JUS
JOURNAL OF USABILITY STUDIES

Issue 2, Vol. 1, February 2006, pp. 64-75

Do usability expert evaluation and test provide novel and useful data for game development?

Sauli Laitinen
Adage Corporation
Hämeentie 153 C
00560 Helsinki, Finland
+358 50 5343164
sauli.laitinen@adage.fi

Abstract

A case study was done to study whether usability expert evaluation and testing are suitable for game development. In the study, a computer game under development was first evaluated and then tested. Game developers were then asked to rate the findings and give other feedback about the methods used and the results gained.

It was found that the usability expert evaluation and testing provided both novel and useful data for game development. Based on these and the other results it is argued that the usability expert evaluation and testing have considerable face validity in game development.

In addition to the usefulness and face validity of the methods it was studied whether the usability experts participating in the game usability expert evaluation should be double experts. It was found that there was no significant difference in the number or the rated relevancy of the problem the gamer and non-gamer usability specialists found.

Keywords

Usability method, validity, expert evaluation, usability test, computer games, video games, games

Introduction

Both computer games and usability have long histories, but only recently have these two been combined. The first reported steps to introduce usability evaluation methods to game development were taken in 1997 at Microsoft (Fulton & Romero, 2004). Since then, other companies have also adopted usability evaluation methods to their game development, but not all have been convinced (H. Desurvire, personal communication, June 29, 2005).

One likely reason for rejecting the usability methods is that the game developers, game producers and marketing departments may be doubtful or ignorant about the usefulness of the usability evaluation methods. In this study it was investigated whether there is a good reason for this or not. It was studied whether usability expert evaluation and usability testing provide data that game developers find novel and useful.

Why does game developers' perception matter?

The game developers' perception of the usability expert evaluation and testing is important for very practical reasons. If game developers think that usability expert evaluation and testing do not provide new and useful data they will not use the methods. The same applies to the problems found with the methods. If the game developers do not find the problems plausible they will choose not to fix them.

The game developers’ perception of the usability expert evaluation and testing is also important because the game producers and marketing departments can use this to make up their minds about whether to use the methods or not.

Face validity

The game developers’ perception of the usability expert evaluation and testing is also interesting because it can be used as a measure of the face validity of the methods. Face validity is a crude measure and does not tell everything about the validity of the usability expert evaluation and testing in game development. In the current situation, however, face validity is an interesting issue and well worth studying. This is because so far only two studies have been published on the validity of the usability expert evaluation and testing in game development. These studies have been done by Desurvire, Caplan and Toth (2004) and Medlock, Wixon, Terrano, Romero and Fulton (2002).

Desurvire, et al. (2004) studied the validity of expert evaluation. In the expert evaluation the experts evaluated a game using a list of heuristics specifically developed for the evaluation of video, computer and board games. They studied the validity by comparing the results of expert evaluation to the results gained through a usability test. They found that the expert evaluation was “…very useful for creating highly usable and playable game design, particularly in the preliminary design phase prior to expensive prototypes” (Desurvire, et al., 2004, p. 4). Yet they concluded that the expert evaluation does not replace usability testing as there is no way of knowing how people really behave.

Medlock, et al. (2002) studied the validity of the rapid iterative testing and evaluation method. The measures they used included a web diary of the lead game designer, game reviews, the awards the game claimed and the sales figures. Based on this data it was concluded that the method was highly effective in terms of finding and fixing problems and resulted in positive industry reviews for the part of the game the method was applied to.

The two studies mentioned above support the view that usability expert evaluation and testing are valid methods for game development. Two case studies, however, are not enough to establish the validity and more studies are needed (John, 1998; Medlock et al., 2002). This study was performed to meet this need.

