, and their behavior and experiences are collected either automatically or by observers. Even though this method enables developers to see how the product is used and perceived by the target users, arranging a user testing is time and resource consuming [25].

There are multiple inspection methods available that can be used for product evaluations. Nielsen calls these inspection methods discount methods, since they do not usually include users from the target user group, but the evaluation is conducted by experts [13]. The most popular inspection method is the expert review method (a.k.a. Heuristic Evaluation) developed by Nielsen and Molich [20]. In this method, usability experts evaluate a product by using usability heuristics (e.g. [18]) as principles to check the user interface of the product. The benefit of the expert review method is that it is fast, and knowledgeable experts can conduct an evaluation within couple of hours by using functional prototypes, low-fidelity prototypes (i.e. paper mockups) or even concept and interaction descriptions.

Despite its popularity in productivity software evaluations, the expert review method has not received much attention among game designers, or at least the usage of the method has not been reported extensively. In game design literature, focus groups and playtesting are mentioned frequently as evaluation methods for evaluating game designs [7], [24], [26]. Recently, game researchers have started to develop the expert review method for game evaluations [3], [6], [10], [23], and their focus has been in developing required heuristics. Comparing different evaluation methods in game evaluations has not been studied extensively.

Games differ in many respects from productivity software, which should be taken into account when applying evaluation methods developed for productivity software evaluations to game evaluations. Pagulayan et al. describe differences between games and productivity software [14]. According to them, productivity applications are tools, and the design intention is to make tasks easier, efficient, less error-prone, and increase the quality of the results. Games, instead, are intended to be pleasurable to play and sufficiently challenging in order to provide a good gaming experience. In addition, learning the goals, strategies and tactics to succeed in a game is part of the fun [14]. The intention of the game evaluation is to reduce the obstacles of fun, rather than the obstacles of accomplishments. This is a remarkable difference compared to productivity software evaluation, which focuses on the efficiency and ease of use of the product.

The contribution of the paper to the game research domain is that it provides empirical results on how the expert review method results correspond to playtesting results. In this study, we evaluated a mobile game by using two methods, and the results indicate that the expert review method very accurately identifies

Fun and Games 2010, September 15-17, 2010, Leuven, Belgium.
Copyright 2010 ACM 978-1-60558-907-7/10/09…$10.00.

the playability problems that came up in playtesting. Therefore, using the expert review method should be included in the evaluation method toolbox of game designs, and could be used to evaluate playability problems in games.

2. RELATED WORK

In this section, we review some studies that have applied the user testing and expert review methods to evaluating productivity software or web pages, and have compared evaluation results. Then we take a look at published game evaluations that have used several evaluation methods to evaluate digital games.

2.1 Comparative Productivity Software Evaluations

Karat [17], (p.204) describes that the number of identified usability problems and their type is one of the issues that define the usefulness of an evaluation method when comparing different evaluation methods. Molich et al. have conducted a series of comparative usability evaluations to find out what kinds of usability problems are found in the same application or web page [14]. In these studies, voluntary usability teams plan and conduct an evaluation of an application or web site by using the method they prefer and report results. Molich et al. found that the correspondence of the results is not very high, since 75% of the identified usability problems were reported only by a single team [14]. Doubleday et al. have reported that in their study, 39 % of usability problems that were identified in the user testing were not discovered in the heuristic evaluation [2].

Desurvire [17], (p.175) states that if the majority of identified problems are minor problems and there are more problems identified than in the laboratory test, an inspection method should not be considered very useful. The results from Jeffries et al. [5] showed that the heuristic evaluation method identified the most usability problems and more of the serious problems. However, the results of the study are inconclusive because different evaluation methods seemed to identify different usability problems. The overlap of the identified problems was only 10-15 percent between the methods [17], (p.210). Other researchers have reported similar results when evaluating productivity software. Desurvire reported that in their study, usability experts found 44 percent of usability problems that also occurred in the usability testing [17], (p.185). Karat reported that in their study, a third of the significant usability problems were identified with all methods. Gray and Salzman have criticized that many of these older studies contain errors in the experiment design, resulting in the fact that conclusions drawn from the data are not reliable or valid [8]. They state that the experiment should be designed very carefully to achieve reliable and comparable results.

There are at least three reasons to why evaluation results are different. First, in some studies, the evaluated product was not the same. Desurvire et al. reported that the user testing was conducted with the real system, whereas usability experts evaluated a set of flowchart diagrams of the system [17], (p.181). Another reason is that tasks or scenarios that are used in the evaluation are not the same. Molich et al. allowed each evaluation team to develop their own set of tasks, and the result was that almost half of the tasks were unique and used only by one team [14]. The third reason for different results is the "evaluator effect" [4]. Nielsen and Molich also reported that, regardless of the same background, the correlation between the numbers of usability problems identified by individual evaluators was not high. Nielsen [13] (p.59) concluded that it is preferable to use usability specialists to conduct evaluations, and for optimal performance, they should be double experts[1]. Recruiting double experts to conduct an evaluation is not an easy task, and therefore, it is a tempting idea to use actual end users as evaluators. However, research results indicate that this does not work, since the users do not have the required knowledge and understanding of usability principles [13] (p.59). Muller et al. propose the Participatory Heuristic Evaluation method, in which end users participate in the heuristic evaluation [10]. This is reasonable when end users are members of the development team.

Recent studies have indicated that both user testing and expert review can provide similar kinds of results. Molich and Dumas have studied how the usability testing and expert review methods compare. The results from this study show that the expert review method can provide valuable results in terms of identified problems when compared to user testing [13].

