

Expert Review Method in Game Evaluations – Comparison of Two Playability Heuristic Sets

Hannu Korhonen

Nokia Research

P.O. Box 1000

00045 Nokia Group, Finland

hannu.j.korhonen@nokia.com

Janne Paavilainen

University of Tampere

Kanslerinrinne 1

33014 Tampereen Yliopisto, Finland

janne.paavilainen@uta.fi

Hannamari Saarenpää

University of Tampere

Kanslerinrinne 1

33014 Tampereen Yliopisto, Finland

hannamari.saarenpaa@uta.fi

ABSTRACT

The expert review method is not yet widely adopted in game evaluations, although it has been used successfully in productivity software evaluations for years. In order to use the method effectively, there need to be playability heuristics that take into account the characteristics of the videogames. There are a few playability heuristic sets available, but they have several differences, and they have not been compared to discover their strengths and weaknesses in game evaluations. In this paper, we report on a first study comparing two playability heuristic sets when evaluating the playability of a videogame. The results indicate that the heuristics can assist the evaluators in evaluating both the user interface and the gameplay aspects of the game. However, playability heuristics need to be developed further before they can be utilized by the practitioners. Especially, the clarity and comprehensibility of the heuristics need to be improved, and the optimal number of heuristics is still open.

Categories and Subject Descriptors

H.5.3 [Information Interfaces and Presentation]: User Interfaces – *Evaluation/Methodology*.

General Terms

Experimentation, Human Factors

Keywords

Playability, Heuristic, Game Evaluation, Expert Review, Videogame.

1. INTRODUCTION

Competition in the game industry is hard and gaming experience has become a crucial factor in differentiating similar kinds of game titles. If the gaming experience is not optimal, players can easily switch to another game. Typically, gaming experience can be evaluated after there is a working prototype implemented and it is ready for beta testing. At this point, correcting any playability

problems (e.g. UI navigation is complex, goals are not clear, or the challenge level or pace is set incorrectly) is often too expensive, or the project schedule does not allow any delays due to marketing reasons. As a result, there is a need for an evaluation method that can identify these playability problems before beta testing starts and thus provide time for corrections.

Productivity software has been evaluated for years with the expert review method to find usability problems in the design and implementation [20]. The method is cost-efficient and effective, and the design can be evaluated already in early project stages. A skillful and knowledgeable usability expert can identify usability problems as accurately as in user testing [18]. Evaluating games with this method is a tempting idea, but the traditional usability heuristics cannot be applied directly.

The design objectives between productivity software and games are different, and the evaluation methods need to recognize this divergence as well before they can be effectively applied in games domain. Pagulayan et al. describe these differences, and according to them, productivity software is a tool and the design intention is to make tasks easier, more efficient, less error-prone, and increase the quality of the results. Games, instead, are intended to be pleasurable to play and sufficiently challenging [21]. Because of these differences, a set of specifically designed heuristics are needed when videogames are evaluated with the expert review method.

Playability has been studied very little by game researchers and HCI researchers. The research community is lacking a commonly agreed definition for playability, which would describe important issues influencing the game experience and guiding the research work. Egenfield-Nielsen et al. state that the game has good playability when it is easy to use, fun and challenging [3]. Järvinen et al. have defined playability as an evaluation tool, which consists of four components: 1) functional, 2) structural, 3) audiovisual, and 4) social playability [12]. These components can be used to evaluate both the formal and informal aspects of the game. Fabricatore et al. have defined playability in action games as the possibility of understanding and controlling the gameplay. In addition, they state that poor playability cannot be balanced or replaced with non-functional aspects of the design [5]. According to the usability glossary¹, playability is affected by the quality of different aspects, including storyline, controls, pace, and usability.

Playability is related to intuitiveness, unobtrusiveness, fun, and challenge. In addition, it is a combination of user interface and the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MindTrek 2009, September 30th-October 2nd 2009, Tampere, FINLAND.
Copyright 2009 ACM 978-1-60558-633-5/09/09...\$10.00.

¹ http://www.usabilityfirst.com/glossary/term_657.txt

gameplay, i.e. game content aspects of the game. In multiplayer games, players' social interaction will also affect playability. The user interface consists of game menus, controls and an interface through which a player interacts with game items, non-player characters (NPCs), and other players. The game has good playability when the user interface is intuitive and unobtrusive, so that the player can concentrate on playing the game. Gameplay includes, for example, game mechanics, narrative, and goals that the player tries to achieve. Fun and challenge are created by the gameplay; the game has good playability especially when the gameplay is understandable, balanced, suitably difficult, and engaging.