Should the usability experts be also gamers?
An additional goal of this study was to find out whether all of the usability specialists participating in an expert evaluation of a game need to be double experts. A double expert was defined as a specialist who is both a usability specialist and a gamer. This is an interesting issue, because it has been argued that the usability specialists participating in game development should be both gamers and usability specialists (Fulton & Romero, 2004). This question has also practical importance because finding double experts to do the usability expert evaluations is not always easy. The temptation to have non-gamer usability specialists to participate in the evaluation team is often strong.

In this study the difference between the game and non-gamer usability specialists was studied by comparing the number and quality of the problems they found in the usability expert evaluation.

Method and process

Design
A case study was conducted to study the aforementioned issues. A computer game under development was first evaluated and then tested. After this, the game developers answered a web survey where they rated each usability problem found on a multidimensional scale. They also gave general feedback about the process.

Participants
The usability expert evaluation was conducted by six usability specialists. Four of them were classified as gamers and two as non-gamers. The usability specialists classified as gamers reported that they played games weekly. One of them played only on computers and one only on game consoles, the remaining two played on both computers and consoles. The two usability experts classified as non-gamers reported that they did not play games at all. The usability specialist who led both the usability expert evaluation and test was classified as a gamer.

The most experienced usability specialist who participated in the expert evaluation had been working as a usability specialist for 8 years and the least experienced for 3 years. The average experience of the usability specialists was 4 years and 4 months. The evaluation leader was the only specialist who had considerable previous experience in the game user research.

The usability test was conducted by one usability specialist and an assistant. In the usability test there were six test users. They represented the three target groups of the game.

The two game developers who answered the web survey were the lead designer and the project manager. They were also the contact persons on the game developers' side, and the results of both the usability expert evaluation and test were reported to them.

Materials

The game evaluated and tested was Frozenbyte's computer game called Shadowgrounds. Shadowgrounds is an action game viewed from the top-down perspective. See Figure 1 for an illustration.

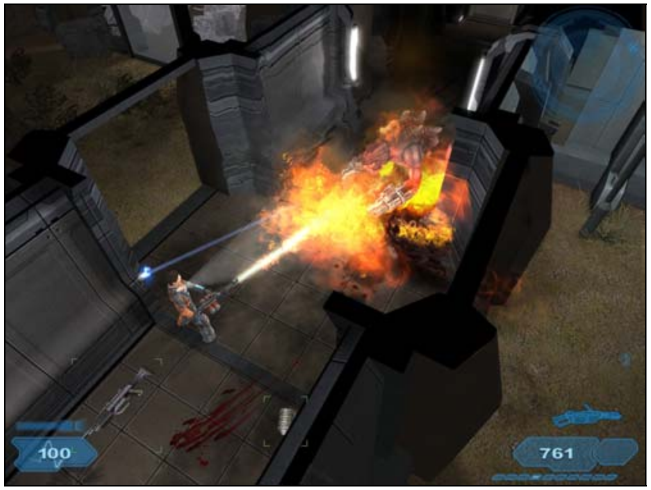


Figure 1. Shadowgrounds is an action game viewed from a top-down perspective.

The expert evaluation was conducted approximately six months before the planned launching of the game. In the version evaluated, there was one playable level and the basic gameplay mechanics had been implemented. However, not everything was completed. For example, the destructible environment was not fully implemented, voice acting was missing, and several smaller bugs were present.

The usability test was conducted in a standard usability laboratory. The developers had the opportunity to observe the tests in a separate room (see Nielsen, 1993, for an example).

In the web questionnaire, seven questions addressed each problem found in the usability expert evaluation and testing. The first two questions measured the novelty of the problem and its description. The following three questions were about the relevancy of the problem, accuracy of the severity classification and usefulness of the suggested solution. The remaining two questions probed whether corrective actions were to be taken to fix the problem and was the problem due to a programming error. The questions and the scale are illustrated in the Figure 2.

Additional seven open-ended questions were asked after the problem specific questions. The questions addressed the pros and cons of the methods and how to improve them.