2.2 Game Evaluations

User testing or playtesting is the most popular evaluation method, and it is also described as the main evaluation method for game designers in game design literature [7], [24], [26]. Pagulayan et al. have conducted playtesting by using open-ended tasks. The purpose of these tests is to gather data on how players prioritize tasks and goals in the game [14]. They can also reveal how players understand the game mechanics or controls in the game.

Analytical evaluation methods and their potential applications to game evaluations have also been studied. The main focus of these studies has been to develop heuristics that are used during the evaluation [3], [10], [23]. More recently, a comparison study for evaluating two heuristic sets has also been conducted [11]. Baauw et al. have introduced an analytical expert evaluation method called the Structured Expert Evaluation Method (SEEM). Instead of using heuristics, the method includes a set of questions based on previous work from Norman and Malone [2]. This method has been used to evaluate games for children.

There is also increasing interest towards using automatic recording of interaction with the game systems to evaluate playability issues and gaming experience. Kim et al. have introduced the TRUE system which records data streams combining events of interest, contextual information and subjective opinions of the players to get a holistic view of what is happening in the game and possible problems the players have in playing the game [9]. Drachen and Canossa have also presented an instrumentation system that is used at IO Interactive [4]. This system collects various data from the play session. These kinds of systems are excellent tools for gathering very detailed information about player behavior in the game.

Studies on the comparison of different methods in game evaluations are very minimal. Laitinen has conducted a usability expert evaluation and user testing in a case study of a computer game [8]. The expert evaluation used traditional usability heuristics originally defined by Nielsen and Molich [12]. In the user testing, moderators introduced the game to the players and

[1] The evaluator has knowledge of general usability principles as well as product domain and task flows of the end users.

explained the background story and the starting point of the game. Players played the game for one and the half hours, and for the last 15 minutes, a cheat mode was activated in order to evaluate features that were not directly available for the players. The results do not directly indicate how well the results of the evaluation methods correspond to each other. However, Laitinen concluded that both methods provide useful data and identified problems in the design that were new to game developers [12].

Desurvire et al. performed a comparative evaluation of a computer game using both heuristic evaluation and playtesting [3]. Heuristics that they used in the inspection were specifically designed for game evaluations. One playability expert performed the heuristic evaluation. The two-hour playtesting was arranged with four players. The evaluation was conducted with an early prototype of the game. Results of the study indicate that both methods identify similar kinds of playability problems. The playtesting identified specific problems in the interface, whereas the expert evaluation identified general interface design issues.

3. The Study

The objective of the study was to compare the playtesting and expert review methods and to see how much evaluation results differ when they are used to evaluate a mobile game. The expert review was conducted before the playtesting. The author analyzed evaluation data after both evaluations were finished.

3.1 The Mobile Game

The mobile game was a commercial 3D action/puzzle game that can be played on smart phones. The game was a new version of one of the first games that started gaming on mobile phones. Therefore, evaluators and players who participated in the playtesting were familiar with the game concept. The evaluators and four out of six players have played the previous version of the game. The game was still under development, but the evaluated version was launched for beta testing, which means that all features of the game were implemented and it could be played on a target device [15]. This allowed the game to be evaluated thoroughly and the evaluation sessions were realistic. Both evaluations were conducted by using the same version of the game. The evaluators and the players used Nokia N73 mobile phones to play the game.

3.2 The Expert Review of the Game

The expert review was conducted by two playability experts and it followed recommended procedure [21], [19]. The evaluators were selected based on their expertise in conducting game evaluations and productivity software evaluations. In addition, both the evaluators play different mobile games regularly, which gave them expertise with mobile games. They were also familiar with the playability heuristics (See Appendix A) that are specifically designed for evaluating mobile games [10].

The evaluators did not belong to the development team of the game, and they did not have any previous experience with the game prior to the evaluation session. This provided a realistic context for the evaluators, and it resembled the situation that players will face when they get a new game for their device. The evaluators were instructed both to explore the user interface of the game and to try to complete as many levels as they could during the evaluation session. The instructions did not include tasks or scenarios that the evaluators should follow, but they were free to explore the game as they liked.

The evaluation started from the moment when the evaluators launched the game for the first time. First, the evaluators examined the game menu and general settings of the game and walked through the first levels that served as a tutorial for the game. Identified playability problems were written down briefly and the violated heuristic was assigned. The purpose was to keep the paper work of the evaluation to a minimum during the first moments, because it would otherwise disturb the evaluators' gaming experience. However, it is extremely important to record these first impressions with the game, because a player will learn and adapt quickly to design problems and valuable information is lost, if it is not recorded immediately.

As the evaluation continued, the evaluators focused thoroughly on both game usability and gameplay issues of the game. Sometimes the evaluators needed to play some levels several times before they could complete them. However, this also allowed them to explore the possibilities of the game and try out different strategies and playing styles.

The evaluation continued until the time required for finding new playability issues increased dramatically. This was based on the evaluators' own judgment. After that, the evaluators walked through identified playability problems one by one and discussed about their findings. The identified playability problems were clarified and duplicates were removed from a combined list. Finally, a violated heuristic and severity of the playability problem were assigned. Recommendations to fix the playability problems were documented as a final step of the evaluation.

The expert review session took approximately 3-4 hours including the evaluation of the game, discussion between the evaluators, and documenting the findings. Time reserved for playing the game was approximately one hour. The evaluators did not finish the whole game, but they completed the first eight levels.