Despite the lack of a commonly agreed definition, researchers have defined playability heuristic sets that could be used to evaluate videogames and their playability. However, the development work is still ongoing and there is very little knowledge about the usefulness and clarity of these heuristic sets. In addition, there are no previously published studies that would use these heuristic sets to evaluate a videogame and compare the results.

In this paper, we report a first experiment in which two playability heuristic sets are used in a videogame evaluation to discover their weaknesses and strengths in identifying playability problems, as well as whether they are helpful to the evaluators in conducting the evaluation. The results indicate that heuristic sets should be improved before they are usable for the practitioners.

The rest of this paper is structured as follows. First, we review the relevant related work regarding the expert review method and introduce playability heuristics that have been developed. Next, we describe an experiment we arranged to compare two playability heuristic sets in a game evaluation and report the results of the experiment, followed by discussion and conclusions.

2. RELATED WORK

In this section, we review the expert review method and look at playability heuristics that have been developed for evaluating videogames.

2.1 The Expert Review Method

Inspection methods are well known and widely used to evaluate the usability of a product. This is due to their effectiveness and cost-efficiency in discovering usability problems. The method probably used most often is heuristic evaluation, developed by Nielsen and Molich [20]. The method is also known as an expert review method, since the evaluators' experience and knowledge will affect the evaluation results [10]. The first version of the usability heuristics was published together with the method, but the revised and currently used version of the heuristics was published in 1994 [19].

Several researchers have extended the original usability heuristics or developed new ones for different application domains. Ling and Salvendy present a summary of these studies [15]. The list contains domains such as websites, e-learning systems, groupware, notification systems, and games.

The applicability of usability heuristics in game evaluations has been questioned by game researchers [2], [7], [14]. The most important reason for this is that usability heuristics concentrate primarily on the user interface of the product and disregard the gameplay. In their study, Johnson and Wiles show how games

contravene the traditional usability heuristics to achieve a good game experience [11]. Hence, game researchers have started to develop heuristics which would include both usability and gameplay issues, to assist game developers in discovering playability problems in the game design.

2.2 Development of Playability Heuristics

In the early 80's, Malone studied videogames and what makes the user interface enjoyable [16]. He identified three principles (challenge, fantasy, and curiosity) that are needed for designing enjoyable user interfaces. Malone also calls these principles heuristics in a design framework. Although the list is very limited and it concentrates only on high level issues in games, it highlights the importance of the game content in the evaluation.

In 2002, Federoff defined the first playability heuristics that are similar to usability heuristics. These heuristics were a result of a case study in a game company [7], but they lack validation, or at least such results have not been published. Fabricatore et al. have studied players and their preferences that will affect the playability of action videogames [5]. Even though these are not described as heuristics, they could be converted into heuristics to evaluate games belonging to this specific genre.

Desurvire et al. published Heuristic Evaluation for Playability (HEP) in 2004 [2]. The heuristics were validated in a study where the heuristic evaluation was compared to the user testing. The results indicated that the heuristics were very good at identifying playability problems from a game prototype.

Korhonen and Koivisto published a playability heuristic set for mobile games in 2006 [14]. However, the heuristics are applicable for evaluating games in other platforms as well because of their modular structure. The playability heuristic set can be extended or limited based on the needs of the evaluation. In addition, the number of the heuristics is smaller than in two previous sets [2], [7]. The heuristics have been validated in several game evaluations.

More recently, Pinelle et al. published game usability heuristics that are based on game reviews [23] and they have been validated in a preliminary study. These heuristics are used to evaluate game usability (user interface) and there are no heuristics concerning gameplay issues.

There are also other guidelines that are targeted for game developers in order to make games more engaging and usable for the players [6], [24].

Based on the literature review, the expert review could be an appropriate method for evaluating the playability of videogames, but there should be specific playability heuristics accompanying the method. Several researchers have started to develop these heuristics, and currently, there are multiple heuristic sets available. However, the work is still ongoing and the heuristic sets are quite different, even though there are some common issues included. This raises the question of which heuristic set should be used in a game evaluation, and if one heuristic set is easier to use than another one from the evaluators' point of view.