Procedure – Initial meeting

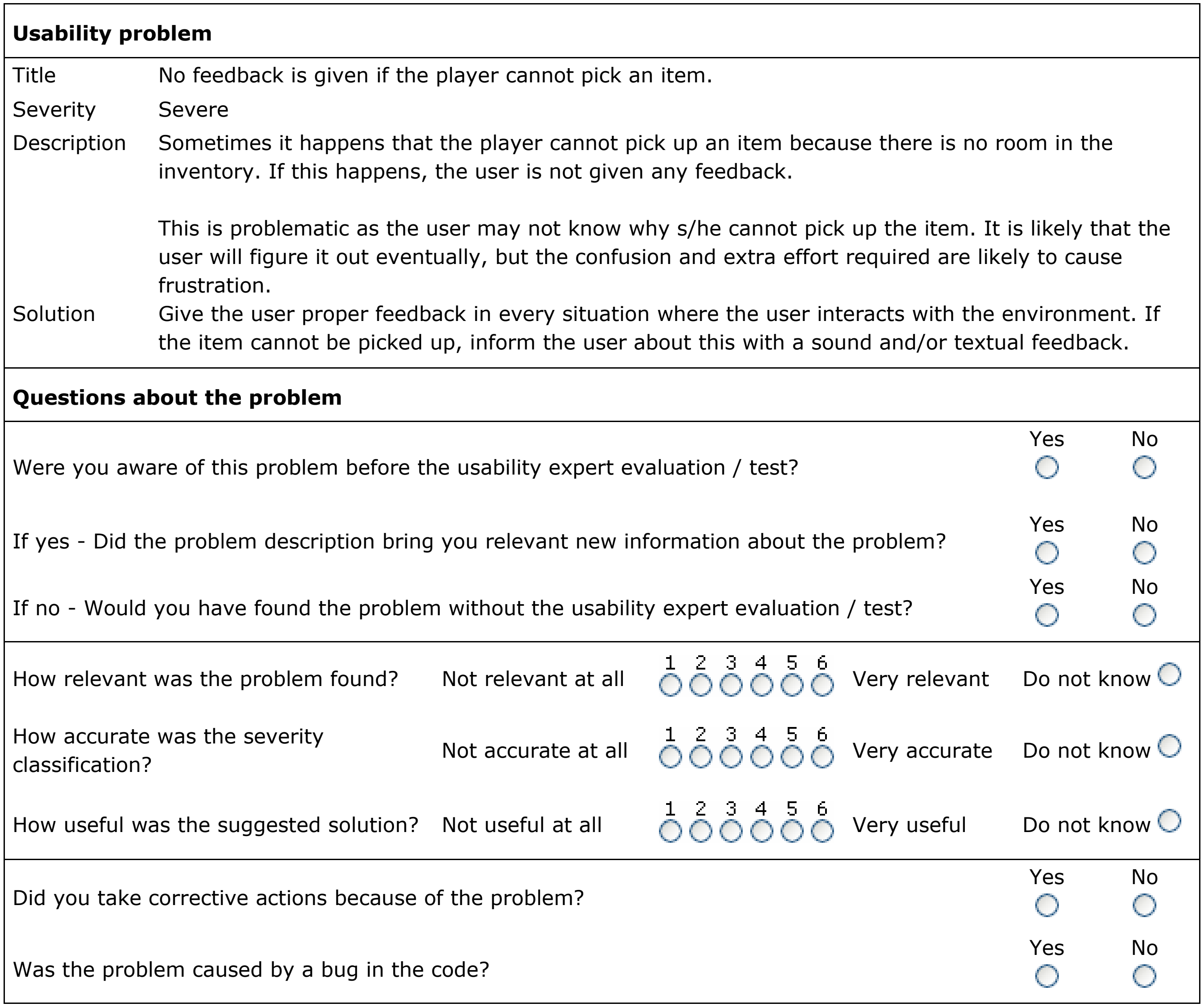
The process began with an initial meeting where game developers presented the game to the usability specialist who led both the usability expert evaluation and testing. In the presentation, the focus was on pinpointing what was supposed to be challenging to the gamers and what was not supposed to be a challenge. The goals of the usability expert evaluation and testing were also defined in the meeting.