3.3 The Playtesting of the Game

The playtesting was conducted in a usability laboratory and one participant was playing the game at a time. Six players were recruited for the evaluation. One participant out of the six was female. The average age of the players was 30 years, ranging from 26 to 35. All participants were experienced mobile phone users and they had played mobile games to some extent. Only two players played mobile games frequently. Others can be categorized as casual players of mobile games, which was the target population of the evaluated game.

The procedure of the playtesting followed the standard procedure of user testing (e.g. [5], [25]). In the beginning, the moderator instructed a player on how to think aloud during the session and collected background information. After that, the player was allowed to start playing the game. The moderator observed the game session and how the player played the game. The moderator also asked questions to verify his observations during the evaluation session. There was also another observer in an observation room to make notes from the game session. The session lasted 60-90 minutes including the introduction phase, playing the game, and a post-test interview. The time for playing the game was approximately 60 minutes, and the difference in the total time was caused by the length of the post-test interview which consisted of open-ended questions and a questionnaire.

Although the playtesting was similar to standard user testing, there was one significant difference. Instead of using specific or predefined tasks, the players were instructed with a single open-ended task which was formulated as follows:

> *"Play the game as you would play it on your own. The moderator will ask you questions and tell you when to stop."*

The same method is also used in game evaluations at Microsoft Games User-Testing Group [22]. The open-ended task does not instruct a player to perform or achieve anything in particular, or to play the game in a certain way. Instead, it allows the moderator to observe whether a player understands the goals and other aspects in the game. In our opinion, this makes the evaluation session more realistic and it also corresponds to the situation where the experts were playing the game.

The playtesting sessions were video recorded for later analysis. The recording system consisted of two cameras and a video mixer. One camera was recording the facial expressions and body movements of the participant, and another camera was mounted to a mobile phone to capture events of the screen and interaction with a keypad (Figure 1). During the sessions, the video stream was also transmitted to a monitor located in front of the moderator. This enabled the moderator to have a clear visibility of the content on the screen and observe the player's actions without disturbing them. Having the moderator next to the player also allowed for natural conversation during the test session.

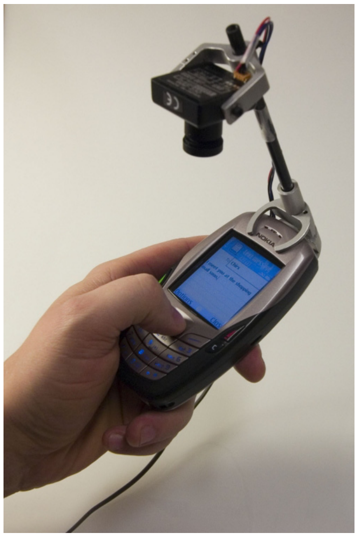

Figure 1. A mini camera mounted to a mobile phone (The mobile phone differs from the actual device used during the playtesting)

3.4 The combined analysis of the evaluation results

Our main research question for the study was how well an expert evaluation and playtesting can identify the same playability problems in the game. In order to do this, we harmonized findings from both evaluations. In the analysis phase, all playability problems were analyzed one by one and duplicates were removed. We then categorized the remaining problems according to our playability heuristics and severity rankings. Finally, results from both evaluations were cross-checked to ensure that the same playability problems were recorded equally.

4. Results

In this section, we describe results from both the expert review and playtesting and go through the differences that were found from the results.

4.1 Playability problems

The number of playability problems reported by the two evaluation methods was quite similar. Six playtesting sessions reported 46 playability problems altogether, whereas the combined list of playability problems from two experts contained 32 playability problems. There were both common problems and unique problems reported by a single method.

The combined list of playability problems from both evaluations contains 53 playability problems. We did not include playability problems violating mobility heuristics in the analysis, because those were reported only by the two playability experts. Evaluating mobility aspects during the playtesting would have made the test sessions more complicated. In addition, we excluded all positive observations from the analysis, as those were only reported by the experts and they do not contribute to the problem identification of the game. Hence, the analysis contains playability issues related to user interface and gameplay of the game.

Both evaluations were heavily user interface oriented, as 70% of the reported playability problems concerned user interface issues and 30% of the problems were identified as gameplay problems. Figure 2 shows problems categorized by the playability heuristics they violated. An interesting observation from the reported problems is that both evaluation methods reported 20 user interface problems, and in addition, playtesting reported 17 user interface problems. The expert review did not report any user interface related playability problems that would not have been reported by the playtesting.

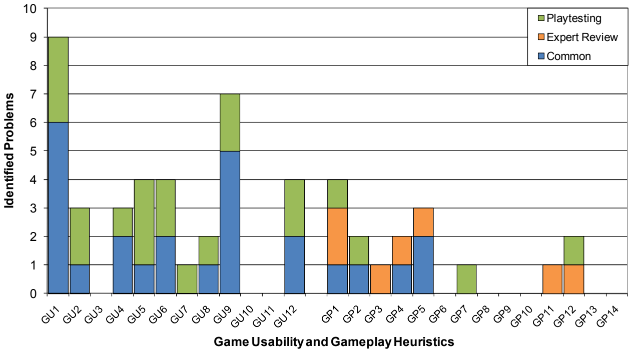

Figure 2. Playability problems categorized by violated heuristics

When looking at playability problems related to gameplay, we can see some differences in what kinds of problems were reported. There are 16 playability problems in total. Five playability problems are reported by both methods, seven problems are reported by two playability experts, and four problems are reported in the playtesting (Table 1).