In our work, we aim to achieve some clarity to the different playability heuristic sets and their usefulness in game evaluations. We compare two playability heuristic sets in a game evaluation to see what their strengths and weakness are and how the evaluators perceive the heuristic sets.

3. EXPERIMENT

We arranged an evaluation session with 8 persons, who are working in the game industry or in academia as game researchers, to explore how the expert review method and two playability heuristic sets operate in a game evaluation. At first, the participants were briefly introduced to the expert review method and the heuristics that are commonly used in productivity software evaluations to give an idea of how the usability specialists usually conduct the evaluation. Three participants had previous experience how to conduct an expert review of a product.

The participants were divided into 4 teams (two persons in each team) forming two groups based on the playability heuristic sets that were given to them. In the evaluation session, the teams played the game for one hour. The evaluators observed the game and wrote down short descriptions for possible playability problems they encountered in the game. After that, the teams went through their own playability problems and assigned violated playability heuristics to these problems. Finally, the observations were walked through with other teams and the participants discussed playability problems, the evaluation method, and the playability heuristics they used. The results section describes the main observations from the discussion.

The evaluated game was EA Mobile's *The Simpsons: Minutes to Meltdown*². It is a rather short mobile game, which can be finished in less than 30 minutes (in real time). The player plays as Homer and tries to save Springfield from nuclear disaster.

3.1 Playability Heuristics

As there are multiple heuristic sets available, it is important to choose heuristic sets that are feasible to compare. As described in the chapter on related work, some heuristics are proposals which have not been validated, others are targeted to specific game genre or they do not consider all aspects of the playability. For this study, we selected playability heuristic sets from Desurvire et al. [2] and Korhonen and Koivisto [14] because they resemble each other and they are probably the most advanced at the moment. Both heuristic sets are based on literature reviews and the initial heuristics were reviewed by game researchers and game designers. The playability heuristics were developed further in game development projects. Although the sets have some similarities in their content, there are major differences in how the heuristics are organized and described.

3.1.1 Heuristic Evaluation for Playability (HEP)

This playability heuristic set contains 43 heuristics and the researchers have defined four categories for organizing them [2]. *Game Play* is related to challenges and problems that the player must face to win a game. *Game Story* includes heuristics for story and character development. *Game Mechanics* involves the structure, which defines how the game units interact with the environment. *Game Usability* addresses the interface and the controls the player utilizes when interacting with the game. Most heuristics are presented as one sentence descriptions and they have been validated in a user study.

Teams that used this playability heuristic set during the evaluation are referred to as Violet 1 and Violet 2 in the results section.

3.1.2 Playability Heuristics for Mobile Games

This playability heuristic set contains 29 heuristics and they have been organized into three modules [14]. Each module can be included or excluded depending on the needs of the evaluation. Two core modules, *Gameplay* and *Game Usability*, are common for all games. The *Mobility* module contains heuristics that are specific for mobile games. Each heuristic is described in detail on a separate document including examples of use [13]. The heuristics are validated in several mobile game evaluations conducted by playability experts.

Teams that used this playability heuristic set during the evaluation are referred to as Orange 1 and Orange 2 in the results section.

4. RESULTS

In this section, we present the main results of the study, which are based on the comments from the group interview as well as the analysis of the data collected from the evaluation reports.

4.1 Heuristics Provide Guidance

The participants commented that the expert review method seemed to be an appropriate method for evaluating videogames, because it helped them to focus on the different aspects of the game during the evaluation. Participants commented that heuristics could also be useful in the design and implementation phases to identify possible playability problems that might exist in the design.

However, the evaluators should always be aware of the creative vision that the designers have and what is used as a design principle when designing a game. Typically, it also guides the experience that the designers want to create for the players [22]. The evaluation should always be relative to this vision, because otherwise the evaluators might focus the evaluation incorrectly and point out issues which are contradicting to the vision.

4.2 Defining a Proper Abstraction Level

Although the participants appreciated the efficiency of the expert review method, they stated that there are certain challenges when the method is applied to game evaluations. Their biggest concern was related to the heuristics and their descriptions. The variety of videogames is enormous and defining playability heuristics that are suitable for evaluating all kinds of games can be a challenge.

"It is a laborious and challenging task to define heuristics that can capture those aspects that are considered to be essential from the game experience point of view. In addition, game environments are changing constantly as they adopt new kinds of technical enablers", Violet 1 evaluator³.