Procedure - Usability expert evaluation

First, the six usability specialists evaluated the game independently of each other. The specialists played the game and wrote notes on the usability issues they found while playing. The findings were based on the Nielsen's usability heuristics (see Nielsen, 1993) and the specialists' knowledge and experience in human computer interaction. The specialists were told to evaluate the game like any other software product. No heuristics specific to computer or video game were used. Neither was any specific instructions given on what to focus in the game. Before the usability specialists started to evaluate the game they were instructed how to play the game and reminded that in games some issues are supposed to be challenging whereas everything else should be as easy as possible. The time the evaluation took varied from two to four hours per specialist.

After the evaluation, the specialists presented their findings to the evaluation leader and discussed the reasons behind the problems, severity classifications and the possible solutions. The leader then collected the problems to a single list. Once the list was ready the problems were grouped within predefined categories. After the categorization, similar problems within each category were grouped together. This categorized and grouped list served as the basis for the final report which was written by the lead specialist.

In the final report each problem had a title, severity classification, detailed description of the problem and suggested solution. A five step scale ranging from cosmetic to catastrophic was used to rate the severity of each problem. There were additional categories for the technical problems and the problems that could not be classified. See Figure 2 for an example of a problem and how they were reported. The final report was then delivered to the game developers, and a results meeting was held where the key findings were presented and discussed.



Usability problem

Title	No feedback is given if the player cannot pick an item.
Severity	Severe
Description	Sometimes it happens that the player cannot pick up an item because there is no room in the inventory. If this happens, the user is not given any feedback. This is problematic as the user may not know why s/he cannot pick up the item. It is likely that the user will figure it out eventually, but the confusion and extra effort required are likely to cause frustration.
Solution	Give the user proper feedback in every situation where the user interacts with the environment. If the item cannot be picked up, inform the user about this with a sound and/or textual feedback.

Questions about the problem

Were you aware of this problem before the usability expert evaluation / test?	Yes ○	No ○
If yes - Did the problem description bring you relevant new information about the problem?	Yes ○	No ○
If no - Would you have found the problem without the usability expert evaluation / test?	Yes ○	No ○

How relevant was the problem found?	Not relevant at all	1 ○	2 ○	3 ○	4 ○	5 ○	6 ○	Very relevant	Do not know ○
How accurate was the severity classification?	Not accurate at all	1 ○	2 ○	3 ○	4 ○	5 ○	6 ○	Very accurate	Do not know ○
How useful was the suggested solution?	Not useful at all	1 ○	2 ○	3 ○	4 ○	5 ○	6 ○	Very useful	Do not know ○

Did you take corrective actions because of the problem?	Yes ○	No ○
Was the problem caused by a bug in the code?	Yes ○	No ○

Figure 2. Feedback about the novelty and usefulness of the individual problems found was gathered with a web questionnaire. In the upper part of the figure there is an example of a usability problem found and way problems were reported.

Procedure - Usability test
The usability test was conducted two weeks after the usability expert evaluation had been reported to the game developers. The same version of the game was used as in the expert evaluation. The usability test was conducted by one usability specialist and an assistant.

In the usability test the six participants were run individually. The test was done in a standard usability laboratory. Each session started with an introduction where the user was told the basics about usability testing. After the introduction the user was taken to the laboratory and was briefed about the game. In the briefing the game's background story was told and the setting where the game started was explained. If the user had no questions the test began.