Table 1. The number of playability problems and their severity identified by different evaluation methods

	Game Usability			Gameplay			
	Major	Medium	Minor	Major	Medium	Minor	Total
Common	9	10	1	3	2	0	25
Experts Only	0	0	0	3	4	0	7
Players Only	0	7	10	1	2	1	21
							53

Although 47% of the playability problems of the game were reported by both evaluation methods, there seem to be lots of playability problems that were reported independently from the other method (Figure 3). This provides an interesting starting point for viewing the differences between the results in more detail, as well as possible causes of these differences.

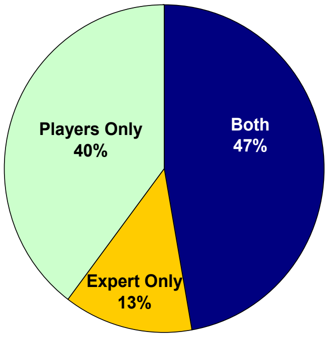

Figure 3. Playability problems identified by different methods

4.2 Playtesting Findings

Both evaluation methods focused on user interface issues, and 70% of the reported problems belonged to this category. The playtesting reported several user interface problems that were not discovered by the playability experts. When looking at these problems in detail, the first observation is that the majority of these problems seem to be quite player specific. The playtesting reported 21 playability problems in total that only players encountered (Table 1). However, in 81% of the cases (17 problems) only one out of six players encountered the problem. Furthermore, the problems did not seem to be severe, since the severity ranking of the problems was either medium or minor. 10 problems were rated as minor problems, indicating that someone finds them disturbing, but they do not really affect the player experience. Seven problems were rated medium, meaning that players initially find them disturbing, but once the players learned them, they are not problematic anymore. Examples of such playability problems are visualization and terminology issues and difficulties with control keys.

However, there were a couple of playability problems that were reported by several players. Since there were only six players participating to the playtesting, a problem reported by more than one player can be a significant issue and problematic to a larger player population as well. Two of them were ranked minor problems, whereas one problem had medium severity.

Two players had problems with in-game instructions, because they did not explain the purpose of some special items in the game clearly enough and the items were introduced to the players too early. The players tried to find these items on basic levels, even though they were available only on advanced levels. The playability experts did not consider this as a problem.

Three players had difficulties in understanding one visualization effect that was used in the game. When the player completed the level or a game character died, the screen went into grayscale mode. The players misinterpreted this to be some sort of special mode, because the game character was still moving in the game world and the player could control the character awhile. The playability experts did not point out this specific problem, because there were other problems affecting in this situation at the same time. The problem was ranked with minor severity.

The most severe playability problem concerning the user interface was related to indicators that were used to present essential game information to the player. Five out of six players had difficulties in recognizing indicators for the time limit, level progress and special abilities indicators on the screen. Fortunately, players learned these indicators during the game session, but they were not clear from the beginning, and it had a tremendous impact on their gaming performance. Usually, players started to investigate what these indicators meant once the moderator had noticed the possible problem and asked about it from the players. We were curious as to why playability experts did not report this problem, because it seemed like an obvious problem that should have been reported. The playability experts said that they did not consider the visualization of the indicators problematic, because it took only a short time for them to learn what they meant in the game.

From the gameplay, the players reported one particularly interesting problem that the playability experts did not report. Three players had problems in recognizing a new long-term goal, which was different than the long-term goal in the previous version of the game. Instead of having the game character avoid any obstacles in the game world, the players' main task is to maintain the energy level of the game character while performing short term goals (e.g. collecting items). The player can gain or lose energy by hitting obstacles in the game world. In addition, without new energy supplies, the energy level will drain slowly. In the previous version of the game, hitting an obstacle is lethal and the game character is self-sustained. The players did not notice this change and they were puzzled as to why the game character seemed to die without a reason. Other player-reported gameplay issues were mostly opinions rather than actual playability problems.

4.3 Expert Findings

Playability problems that were reported by the experts only were all related to the gameplay. The playability experts reported seven problems that were not discovered by the players (Table 2). Three of those problems were ranked as critical, but they were discovered from the levels that the players never played during the play session. However, it is very likely that these problems will materialize when the players reach those levels. The probability of encountering those problems is high, because they are all related to basic game mechanics. Two of these problems are related to collectable items in the game world. The game requires that items on some levels must be collected in a certain order, and if a player fails to do that, completing the level is impossible. These limitations will restrict the player's choice to complete a level and they may cause the game to stagnate. The third gameplay problem is related to substantially increased difficulty in the middle of the level. The game requires very accurate navigation in order to collect items sequentially, and the pace of the game is temporarily increased at the same time. The actual problem is that the collectable items are possibly misplaced in the game world, and the input devices of the mobile phone do not enable the navigation accuracy needed on the level.

Other four playability problems discovered by the experts were related to the goals, rewards, and challenge of the game. The severity ranking of these problems was medium. The playability experts reported two problems related to the goals of the game. The game has a short term goal, which gives a special bonus if a player manages to collect all bonus items from the levels. However, these items are permanently attached to specific levels, and if a player starts a play session other than the first level, it is not possible to collect all bonus items and thus get the bonus.

The second problem with the goals was related to conflicting short term goals. The primary goal of the player is to avoid collision with obstacles which are not energy supplies, and to collect items in the game world. In the current version, this goal alone is a challenging task for the player. However, each level also includes a time limit, and the player fails if the time runs out before all items are collected. The reason why these two goals are conflicting is that controlling the game character to avoid collisions and collecting items is time consuming, and the player cannot really hurry up actions that need to be completed.

The playability problem related to the rewards of the game is that a good high score can only be achieved by playing levels in sequence and starting from level one. Completing more advanced levels does not give significantly more points than easier levels, and it is easier to complete easy levels than to try to play more difficult levels. The lack of proper rewards on the advanced levels can reduce the motivation of the player.