Therefore, it is important that the playability heuristics are on the right abstraction level. Too specific heuristics restrict their applicability to a large number of games, but in contrast, the heuristics that are on very general level lose their power to guide and assist the evaluators during the evaluation. The participants stated that both heuristic sets had problems in this respect.

² <http://www.eamobile.com/Web/mobile-games/the-simpsons-minutes-to-meltdown>

³ Translated from Finnish by the authors

Playability heuristics defined by Desurvire et al. had both detailed heuristics and very broad heuristics, which were difficult to use during the evaluation. For example, there is the Game Play heuristic number 10 (*“The game is fun for the Player first, the designer second and the computer third. That is, if the non-expert player’s experience isn’t put first, excellent game mechanics and graphics programming triumphs are meaningless.”*) This heuristic was considered to be very difficult to apply during an evaluation.

Playability heuristics defined by Korhonen and Koivisto also had some heuristics which were considered to be quite specific, and they could be combined to provide a more concise list. For example, heuristics GP9 (*“The players can express themselves”*) and GP10 (*“The game supports different playing styles”*) are describing similar kinds of issues related to the player’s behavior and playing style in the game world.

4.3 Evaluation Process

The participants commented that the evaluation task influenced their gaming experience, and for that reason, playing the game was different than what it would be normally. The objective of the game design is to immerse players on different levels [4]. The evaluation task, however, prevented the immersion because the evaluators need to be alert all the time and inspect the game for problems in playability. In addition, the evaluators found it difficult to play as any player would play the game, and for that reason, the evaluation session cannot be considered equal to a normal play session.

“There are two dimensions that make the evaluation difficult. First, you should be able to describe the problem that you have identified and it affects your gaming experience negatively. On the other hand, you should play the game as players would play and get a positive gaming experience”, Violet 1 evaluator.

Another issue the participants pointed out was that it is very important for the evaluators to familiarize themselves with the heuristics beforehand. In our study, playability heuristics sets contained 43 or 29 heuristics. When considering Miller’s golden rule of 7 ± 2 [17], the number of heuristics might have been overwhelming and there was too much information about the heuristics to keep in mind. During the evaluation it was time-consuming to browse the whole list through and find a proper heuristic for each playability problem. Due to time constraints, the participants did not study the heuristics beforehand, but there was a playability expert present, if they had any questions concerning the heuristics.

4.4 Revision for Playability Heuristic Sets

The participants found several issues troublesome with the playability heuristics defined by Desurvire et al. These issues made utilization of the heuristics difficult during the evaluation. There are a total of 43 heuristics on the set and the participants thought it is too much. The heuristics are organized into four categories, but the participants did not find them helpful because some heuristics were in a different category from what they expected. For example, some Game Story heuristics were located in the Game Play category and vice versa. The Violet team evaluators pointed out that Game Play heuristic number 8 (*“Players discover the story as part of game play”*) would belong to the Game Story category rather than the Game Play category, and that Game Story heuristic number 6 (*“Player experiences fairness of outcomes”*) sounds more like a heuristic belonging to

the Game Play category. There were also some overlapping heuristics in the set.

Another problem that the participants noticed was the descriptions of the heuristics, as they were presumably influenced by the game that was used as a basis during the development work. Some heuristics were seen as too specific to apply in practice. In addition, the descriptions were not consistent in terms of wording and level of generalization. Some heuristics clearly set requirements for the game design and state explicitly how the game design should be done. The example of this kind of heuristic is Game Play heuristic number 3 (*“Provide clear goals, present overriding goal clearly as well as short-term goals throughout the play”*), whereas some heuristics are more like recommendations for designers. For example, Game Play heuristic number 5 (*“The game is enjoyable to replay”*) is a too general and subjective issue to evaluate with the expert-based method. There were also some heuristics which were difficult to understand and apply during the evaluation. The participants pointed out Game Play heuristic number 10 (*“The game is fun for the Player first, the designer second and the computer third. That is, if the non-expert player’s experience isn’t put first, excellent game mechanics and graphics programming triumphs are meaningless”*) to be an example of such a heuristic. Finally, the participants commented that the current writing style and format makes understanding of the heuristics more difficult, because they are not consistent and are missing a heading or description.

