A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games

Ester Baauw, Mathilde M. Bekker, and Wolmet Barendregt

TU Eindhoven, Department of Industrial Design,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{E.Baauw, M.M.Bekker, W.Barendregt}@tue.nl

Abstract. Inspection-based evaluation methods predicting usability problems can be applied for evaluating products without involving users. A new method (named SEEM), inspired by Norman's theory-of-action model [18] and Malone's concepts of fun [15], is described for predicting usability and fun problems in children's computer games. This paper describes a study to assess SEEM's quality. The results show that the experts in the study predicted about 76% of the problems found in a user test. The validity of SEEM is quite promising. Furthermore, the participating experts were able to apply the inspection-questions in an appropriate manner. Based on this first study ideas for improving the method are presented.

1 Introduction

1.1 Evaluation Approaches

Evaluation plays a crucial role in user-centred design in general, and also in the development of computer games for children. Globally two types of evaluation approaches exist for assessing interactive products: empirical evaluation methods and predictive or analytical evaluation methods. The main advantage of applying empirical methods is that real users are likely to find real problems. Overall, the strengths of predictive methods are that they are cheap to apply and they can be applied more easily early in the design process when only prototypes of products exist [17]. We are developing a predictive method for assessing usability and fun of children's computer games. This paper describes a study in which we assess the quality of the proposed method.

1.2 Predicting Problems in Computer Games

When developing games, the most important evaluation criterion is whether the game provides a fun experience. However, as indicated by Pagulayan et al. [19] usability should also be taken into account. They stated: 'The ease of use of a game's controls and interface is closely related to fun ratings for that game. Think of this factor as the gatekeeper on the fun of a game'. Thus, our predictive method should focus both on usability and fun.

Some well-known and frequently used predictive evaluation methods, focusing on usability are the Cognitive Walkthrough, Heuristic Evaluation and Guideline-based

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 457 – 469, 2005.

evaluation [6]. While the Cognitive Walkthrough is based on an underlying theory of exploratory learning, Heuristic Evaluation and Guideline-based evaluation are based on exploring an interface for breaches of design guidelines. Globally, these methods can be divided into two types of methods: the first group of types is based on underlying models of human behaviour (like the Cognitive Walkthrough) and the second group is based on collections of separate guidelines.

Not many specific predictive evaluation methods exist for evaluating computer games. Some of the existing methods are heuristic-based and focus on computer games for adults [9]. Other heuristics have been specifically developed for children's computer games, but are intended for design purposes, and not specifically for evaluating games [15]. Federoff [10] organized many of these existing guidelines. This set is quite large which is a disadvantage since the probability of conflicting statements increases with an increasing number of guidelines [16]. Another issue is that most guidelines are at a high level of abstraction, e.g. 'get the player involved quickly and easily', while others cover very specific design issues, e.g. 'minimize control options'. This makes these guidelines hard to use for predicting problems.

As far as we know, there are, to date, no existing predictive evaluation methods based on theory for the evaluation of (children's) computer games. Considering the drawbacks of applying guideline-based methods, we decided to develop a predictive method for identifying problems in children's computer games.

1.3 Predictive Method

First, a pilot study was executed to test the assumption that it is possible at all for adults to predict problems that children will encounter in computer games. Two adults, both with a good understanding of evaluation methods in general, (usability testing with) children and computer games, predicted problems in children's computer games without the use of a standard predictive method. The results of this pilot study will not be discussed in detail in this paper; however they showed that it is indeed possible for adults to predict problems in computer games for children. Even without the use of a standard evaluation technique the evaluators predicted about 40% of the problems that children encountered during user tests of this game.

In search of an appropriate theoretical basis for a predictive method, Norman's theory-of-action [18] was selected. This general model allows a systematic analysis of user product interaction. The model has two main aspects: the first aspect is Execution that covers planning the actions, translating the plans into actions, and executing the actions on the product. The second aspect is Evaluation, which covers both perceiving and interpreting the feedback and evaluating the outcome of the previous actions on the product. The model has the assumption of goal-driven behaviour. Goal-driven behaviour is also applicable for both children and computer games. To play a game successfully children have to reach certain goals (e.g. to collect all the right tools from various parts in the game in order to free the princess).

This model was employed for the construction of our predictive method, called Structured Expert Evaluation Method (SEEM). SEEM's checklist consists of questions based on Norman's stages complemented with questions based on the fun-related concepts from Malone [15], Challenge, Curiosity and Fantasy.[1] The general predictive questions were divided in a) questions for each screen of a computer game and b) more global questions that should be answered after evaluating the game. The following questions have to be checked at all screens:

1. Do children understand the goal?
2. Do children know what to do in order to accomplish the goal?
3. Are children able to perform the physical actions easily?
4. Can children perceive the feedback? This includes feedback (if any) from both correct and wrong actions, and whether children can click to stop the feedback.
5. Do children understand the feedback? This holds for both visual and auditory feedback from correct and wrong actions.
6. Will children keep on going until they reach the goal? This includes whether children will like the sub game and if the level of difficulty is okay for young children.
7. Are the navigation possibilities and the exits from a (sub) game clear?
8. Are there other objects in the Game Interface that will cause problems?