The playability problem related to challenge was found on the level that players never reached in the playtesting. The level contained an enemy which could destroy the game character instantly. This was unexpected feature, because the game character has an energy level and hitting an obstacle or an enemy will always decrease the energy level, but never drain it completely. Therefore, encountering such a strong enemy sounds like an unbalanced game feature and thus, it was reported.

4.4 Effectiveness of Evaluation Sessions

In playtesting, the participants' performance in terms of evaluation efficiency and how many playability problems they encountered was quite similar. On average, one session reported 20.17 (σ 2.23) playability problems and tendency was towards critical problems than minor problems (Table 2).

Table 2. Distribution of identified playability problems

	Mean	SD
Major	9.33	2.07
Medium	8.33	1.03
Minor	2.50	0.55

When evaluating the efficiency of the playability experts, we compared the average number of players who encountered a playability problem in playtesting which was also reported by the experts to those problems that were only reported by the players. The independent t-test indicates that there is a statistically significant difference in means, $t(43)=2.02$, $p<0.01$.

5. Discussion

Recent studies indicate that both the expert review and user testing have provided similar kinds of results [13], [27]. Our results also show that the expert review and playtesting identified playability problems quite consistently. There were altogether 53 playability problems identified from the game. 70% of them were related to user interface aspects such as visualization, feedback and controls, and 30% of the playability problems were related to gameplay issues. Distribution between the user interface and gameplay problems corresponds to the previous study [6]. One probable reason for this is that evaluating user interface issues is easier than finding gameplay, issues because problems in information visualization, feedback and navigation are apparent to both the players and the experts. An interesting observation from the results is that the expert review identified fewer problems than playtesting. Usually, Human-Computer Interaction (HCI) literature states that the expert review method tends to find more problems than user testing [13] (p.56), and quite many of them are false alarms that are not verified in the user testing [16], [11].

In this study, the expert review method was very well able to predict playability issues that cause problems for players. Both evaluation methods jointly reported 20 playability problems from the user interface of the game. Nine problems were ranked as critical problems, and 10 problems had medium severity. This indicates that both methods have found the most serious problems from the user interface. In addition, the playtesting identified 17 user interface problems. However, a detailed analysis of these problems indicated that usually only one out of six players encountered these problems, and more than half of them were ranked as minor in the severity classification.

For gameplay related issues, the diversity of the problems was greater. Both methods identified 5 common playability problems in the gameplay. Three of them were major problems, and 5-6 players encountered problems which were also reported by the experts. Two other gameplay problems were ranked with medium severity. In addition to jointly reported problems, the playability experts reported seven additional problems. During the evaluation session, the experts were able to progress further in the game, and they uncovered playability problems from levels that players never played in the playtesting. In addition, the experts uncovered playability problems related to goals and rewards that were not identified by the players. Reporting these kinds of problems can be regarded as specific benefit of the expert review method, because they require that the game is played repeatedly and the game mechanics are explored systematically.

The playtesting reported four playability problems that were not identified by the experts. One of them is a particularly interesting problem, because the players had difficulties understanding the long-term goal of the game as it conflicted with their previous experience of the earlier versions of the game. The game designers had dramatically altered the behavior of the game character and what the player must do to keep the character alive. The playability experts did not report such problem, because it was not directly related to the current game design. Of course, the experts should consider the previous experience and knowledge of the target users and how it might influence on gaming experience, but in this case, it was not a problem that everybody would face when playing the game. During the playtesting, half of the players reported this particular problem.

The results of our study indicate that expert review is an efficient evaluation method for evaluating the playability of mobile games, and the method can be used to complement results from playtesting. In our study, the evaluation procedure was similar in both cases, which will make the comparison of the results feasible. In both evaluations, the experts and players were able to play the game as they would normally do. We did not give any specific tasks that should be completed during the evaluation, but they were allowed to explore the game world as they would like to. Games, in general, are quite linear at the beginning and players are guided through the first missions or levels by the game design [1]. Therefore, the progress of the expert review will correspond to the progress of the players, and the experts can explore issues that are possibly problematic for players.

There are several advantages to using experts to evaluate game design instead of inviting players for playtesting sessions. Quite often, the playability experts have required domain expertise which is recommended by literature when recruiting evaluators [18]. The evaluators may belong to the target audience of the game, and playing games regularly will provide good baseline knowledge of different games. This is also a distinguishing aspect when compared to productivity software evaluations, where finding a domain experts can be much harder.

Another advantage of using playability experts is that although they are playing the game, their main focus is still in the evaluation, and they are analyzing their own behavior in order to identify possible playability problems in the game design. In this evaluation mode, they are more sensitive to recognizing problems and writing them down immediately. In playtesting, players are focusing on playing the game and more easily immersed by the gameplay. In this mode, they do not necessarily recognize problems they are facing, but instead they try to overcome them. This is a contradicting situation, since the main focus should be in identifying playability problems from the game, but if the game is good, players will be immersed by the game and it is difficult for them to express problems.

In our study, the players were frequently silent for a long period and playing intensively. Even though it is the moderator's responsibility to observe the behavior of the players, it is impossible to recognize all problems that the players facing in the game if they do not express them to the moderators. In addition, it is often very difficult for the moderator to start asking questions because it will break the immersion and disturb the player experience [26]. For many players, it seemed to be very difficult not to be too immersed and think aloud while playing the game. Therefore, we tried to direct our questions to proper moments (e.g. completing the level or restarting the level) that would disturb the game play the least. However, the problem with this approach was that the problematic moment had already passed, and the players were not always able to recall the problem that they were struggling with. Therefore, our observation supports Laitinen's findings [8].