Playability heuristics developed by Korhonen and Koivisto were not optimal either. Even though each heuristic clearly had the heading and the description, they were presented in two documents which made using them difficult. The first document described the heuristics on a heading level, similar to the other heuristic set. There was a separate document available that contained the descriptions and practical examples. Some descriptions were also long, and reading the entire description and examples was time-consuming. The participants commented that this playability heuristic set was in a better shape and the wording of the heuristics was more consistent and on a more generic level than on the other heuristic set. However, there were still some heuristics such as GP8 (*“There are no repetitive or boring tasks”*) and GP11 (*“The game does not stagnate”*) that sounded similar on the heading level and they could be possibly combined.

4.5 Evaluation Statistics

Surprisingly, there was very little consistency in reporting playability problems between the four teams. Only a few playability problems were identified by more than one team. Even if the teams identified the same playability problem, they assigned a different playability heuristic to describe the problem. This also happened within the two teams which used the same playability heuristic set. The teams reported 69 playability problems in total. 13 playability problems were reported by two or more teams and 52 playability problems were unique reported by a single team. In addition, there were 13 duplicate playability problems (i.e. reported multiple times by a single team), but these problems have been excluded from the analysis. There was a difference between groups in how many playability problems they reported. Teams Orange 1 (O1) and Orange 2 (O2) identified 13 and 12 playability problems respectively. Teams Violet 1 (V1) and Violet 2 (V2) identified substantially larger number of playability problems, 20 and 24 playability problems respectively.

Table 1 playability problems concerning different heuristic categories.

| Teams | Orange | | | | Violet | | | |
|----------------|--------|----|-------|------|--------|----|-------|------|
| | O1 | O2 | Total | % | V1 | V2 | Total | % |
| Game Usability | 3 | 4 | 7 | 28% | 6 | 3 | 9 | 20% |
| Gameplay | 8 | 5 | 13 | 52% | 3 | 11 | 14 | 32% |
| Mobility | 1 | 0 | 1 | 4% | - | - | - | - |
| Game Story | - | - | - | - | 1 | 2 | 3 | 7% |
| Game Mechanics | - | - | - | - | 5 | 0 | 5 | 11% |
| Unassigned | 1 | 3 | 4 | 16% | 5 | 8 | 13 | 30% |
| Total | 13 | 12 | 25 | 100% | 20 | 24 | 44 | 100% |

The most problems reported by both teams were related to gameplay issues. Teams O1 and O2 reported more than a half of the problems belonging to this category. The second most common problem category was game usability. Playability problem distribution in the heuristic categories⁴ is illustrated in Table 1. Some user interface problems were due to the mobile phones the participants used. The game looked and sounded different on their devices, and there were some minor changes in the game content because of the smaller screen resolution and the memory capacity of the device.

The teams seemed to have difficulties in assigning violated heuristics to the identified playability problems, and the participants commented that they could not always find a proper playability heuristic from the set. Especially for teams V1 and V2, assigning a violated playability heuristic was difficult, and they left 30% of the playability problems open (Table 1). Teams O1 and O2 were able to do it more accurately, and they left only 16% of reported playability problems open.

Usually the teams assigned only one violated heuristic per problem, but there were a few cases when they assigned several heuristics (Table 2). The teams reported nine playability problems to which they assigned several heuristics from the same category that the problem violated. Three of them were related to Game Usability and the rest were Gameplay problems. There were also three playability problems to which the teams assigned playability problems from different categories. These problems were combinations of Gameplay, Game Usability, and Game Story related issues.

Table 2 Assigning heuristics to playability problems

| Heuristics | Orange Teams | | Violet Teams | |
|-------------------------|--------------|------|--------------|------|
| | Count | % | Count | % |
| Single | 12 | 48% | 28 | 64% |
| Many Same Category | 8 | 32% | 1 | 2% |
| Many Different Category | 1 | 4% | 2 | 5% |
| Unassigned | 4 | 16% | 13 | 30% |
| Total | 25 | 100% | 44 | 100% |

⁴ It should be noted that the heuristic categories are not comparable because they contain different heuristics. In addition, some categories exist only in one playability heuristic set and in the other set those categories are left empty on the table.