The test consisted of three parts. In the first part the user got acquainted with the controls and interacting with the game environment. In the second part the user played the game for 1.5 hours. The last fifteen minutes were spent playing the game with the cheat mode activated. This was done to evaluate features that were not directly available in the level played. After the test the user filled in a questionnaire measuring usability and user experience issues. In total, each session lasted for approximately two hours.

The think aloud method was used in the test. The user was instructed to tell what s/he was doing and why. The user was also encouraged to tell if s/he did not understand something plus any positive or negative thoughts that came to his/her mind. During the test the instructor interrupted the player every now and then either with a question or to give the user a task.

Each test was recorded, and the recordings were reviewed. When reviewing the tapes the usability problems were written down. The analysis and reporting of the problems was similar to the usability expert evaluation.

Procedure – Web survey
One month after the final meeting with the game developers, they were sent a link to the web survey. Two game developers answered the questionnaires independently of each other. In the instructions they were encouraged to be critical about the methods and the results.

Results

Results of the usability expert evaluation and testing
The number of problems found is summarized in Table 1. The same table also shows the distribution of the problems by severity.

	Non-applicable	Technical	Cosmetic	Minor	Intermediate	Severe	Catastrophic
Expert evaluation	23	2	1	42	60	30	2
Usability test	17	4	0	33	42	26	1
Total	40	6	1	75	102	56	3

Table 1. The total number of usability problems found and their distribution by the severity classification.

Novelty of the problems found
Out of all the problems found, 43% (N = 122) were new. Neither of the game developers knew about them prior to the usability expert evaluation and test. The remaining 57% (N = 161) were such that at least one developer had prior knowledge of it.

For each new problem, the game developers were asked if they believed they would have found the problem without the usability methods. For 74% (N = 88) of these problems both of the game developers answered that they would not have found the problem.

Relevancy of the problems found and the usefulness of the suggested solutions
The game developers were asked to rate each problem for relevancy. The mean rating was 3.68 (SD = 1.43) on a scale of 1 (not relevant at all) to 6 (very relevant). The summary of the results are presented in Table 2 and illustrated further in Figure 3.

	(1–not relevant at all, 6–very relevant)					
	1	2	3	4	5	6
Expert evaluation	8%	15%	19%	28%	19%	11%
Usability test	5%	17%	25%	22%	17%	14%
Total	7%	16%	21%	25%	18%	12%

Table 2. The distribution of the relevancy ratings the game developers gave to each problem found.

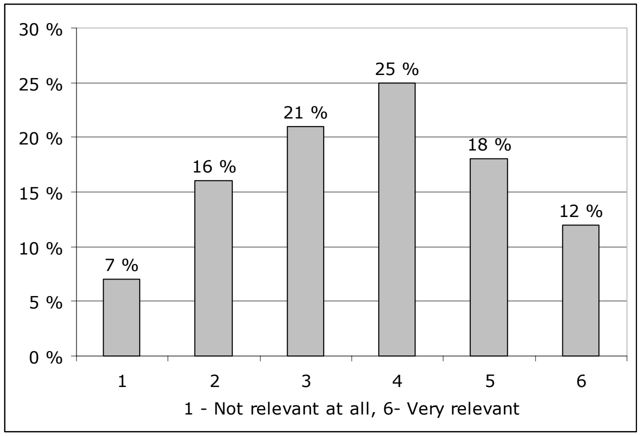


Figure 3. The distribution of the relevancy ratings for all the problems found in the usability expert evaluation and test.

When asked whether corrective actions were to be taken to fix the problems found in the usability expert evaluation and testing, the answers were as follows. Both of the developers answered yes for 31% (N = 89) of the problems. For 41% (N = 115) of the problems one developer answered yes and the other answered no. For the remaining 28% (N = 79) of the problems both developers answered no.

According to the game developers 85% (N = 241) of the problems were not due to a programming error. When asked how useful the suggested solutions were the mean was 4.31 (SD = 1.08) on a scale of 1 (not useful at all) to 6 (very useful).