The third advantage of using playability experts is that they will use the time reserved for evaluation efficiently, and from the evaluation point of view, they can evaluate the game more thoroughly than is possible in playtesting. In our study the expert review lasted three to four hours in total, and the experts managed to explore the game more thoroughly than the players in playtesting, although time for playing was approximately same. In our study, the playtesting sessions were limited to two hours. Completing all playtesting sessions took 12 hours. During playtesting, the players did not progress as far in the game as the playability experts. Especially if the players encountered a very challenging situation or a level, they spent lot of time trying to solve it. These challenges are part of the game, but if they do not contain any playability problems, they are only consuming valuable evaluation time. One solution to this problem is to provide cheat codes to the players, but this will probably distort evaluation results. In our study, the evaluators faced the same challenges as the players, but since they were also skilled players, they solved the challenges quicker and were then able to move ahead in the game. However, it should be noted that it is good to reserve more time for game evaluations than what is usually reserved for expert reviews of productivity software. Korhonen and Koivisto have previously noted that game evaluations will take longer, because the evaluators need to play the game and solve challenges that are included in the game design [6].

The fourth advantage of the expert review method is that the experts pay attention to issues that the players might ignore, but they are still important from an evaluation point of view. In our study, the experts reported a few problems related to goals and rewards which were not covered by the playtesting. The players were so keen on completing levels that they did not pay attention to rewards or short term goals at all. The players commented on rewards after the moderator asked about them, but responses indicated that they were not interesting at this point. The motivation for playing the game was targeted towards seeing new levels than analyzing the results. The playability experts, instead, explored how players will progress in the game and what is achieved by completing levels. These observations discovered new playability problems in the game design.

The expert review method does not make playtesting obsolete, but it should be seen as a complementary evaluation method which can provide useful information for game developers with less effort and pinpoint obvious problems before playtesting. In our study, the expert review identified playability problems consistently with the playtesting method. Moreover, the expert review was able to conduct a more thorough evaluation of the game than playtesting. However, the playtesting also reported some playability problems which were not identified by the playability experts. The most important finding was related to the conflict between the players' previous knowledge and assumptions and the current design of the long-term goal of the game. These kinds of findings are very difficult to achieve with the expert review method, because the evaluators' previous experience and knowledge influence what kinds of playability problems are identified. Another issue with the expert review is that it cannot describe the feelings and experiences that a game elicits from players, because the experts can only describe their own experiences. Therefore, playtesting is needed to explore the experiences of the game.

The limitations of the study are that we have compared the evaluation results of one mobile game. The results are, however, very encouraging and they provide initial findings that the expert review method provides useful data for game developers. Another limitation of the study is that we evaluated a game in which completing the first levels of the game is controlled very tightly by the game system. For other kinds of games, (e.g. MMORPGs or sandbox games) the evaluation procedure might need tighter control as regards what aspects the evaluators should focus on during the evaluation.

In the future, we plan to continue these game evaluations and collect data from several game evaluations to validate the results presented in this study. In addition, we need to investigate the playtesting method and try to find ways of improving the effectiveness of the method in game evaluations and to overcome challenges that immersed players will set. Especially, we need to concentrate on how to identify playability problems in the game content, which is the most important part of the game.

6. Conclusion

In this paper, we have presented an evaluation study that compared the effectiveness of the expert review and playtesting methods in a mobile game evaluation. Playtesting is a commonly used evaluation method, but the expert review method is mainly ignored. In this study, we explored how the expert review method compares to playtesting in terms of efficiency in finding playability problems. The expert evaluation was conducted by two playability experts, and six players participated in the playtesting. The expert review used playability heuristics that are specifically designed for mobile game evaluations. Playability problems were analyzed and categorized based on their severity and heuristic that they violated. The results indicate that the expert review was accurately predicting playability problems that players faced when playing the game. The expert review discovered the most serious playability problems from the user interface that were also reported by the players. Playtesting reported many playability problems that are very detailed and specific to a certain player and which were not reported by the experts. Playability problems related to gameplay were much harder to discover in playtesting, and expert review found several serious problems that were not discovered by playtesting. We observed several benefits of the expert review method that will make it an attractive method for game evaluations. The length of the playtesting session will limit the scope of the evaluation, especially if the players encounter a tough challenge and cannot proceed in the game. Immersion into the game makes recognizing playability problems harder for players and expressing their difficulties to the moderator is often inadequate or even missing completely, because they are engaged in the gameplay. For these reasons, the expert review with playability heuristics can provide a cost efficient and fast method for evaluating the playability of a game.

7. ACKNOWLEDGMENTS

We would like to thank Elina Ollila for her help in arranging both evaluation sessions. In addition, we would like to thank Dr. Juha Arrasvuori and Dr. Matt Jones for their valuable comments on the previous versions of the manuscript.