Finding the same playability problems seemed to be difficult, and the majority of the playability problems (75%) are reported only by a single team (Figure 1). However, there was one playability problem which all teams reported. The playability problem concerns player progression in the game. If Homer dies in the game, the player has to start over. Teams were also consistent when assigning the violated heuristic for this problem. Teams O1 and O2 assigned the gameplay heuristic GP14 (“*The player does not lose any hard-won possessions*”). In addition, Team O1 said that the problem violated Gameplay heuristic GP8 (“*There are no repetitive or boring tasks*”). Teams V1 and V2 also had a consensus on the violated heuristic. They assigned Game Story heuristic GS6 (“*Player experiences fairness of outcomes*”) to describe the problem. In addition, team V1 assigned Game Play heuristic GP5 (“*The game is enjoyable to replay*”).

Playability Problems Reported by Teams

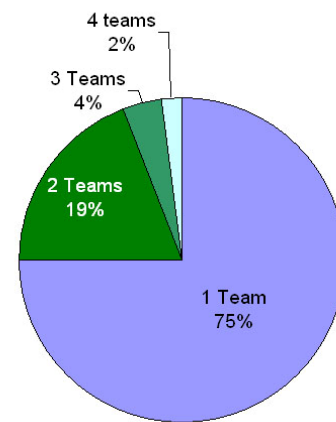


Figure 1 Playability problems reported by teams

There were two problems that were identified by three teams. The first problem concerned navigation in the game world, due to fact that the player gets lost really easily on the second level. The second problem was related to the game menu design. Even though the three teams identified the same problem, each team assigned a different heuristic to describe the problem, or left the problem open. For the playability problems identified by two teams, there was hardly any consistency in the assigned

heuristics. One explanation for different evaluation results between teams O1 and O2 might be that team O1 reported very specific playability problems such as “*catching the pig is hard and it is not clear how it should be done*”, whereas team O2 reported more general level problems like “*the game is too linear and prone to stagnate*” or “*game features boring repetition without optional ways to advance*”.

Similarly, teams V1 and V2 used quite different heuristics to describe the playability problems. Team V2 did not assign any playability problems into the Mechanics category, even though team V1 used Mechanics quite extensively. They found three playability problems that violated heuristic ME1 (“*game should react in a consistent, challenging, and exciting way to the player’s actions (e.g., appropriate music with the action)*”). In addition, they assigned two other heuristics from the Mechanics category to describe identified problems. Correspondingly, team V2 concluded that five playability problems violated Game Play heuristic GP2 (“*provide consistency between the game elements and the overarching setting and story to suspend disbelief*”), while team V1 thought that none of their playability problems violated this heuristic.

Both teams reported playability problems with different abstraction levels. Team V1 identified both specific and general level problems, whereas team V2 concentrated on criticizing the illogical gameplay. Examples of such playability problems in the gameplay were “*Barney opens up a gate when you bring him coffee*” and “*The player can only go through certain bushes*”.

5. DISCUSSION

The evaluators’ comments indicate that the expert review method is applicable to game evaluations. The evaluators liked the method as it is not too time-consuming or laborious to execute. They thought that the method could also be used at earlier development phases, when there are only design sketches or low fidelity prototypes available. Playability heuristics, however, need to be developed further before the method can be widely adopted by the practitioners. The playability heuristics should be presented in a similar manner to how Nielsen has presented traditional playability heuristics [19]. Ling and Salvendy have also concluded that domain specific heuristic sets should be structured and they should not contain too many heuristics [15].

In this study, we used playability heuristics sets developed by Desurvire et al. [2] and by Korhonen and Koivisto [14]. The study revealed that both heuristic sets need to be improved in order for them to be usable and easily understandable. The evaluators considered that there were too many heuristics on the set developed by Desurvire et al. In addition, their organization into categories as well as their descriptions needs to be developed further as they were inconsistent and overlapping. This was visible in the evaluation data, as the teams who used this heuristic set did not assign any violated heuristic to 30% of the identified playability problems. The playability heuristic set developed by Korhonen and Koivisto was more consistent in wording and organization, but the evaluators thought that the heuristics should be accompanied by a short and compact description, since currently, the descriptions were presented in a separate document.

In the study design, it is important to think about the hardware that will be used, since it can have remarkable influence on what kinds of playability problems are reported. Especially mobile

phones can be very different in their technical capabilities and there are many device generations on the market. We did not anticipate that the game would vary so much on different devices. In our study, the evaluators used their personal mobile phones in the evaluation and therefore, we did not have sufficient control over the hardware. Some teams reported playability problems which were somewhat specific for the device they used. These problems were related to audio and the amount of the content on the screen. Gray and Salzman call this as an internal validity problem [9].