Applying the usability methods to game development
A summary of the game developers' answers to the open-ended questions about applying the usability methods to game development is presented in Tables 3 and 4. In table 3 the positive feedback is presented and

in table 4 negative comments and the developers' suggestions for improvement are presented.

Positive comments
"The reports clarified many design issues that were known to be problematic already in the design phase."
"The findings helped improve numerous small details in the game, and to avoid a couple of potential pitfalls in designing and implementing new features."
"It is difficult to know how the game is played without testing it with the real users - gamers are not predictable."
"Expert evaluation is a fast and effective way to check the usability of a game"

Table 3. A summary of the positive feedback the game developers gave regarding the usability methods used.

Negative comments and ideas for improvement
"Some of the problems were known issues. Discussing these in more detail before the evaluation work would save time."
"The closer the release of the game the more important it is also to study the user experience."
"Level designers would do well to study the various player behaviors."

Table 4. A summary of the game developers' negative comments about the usability methods and their suggestions for future development.

The usability specialist who led the usability tests commented that the participants had sometimes difficulties answering the questions the specialist asked. This was because the game play was often hectic which made answering the questions difficult. For the very same reason thinking aloud was difficult for the participant from time to time.

Gamer versus non-gamer usability experts
The average relevancy rating given to the problems found by the usability specialists who were also gamers was 3.72 ($SD = 1.51$). The mean for the non-gamer usability specialists was 3.52 ($SD = 1.49$). There was no significant difference between these two groups, $t(170) = 0.53$, $p > 0.05$.

When the numbers of the usability problems that did not lead to corrective actions were compared, it was found that there was no significant difference between the gamer and non-gamer usability specialists ($Z = -0.78$, $p > 0.05$).

There was no significant difference between the number of usability problems found by the gamer and non-gamer usability specialists $t(4) = 0.27$, $p > 0.05$.

Discussion

Novel and useful data
Based on the results it can be concluded that usability expert evaluation and testing provide both novel and useful data. The view that the results were novel is supported by the finding that 43% of all the usability problems found were new to the game developers.

Additional support to the view that the results were novel is provided by the finding that the developers reported that they would not have found 74% of the new problems without the help of the usability methods used. This measure, of course, is not without problems. It can be argued that the developers cannot know for sure whether the problem would have been found later or not. Despite this, the result can be taken as further evidence about the novelty of the results at the time when they were reported.

The view that the data was useful is supported by three findings. First, the mean relevancy rating for the problems found was 3.68 (1-not relevant at all, 6-very relevant), which can also be considered good. Second, only 28% of all the problems found were such that the game developers had no intention to address them. Third, the game developers rated the usefulness of the suggested solutions high; the mean was 4.31 (1-not useful at all, 6-very useful). Together these findings indicate that the results are useful for the game developers.

The finding that the usability expert evaluation and testing provide useful results is in line with the previous studies (Desurvire et al. 2004; Medlock et al., 2002). Thus, there is now growing evidence about the usefulness of these methods in game development.

Face validity

The game developers reported that the usability expert evaluation and test helped them to improve numerous details in the game, avoid potential pitfalls when developing new features and to solve issues that they knew were problematic. The game developers also found observing the usability test informative because it gave a real life example how the gamers really play the game. These positive comments suggest that the usability expert evaluation and testing have face validity in game development.

The view that the methods are valid is supported even further by the finding that the game developers reported that 85% of the problems found were not due to a programming error. This is important because the usability expert evaluation and test are not supposed to replace the traditional quality assurance methods which are used for finding and fixing bugs.

When the game developers’ positive comments and the finding that the reported problems were not due to programming mistakes are combined with the findings that the results were both novel and useful, it can be concluded that the usability expert evaluation and testing have considerable face validity in game development.