The global questions are:

1 a. Is the challenge right for the target group?
  b. Is the curiosity of children stimulated?
  c. Are the story and the interface tuned to the fantasies of children?
2 a. Is it clear whether a sub game is optional or obligatory?
  b. Does the flow of the game meet the expectations? Is the story line logical?
  c. Is it clear when a child should be either passive or active during the game?

As preparation for applying these questions to predict problems in children's computer games, a tutorial is provided. To get acquainted with the predictive method, the tutorial also provides many examples related to each of the predictive question.

1.4 Assessing the Quality of SEEM

To assess the quality of SEEM, two performance measures were used: thoroughness and validity [12] [7]. These measures were determined by having experts apply SEEM to evaluate two computer games. The resulting lists of problems were compared to the lists of problems obtained from User Tests (UT) of these games. Furthermore, the experts' understanding of SEEM's questions was examined by checking the appropriateness of the questions they used to identify problems. The problem predictions that did not match problems uncovered in UT were analyzed in detail. Based on these analyses suggestions for improvements of SEEM are made.

2 Method

2.1 Procedure

The participants in this study were experienced in at least one of the following areas: children, usability and/or usability testing methods and computer games. Their

[1] We have also used the combination of Norman's model and Malone's concepts of fun to structure design guidelines for children's computer games [3].

experience varied from 6 months to 20 years, with a mean of 4.4 years. In sum 10 male and 8 female participants (from now on called experts) from 11 different companies participated in the study. The youngest expert was 25 years old; the oldest was just above 40 (mean age was 30.3 years).

The experts evaluated two different Dutch computer games, ‘Milo and the magical stones’ [1] (from now on referred to as Milo) and ‘Roger Rabbit, Group 3: Fun in the Clouds’ [2] (from now on referred to as Roger).

The experts read a written tutorial before the test took place. This tutorial contained SEEM’s questions with corresponding examples from other computer games, descriptions of the computer games to be evaluated and the procedure experts had to follow. During the first meetings with experts (in small groups at the same time) they received a short training with another computer game, in which two sub games had to be evaluated in the correct manner. The problems obtained from UT of these sub games were shown and discussed. The aim of the tutorial and training was to increase experts’ understanding of SEEM. Understanding of the method is important since analysts who do not understand an inspection method can readily both come up with many false positives and fail to predict problems [8].

After this training, the actual evaluation started. The first game had to be evaluated for one hour. Evaluators were told to go at least once through the questions at each screen. After 55 minutes experts were asked to take a closer look at the global questions. Experts took a short break when they were finished analyzing and reporting problems that related to the two final questions. After the break experts were given the instruction to look specifically at predetermined sub games for a further 45 minutes. These sub games were selected because about 50% or more of the children visited these screens during UT and these screens contained many uncovered problems. After 40 minutes experts once again were requested to focus on the two global questions.

The order of the games was randomly determined. Experts took the second game home and evaluated it there. The instructions for the evaluation did not change.

2.2 Problem Report

For each predicted problem evaluators filled in an Interaction Problem Report (IPR). The format for IPR’s is based on Lavery et al. [14]. Experts had to fill in the screen number, the predictive question the problem referred to, a short problem description, expected causes of the problem and expected outcomes of the problem. By constraining experts to use this format, the comparison of their predictions to the problems uncovered by children became easier.

3 Analysis of the Data

3.1 Creating the Actual Problem Sets

To determine the thoroughness and validity of SEEM, standard problem sets were needed for both computer games. UT was used to generate these touchstone sets of usability problems [12]. For both computer games children participants played the game as they liked in sessions that lasted about 30 minutes.

Twenty-six children participated in the UT of Milo, which makes it likely that almost all problems were detected because the total number of usability problems found levels off asymptotically as the number of participants increases [12]. Only seven children participated in the UT of Roger. All children were between 5 and 7 years old. For information of the study set-up and data analysis see Barendregt et al. [4]. The data analysis resulted in a list of 86 actual problems for Milo, and a list of 39 actual problems for Roger.

3.2 Determining the Thoroughness

The thoroughness of SEEM was assessed with the following formula [7]:

$$\text{Thoroughness} = \frac{\text{hits}}{(\text{hits} + \text{misses})} \tag{1}$$

Hits are predictions matched to problems found in UT, and misses are problems found in UT that are not predicted. Two researchers judged whether a prediction from an expert matched with an actual problem (which made the problem prediction a hit). When they disagreed, they discussed the problem prediction until they both agreed whether it was a hit or not. The thoroughness of SEEM was compared to the thoroughness for another inspection method.

3.3 Determining the Validity

The validity of SEEM was determined with the following formula [7]:

$$\text{Validity} = \frac{\text{hits}}{(\text{hits} + \text{false positives})} \tag{2}$$

As stated before, hits are predictions from experts that are matched to actual problems from UT. False positives are predicted problems that do not occur in the actual problem set derived from UT. The validity of SEEM was compared to the scores of another existing method.