8. REFERENCES

[1] Adams, E., Rollings, A. *Game Design and Development: Fundamentals of Game Design*. Prentice Hall, 2007

[2] Baauw, E., Bekker, M. M., Barendregt, W. A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games. In proc INTERACT, (2005), 457-469

[3] Desurvire, H., Caplan, M., Toth, J. A. Using heuristics to evaluate the playability of games. In proc CHI Extended Abstracts, ACM Press (2004), 1509 - 1512

[4] Drachen, A., Canossa, A. Towards gameplay analysis via gameplay metrics. In proc MindTrek, ACM (2009), 202-209

[5] Dumas, J. S., Loring, B. *Moderating Usability Tests: Principles & Practices for Interacting*. Morgan-Kaufmann, 2008

[6] Federoff, M. A. Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games. Indiana University, 2002

[7] Fullerton, T., Swain, C., Hoffman, S. *Game Design Workshop: Designing, Prototyping, and Playtesting Games*. CMP Books, 2004

[8] Gray, W. D., Salzman, M. C. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *13,* 3 (1998), 203-261

[9] Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B., Pagulayan, R. J., Wixon, D. Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In proc CHI, ACM (2008), 443-452

[10] Korhonen, H., Koivisto, E. M. I. Playability Heuristics for Mobile Game. In proc MobileHCI, ACM Press (2006), 9-16

[11] Korhonen, H., Paavilainen, J., Saarenpää, H. Expert Review Method in Game Evaluations – Comparison of Two Playability Heuristic Sets. In proc MindTrek, ACM Press (2009), 74-81

[12] Laitinen, S. Do usability expert evaluation and test provide novel and useful data for game development? *JUS 1,* 2 (2006), 64-75

[13] Molich, R., Dumas, J. S. Comparative Usability Evaluation (CUE-4). *BIT 27,* 3 (2008), 263-281

[14] Molich, R., Kaasgaard, K., Karyukin, B. Comparative Usability Evaluation. *BIT 23,* 1 (2004), 65-74

[15] Mulligan, J., Patrovsky, B. *Developing Online Games: An Insider's Guide*. New Riders, 2003

[16] Nielsen, J. Finding usability problems through heuristic evaluation. In proc CHI, ACM Press (1992), 373-380

[17] Nielsen, J. Heuristic Evaluation. in Usability Inspection Methods, Nielsen, J., Mack, R. (Eds) Wiley & Sons (1994), 25-62

[18] Nielsen, J. *Usability Engineering*. Academic Press, 1994

[19] Nielsen, J. Usability inspection methods. In proc CHI Extended Abstracts, ACM (1994),

[20] Nielsen, J., Molich, R. Heuristic evaluation of user interfaces. In proc CHI, ACM (1990), 249-256

[21] Nielsen, J., Phillips, V. L. Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In proc CHI, ACM (1993), 214-221

[22] Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., Fuller, T. User-centered Design in Games. in Handbook for Human-Computer Interaction in Interactive Systems., Jacko, J., Sears, A. (Eds) Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (2003), 883-906

[23] Pinelle, S., Wong, N., Stach, T. Heuristic evaluation for games: usability principles for video game design. In proc CHI, ACM Press (2008), 1453-1462

[24] Rouse, R. *Game Design: Theory and Practice*. Wordware Publishing, 2001

[25] Rubin, J. *Handbook of Usability Testing: How to plan, design and conduct effective tests*. John Wiley & Sons, 1994

[26] Schell, J. *The Art of Game Design*. Morgan Kaufmann, 2008

[27] Tan, W.-s., Liu, D., Bishu, R. Web evaluation: Heuristic evaluation vs. User Testing. *39,* 4 (2009), 621-627

9. Appendix A

Game Usability Heuristics	
GU1	**Audio-visual representation supports the game**
	The game graphics should support gameplay and story and be informative for the player. In addition, the graphical look and feel should be consistent throughout the game. Audio can be used to evoke emotions and increase immersion. A good sound environment in the game supports a positive gaming experience. The graphics or audio should not prevent the player from performing actions or make them unnecessarily difficult.
GU2	**Screen layout is efficient and visually pleasing**
	The layout should present all necessary information for the player, but on the other hand, if the screen is filled with all kinds of information, it starts to look crowded. It is important that the player finds the navigation controls and they should not be mixed with the information that needs to be visible on the screen.
GU3	**Device UI and game UI are used for their own purposes**
	It should always be noticeable whether the player is dealing with the game user interface or device functions. The game interface should not use the device's user interface widgets in the game interface, because it breaks the immersion. The most impressive immersion is achieved when the game uses full-screen mode hiding other features.
GU4	**Indicators are visible**
	The player should see the information such as the current state of the game and status of the game character that is required for being able to play the game. Information that is frequently needed should be visible for the player all the time — if possible.
GU5	**The player understands the terminology**
	The terminology that is used in the game should be understandable and not misleading or unfamiliar for the players. Technical jargon should be avoided. For instance, terminology that is related to the game concept or features that the game needs from the device should be translated into more understandable language.
GU6	**Navigation is consistent, logical, and minimalist**
	Navigation consists of the game menu and the game world. The game menu consists of settings and selections for the desired game session. Different functions should be organized reasonably and possibly on different screens. However, long navigation paths should be avoided. Short navigation paths provide more clarity and are easier to remember. In the main game menu, the player should be able to start a game and have access to other important game features. In the game world, navigation should be intuitive and natural. Regardless of the complexity of the game world, players should be able to navigate there smoothly. With a proper set of control keys, navigation can be very intuitive and almost invisible.
GU7	**Control keys are consistent and follow standard conventions**
	Using common conventions in control keys reduces the time that is needed to learn to play the game since the player can use his or her knowledge from other games. Game devices usually have specific keys for certain actions and every game should follow them.
GU8	**Game controls are convenient and flexible**
	Novice players usually need only a subset of the controls when they start playing the game. On the other hand, veteran players often need shortcuts and more advanced commands. It should be possible to customize the game controls or use shortcuts or macros. However, using shortcuts should not provide a major edge in a competitive player vs. player game. The configurability and amount of controls needed to play the game should be kept at the minimum, but they need to be sufficient. In addition, the controls should be designed according to the device's capabilities.
GU9	**The game gives feedback on the player's actions**
	A good user interface has a low response time on the player's actions. An action can be either a single key press or a more complicated input sequence. The player should notice immediately that the game has recognized the action by providing feedback. The most common way of providing feedback is to present it graphically. Other alternatives are to use audio or tactile feedback. Providing only auditory feedback is not acceptable since a player may be playing the game without sounds. Although the game needs to respond immediately to the player's actions, the consequences of the action can be shown to the player later. If an action cannot be performed immediately, the game should notify the player about the delay.
GU10	**The player cannot make irreversible errors**
	The game UI should confirm actions that can cause serious and irreversible damage, which affects the player's ability to play the game. Such errors are typically related to the game character or player's progress in the game. When mistakes happen, it is helpful to enable recovery.
GU11	**The player does not have to memorize things unnecessarily**
	The game should not stress the user's memory unnecessarily, unless it is part of the gameplay.
GU12	**The game contains help**
	The players do not often read manuals. Instead, the game should teach the player what he or she needs to know to start playing the game. This can be done through a tutorial mode at the beginning of the game. The tutorial mode should be divided into chapters that teach a couple of things at the beginning. Ideally, the tutorial could be embedded completely in the game so that help would be provided every time when it is really needed. Help is also often needed in error situations. If the game provides useful error messages, the player can understand better what caused the problem.