All the usability specialists do not need to be double experts

There was no statistically significant difference in the number or the rated relevancy of the problems the gamer and non-gamer usability specialists found. Because of this it can be argued that all of the usability specialists who participate in the expert evaluation do not necessarily need to be double experts. This applies at least to the action adventure games that do not require extensive previous knowledge about the game type. More research is needed where the other game types are tested.

The finding that there was no difference between the problems the gamer and non-gamer usability specialists found in the expert evaluation does not mean that expertise would be useless. It is likely that if the specialists who participated this study would have had more experience in the game user research the results could have been even better than they were now. This is because it is through the experience usability specialists learn how the players play the games and what is really important in the games from the usability point of view and what is not. Studying how the game user research specific experience affects the quality of the problems found is an interesting topic for future studies.

Other findings

In the traditional usability test the instructor interrupts the participant every now and then to ask questions. In this study it was found that this is not always plausible in games. This is because the interruptions cause unnecessary difficulties to the participants. One potential way to avoid this problem is to have a mixture of think aloud and uninterrupted play.

In this study no survey methods were used to gather information about the user experience. After the usability test the game developers commented that they would have liked to learn more about the user experience. Because of this, it is recommended that post-test questionnaires are used to measure the user experience in the game usability tests. If detailed and statistically reliable measurements are needed, then playtesting methods developed especially for measuring the game user experience should be used (see Pagulayan, Keeker, Wixon, Romero and Fuller, 2003 for a review)

Suggestions for further studies

So far all the studies addressing the usefulness of usability expert evaluation and testing in game development have been case studies. Making generalizations based on case studies is difficult. Because of this more studies on this topic are needed. In the future studies other game genres and other usability evaluation methods should be studied. In addition to game developers, feedback should also be gathered from the game producers.

Another interesting topic for the future studies would be to compare the usability methods to the traditional quality assurance methods used in game development. This study and the common view suggest that these two traditions serve different needs and have different focuses, but there is little experimental information available whether this is the case or not.

A final suggestion for future studies is to start developing new user research methods that would support the level designers. This is because the goal of the level designers is to make levels that the players like. As the game developers suggested in their feedback, user research data could be potentially very useful in this work.

Practioner’s take away

- Traditional usability expert evaluation and testing provide novel and useful data for game development.
- All the usability specialists who participate in the usability expert evaluation of a game do not necessarily have to be double experts.
- When designing a game usability test it is important to notice that thinking aloud and interrupting the player are not always possible. Design the test so that there is a mixture of think aloud and uninterrupted play.

- The game developers are interested to learn about the user experience. Use post-test questionnaires and other survey methods to study the user experience.

Acknowledgements

I would like to thank Janne Sinkkonen for helping with the web questionnaire and senior usability specialist Rolf Södergård for commenting the manuscript.

References

Desurvire, H., Caplan, M., Toth, J. (2004). Using Heuristics to Improve the Playability of Games. CHI Conference, 2004, Vienna Austria (In the collection of Abstracts).

Federoff, M. (2002). Heuristics and usability guidelines for the creation and evaluation of fun in video games. Thesis at the University graduate school of Indiana university.

Fulton, B., & Romero, R. (2004). User-testing in a Hostile Environment: Overcoming Resistance and Apathy in your Game Company. Presented at the Game Developer's Conference, San Jose CA, March 2004.

John, B. E. (1998) A case for cases. In Commentary on "Damaged Merchandize?" Human-Computer Interaction 13(3), 289-296.

Nielsen, J. (1993). Usability engineering. San Francisco: Morgan Kaufmann.

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

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R., & Fuller, T. (2003). User-centered design in games. In J. Jacko and A. Sears (Eds.), Handbook for Human-Computer Interaction in Interactive Systems (pp. 883-906). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sauli Laitinen works as a project manager at Adage Corporation, a usability consultancy located in Helsinki, Finland. He is a psychologist (M. Psych.) and has a special interest in game user research. Sauli has also worked as a game designer on the X-Men and The Lord of the Rings mobile games.