In game evaluations, the evaluators seem to face similar challenges in identifying the same playability problems. This result is consistent with comparison studies conducted with productivity software. However, the results from this study are slightly better than those reported by Molich and Dumas [18]. The majority of the playability problems (75%) were reported only by a single team. However, one playability problem was commonly reported by all four teams, and the violated heuristic was assigned consistently within the teams as well. Furthermore, there were 12 playability problems which were reported by at least two teams. It is an interesting question for future work why the evaluators do not identify same problems from the game. Unlike productivity software, videogames in general are quite linear at the beginning and the players are guided through first missions or levels by the game design [1]. Therefore, the evaluators should have gone through same aspects of the game and presumably identify the same problems. The problem that all teams identified in this study was critical for the game experience, and this is probably the reason why it was reported.

There are several possible reasons for the inconsistency of the reported problems. One obvious explanation is the evaluator effect [10] and its influence to the results. It has been concluded in many previous evaluation studies of the productivity software that the evaluation results differ quite a lot because of this factor (e.g. [10], [18]). In this study, our evaluators had different backgrounds, game design, and evaluation experience. Although we tried to balance teams in their evaluation experience and game design experience, it did not seem to be enough.

Another possible explanation for the inconsistency might be the heuristic sets that the evaluators used in the study. The purpose of the heuristic sets is to guide the evaluation and remind the evaluators to pay attention to important aspects of the playability. The results indicate that using the heuristic sets was not straightforward and the evaluators had some problems with them, which might also explain the difference in reported playability problems. However, one interesting observation from the data is that most of the playability problems that were reported by two or three teams included teams from both groups. There were only few problems which were reported only by one group. Unfortunately, there is not sufficient data from this study to make any deeper analysis how a playability heuristic set influences in finding playability problems from a videogame.

Third possible explanation for this inconsistency might be that the evaluators had a different baseline for reporting. Some teams reported mainly general problems, focusing on certain aspects of the game, while the others reported very specific problems. In the Violet group, team V2 did not report any playability problems which would violate heuristics from Mechanics category, whereas team V1 assigned five problems to this category.

Correspondingly, team V1 assigned five playability problems to one Game Play heuristic which was not used by team V2 at all. Otherwise, the teams reported problems that violated a large number of the playability heuristics from different categories. This difference is probably due to evaluation experience that the teams had. In addition, we probably did not instruct the teams clearly enough on what kinds of issues they should pay attention to and how to report those findings.

One characteristic of the game evaluations is to think about the origin of the playability problem, and whether the problem is in the user interface or in the game content. This problem does not usually exist in productivity software evaluations, as the evaluation concerns only user interface aspects of the product. Evaluating the content and the user interface together has been studied on other domains [e.g. [8]]. In our study, the evaluators identified 12 playability problems to which they assigned multiple heuristics, and in three cases they were from different categories. We do not know for sure why the evaluators did it this way. Possibly, they did not have time to analyze the problem thoroughly to find the origin of the problem.

In the future, we are planning to continue these comparison studies to find out the optimal set of playability heuristics. The shortcoming of this study was that we could not compare which playability heuristics are used to describe same playability problems because there too was little data for this. In the next study, we should also eliminate internal validity errors, which were related to the evaluators' experience in using the evaluation method, their familiarity with the playability heuristics, and the devices that the evaluators used in the evaluation. In addition, there is a need to think about a new presentation format for the heuristics, which would better support the evaluation. In the discussion it became obvious that presenting heuristics as a list is not easily utilized during the evaluation. The heuristics could be improved by using keywords, color coding for the categories and presenting them in a compact format.

6. CONCLUSION

In this paper, we have explored two different playability heuristic sets to discover their strengths and weaknesses when they are used to evaluate a mobile game using the expert review method. This kind of comparison study has not been reported previously, although there are several playability heuristic sets available. The results indicate that both heuristic sets should be improved, as there were problems in clarity and comprehensibility. This study is the first attempt to develop playability heuristics that would help the evaluators to conduct the game evaluations, and to provide precise and relevant evaluations results when evaluating videogames with an analytical evaluation method.

7. ACKNOWLEDGMENTS

GameSpace project was funded by Tekes (Finnish Funding Agency for Technology and Innovation), Veikkaus, TeliaSonera Finland, Nokia, Sulake Corporation, and Digital Chocolate. We thank all project members and the evaluators in this study.