Gameplay Heuristics	
GP1	**The game provides clear goals or supports player-created goals**
	The players should be able to understand goals that exist in the game. The goals can be either set by the game or created by the players. The game should contain both short-term and long-term goals. Short-term goals provide repeated opportunities for reinforcement and keep players motivated to play the game. Long-term goals are usually more difficult to achieve and they can consist of several short term goals.
GP2	**The player sees the progress in the game and can compare the results**
	The players should have enough information so that they can see their progress towards the goals in the game. The progress can be shown to the player explicitly, for instance with numbers, or implicitly, for instance, by changing the behavior of non-payer characters or the game world. The players feel more motivated if they can compare themselves with the other players or the previous achievements. Traditionally, this has been done with high-score lists, rankings, character levels, or different titles.
GP3	**The players are rewarded and rewards are meaningful**
	The players should receive a meaningful reward as they progress in the game. In addition, the reward should be adjusted to the challenge that the player had to face in order to get it. The rewards schedule should be varying and frequent, but still unpredictable.
GP4	**The player is in control**
	The game should provide at least an illusion that the player is in control of what is happening in the game world. The players should be able to decide on actions they want to take and these actions should have an influence on the game world.
GP5	**Challenge, strategy, and pace are in balance**
	The game should be designed so that the challenge is comparable to player's current skills, then the players do not feel frustrated or bored with the game. In single-player games, the player can often choose the difficulty level and thus affect the challenge. The players learn new strategies as they play the game. There should not be dominating strategies for any part of the game. The pace should be adjusted to the game style and it can be intensive or deliberate. The game should allow the player to take a deep breath once in a while during the play sessions.
GP6	**The first-time experience is encouraging**
	The first impression of the game is formed within a few minutes and it is very difficult to change. The players should feel that they have learned the basics and have accomplished something. The first play session should make the player desire for the next play session.
GP7	**The game story supports the gameplay and is meaningful**
	Even though the story plays an important role in many games, it should not dominate the gameplay. Some games do not even have or need a game story. If the game has a story, it should fit the other elements in the game and sound plausible to the player. The dialogue with non-player characters (NPC) should be meaningful and interesting for the player.
GP8	**There are no repetitive or boring tasks**
	The game should not require repetition of tasks without changing any conditions. Often, this repetition happens when the player needs to reach a certain goal before the game becomes interesting or challenging. However, during the training phase (tutorials), it is useful to repeat certain tasks so that the player learns and practices for example how the character is controlled in the game.
GP9	**The players can express themselves**
	The players should be able express themselves by, for instance, customizing their characters, acting in a certain way, or modifying the game world. Allowing the players to customize and personalize their game characters makes it more probable that they feel attachment to a game.
GP10	**The game supports different playing styles**
	The players can vary a lot in terms of both experience and preferred play styles. There are also different playing styles that should be supported at least in the more complex games. The player types are defined based on how the players prefer to interact with the game world or with the other players, Four common player types are: A) Achievers, who like to compete with the game mechanics. B) Explorers, who wish to explore different aspects of the game. C) Socializers, who prefer to socialize with other players D) Killers, who enjoy dominating other players.
GP11	**The game does not stagnate**
	The players should always feel that it is possible to reach the goals and the game progresses. The game should recognize immediately when the game is over and inform the players. Ending of the play session should be clearly indicated and restarting the game should be possible.
GP12	**The game is consistent**
	The game world and actions should be consistent and logical for the player. If something works in the beginning, the player assumes that it also works later on. Correspondingly, if the player is able to perform a certain action in the game world or for a game item, the player assumes that similar kind of action is possible for other similar objects or in the similar situation as well. Moreover, if the game world resembles the real world, the player assumes that the same principles also work in the game world. The game should not contain invisible walls.
GP13	**The game uses orthogonal unit differentiation**
	Each game item should have a purpose in the game world and it should be notably different to other similar game items. In addition, if the player needs to select character classes or roles in the game, they should be functionally different.
GP14	**The player does not lose any hard-won possessions**
	The game should maintain possessions that the player has earned while playing the game and the player cannot lose them accidentally. However, in some cases the game can provide very high risks and the player can stake valuable game items which can be lost during the gameplay.