8. REFERENCES

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

- [2] Desurvire, H., Caplan, M., Toth, J.A. 2004. Using heuristics to evaluate the playability of games. In proceedings of ACM SIGCHI '04 extended abstracts, 1509-1512. DOI=<http://doi.acm.org/10.1145/985921.986102>
- [3] Egenfield-Nielsen, S., Smith, J.H. and Tosca, S.P. 2008. *Understanding Video Games: The Essential Introduction*. Routledge.
- [4] Ermi, L., Mäyrä, F. 2005. Fundamental Components of the Gameplay Experience: Analysing Immersion. In proceedings of DIGRA 2005 Conference: Changing Views - Worlds in Play.
- [5] Fabricatore, C., Nussbaum, M., Rosas, R. 2002. Playability in Action Videogames: A Qualitative Design Model. *Human-Computer Interaction* 17, 311-368. DOI=http://dx.doi.org/10.1207/S15327051HCI1704_1
- [6] Falstein, N. and Barwood, H. The 400 Project, http://theinspiracy.com/400_project.htm.
- [7] Federoff, M.A. 2002. *Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games*. Department of Telecommunications, Master of Science. Indiana University, Indiana.
- [8] Galagher K., Foster D., Parsons J. 2001. The Medium Is Not the Message: Advertising Effectiveness and Content Evaluation in Print and on the Web, *Journal of Advertising Research* 41(4), 57-70.
- [9] Gray, W.D., Salzman, M.C. 1998. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction* 13 (3), 203-261. DOI= http://dx.doi.org/10.1207/s15327051hci1303_2
- [10] Jacobsen N.E., Hertzum M., John B.E. 1998. The evaluator Effect in Usability Tests, In *Proceedings of ACM SIGCHI*, 255-256. DOI= <http://doi.acm.org/10.1145/286498.286737>
- [11] Johnson D., Wiles J. 2003. Effective Affective User Interface Design in Games, *Ergonomics* 46 (13/14), 1332-1345. DOI= <http://dx.doi.org/10.1080/00140130310001610865>
- [12] Järvinen, A., Heliö, S., Mäyrä, F. 2002. *Communication and Community in Digital Entertainment Services*. University of Tampere.
- [13] Koivisto E.M.I, Korhonen H. 2006. *Mobile Game Playability Heuristics*, www.forum.nokia.com.
- [14] Korhonen, H., Koivisto, E.M.I. 2006. Playability Heuristics for Mobile Games. In proceedings of *MobileHCI'06*, 9-16. DOI= <http://doi.acm.org/10.1145/1152215.1152218>
- [15] Ling, C., Salvendy, G. 2005. Extension of Heuristic Evaluation Method: a Review and Reappraisal. *Ergonomia*. An International Journal of Ergonomics and Human Factors 27 (3). 179-197.
- [16] Malone, T.W. 1982. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In proceedings of ACM SIGCHI, 63-68. DOI= <http://doi.acm.org/10.1145/800049.801756>
- [17] Miller, G. A. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63. 81-97.

- [18] Molich, R., Dumas, J.S. 2008. Comparative Usability Evaluation (CUE-4). *Behaviour & Information Technology* 27, 263-281.
- [19] Nielsen J. 1993. *Usability Engineering*, Academic Press.
- [20] Nielsen, J., Molich, R. 1990. Heuristic evaluation of user interfaces. In *proceedings of ACM SIGCHI*, 249-256. DOI=<http://doi.acm.org/10.1145/97243.97281>
- [21] Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L. and Fuller, T. 2008. User-centered Design in Games. In Jacko, J. Sears A., (Eds.). *Handbook for Human-Computer Interaction in Interactive Systems*, Second Edition, Lawrence Erlbaum Associates, Inc., 741-759.
- [22] Pagulayan R., Steury K. 2004. Beyond Usability in Games, *Interactions* (11), 5, 70-71. DOI=<http://doi.acm.org/10.1145/1015530.1015566>
- [23] Pinelle, S., Wong, N., Stach, T. 2008. Heuristic evaluation for games: usability principles for video game design. In *proceedings of ACM SIGCHI*, 1453-1462. DOI=<http://doi.acm.org/10.1145/1357054.1357282>
- [24] Snow, B. 2007. Game Usability 101. *BusinessWeek*, http://www.businessweek.com/innovate/content/oct2007/id20071012_041625.htm.