3.4 Determining the Appropriateness of SEEM's Questions

As Cockton and Woolrych [8] stated: “For heuristics to be shown to have a role in problem discovery or analysis, appropriate heuristics must be associated with problems”. The same applies for the questions of SEEM. Therefore two evaluators checked the appropriateness of the questions that the experts had filled in on the IPR. The following categories were used:

1. Correct use: when an expert used a correct question, or when the choice was not optimal, but this question was possible in relation to the problem
2. Incorrect use: when experts did not fill in any question, or when they used a wrong question in relation to the problem.

4 Results

4.1 Thoroughness

Table 1 shows the thoroughness of SEEM. Thoroughness scores have a range from 0 to 1, with an optimal value of 1.

Table 1. Thoroughness of SEEM

	Lowest score	Highest score	Median	Mean	Sum (n=18)
SEEM Milo	0.08	0.29	0.15	0.17	0.77
SEEM Roger	0.05	0.28	0.15	0.18	0.74
SEEM	0.10	0.29	0.16	0.17	0.76

The third row is not simply an average of the two computer games, e.g. the lowest score for SEEM is higher than the average of the two scores above. An explanation is that the expert with the lowest thoroughness at Milo compensated this with a higher thoroughness at Roger. The lowest scores are from different experts. The same goes for the other numbers. The sum shows that for the two computer games together about 76% of the actual problems were predicted by experts while using SEEM. The sum is much higher than the individual scores from experts. Based on a study that compared three predictive evaluation methods, Sears [20] also concluded that the thoroughness increases as more evaluators participate. This is due to the increasing number of hits whereas the number of hits plus misses (the actual problem set) remains unchanged.

These numbers can be compared to values from other studies to give a global impression of their quality. Cockton et al. [7] conducted a Heuristic Evaluation with many participants (31 analysts divided over 10 groups in the latest study compared to 96 analysts divided over 16 groups in an earlier version of the study) to validate the DARe model. They compared the predictions from the Heuristic Evaluation with problems uncovered with UT. In the latest study 5 users were included in UT; in the earlier study 15 users were included. They found a thoroughness of 0.70 in the latest study; in the earlier study they found a thoroughness of 0.63. Compared to these thoroughness scores for a Heuristic Evaluation, the results from SEEM are promising.

Apart from overall thoroughness, we determined separate thoroughness scores per severity category, similar to Hartson's approach [12]. Frequency severity stands for the number of children that experience a problem, impact severity stands for the seriousness of the consequences a problem has for children [4]. Table 2 shows the thoroughness of the different severity categories.

Table 2. Thoroughness of SEEM per severity category

	Frequency severity			**Impact severity**		
	High	Medium	Low	High	Medium	Low
SEEM Milo	1	0.88	0.70	1	0.94	0.70
SEEM Roger	1	1	0.70	0.80	0.78	0.72
SEEM	1	0.90	0.70	0.91	0.88	0.71

Striking is that for both computer games none of the high frequency severity problems were missed. The two medium frequency severity problems that were not predicted both have a low impact severity.

Experts failed to predict only one high impact severity problem, however only one child found this problem. In Milo one medium impact severity problem was not predicted, however only one child out of 26 found this problem. The two problems with medium impact severity from Roger also have a low frequency severity; only one child found them. Besides this, these problems were very detailed and therefore hard to predict. So by far most misses have a low frequency or impact severity.

Separate thoroughness scores per question were determined to examine SEEM's quality in more detail. All problems uncovered with UT were distributed over the different predictive questions. Table 3 shows the thoroughness scores only for the predictive questions which had related misses.

Table 3. Thoroughness of SEEM per predictive question

	1 Goal	2 Translation	3 Physical actions	5 Understanding feedback	8 Other
SEEM Milo	0.88	0.80	1	0.38	0.78
SEEM Roger	1	0.56	0.80	0.67	- *
SEEM	0.92	0.74	0.83	0.47	0.78

* Because there were no actual problems for this category, no thoroughness score could be determined.

Only a few problems regarding the Goal, Translation, Physical Actions, and Other (they were all technical problems) were not predicted by SEEM. The number of misses regarding Understanding Feedback is relatively high. Zapf et al. [21] investigated the error detection of computer users using word processing. They also found that errors in the feedback phase were particularly hard to detect for users. A possible explanation is that experts experienced a problem but did not realize it. Only when a person really knows a computer game very well, it is possible to predict whether someone else will see and understand the feedback.

4.2 Validity

Table 4 shows the validity of SEEM, the validity scores range from 0 to 1 with 1 being the optimal score.

Table 4. Validity of SEEM

	Lowest score	Highest score	Median	Mean	Sum (n=18)
SEEM Milo	0.54	0.90	0.71	0.69	0.47
SEEM Roger	0.15	0.63	0.39	0.37	0.19
SEEM	0.34	0.66	0.57	0.53	0.33

The validity of SEEM for Roger is much lower than the validity for Milo. This is due to the number of false positives; for Roger this number (124) is much higher than the number for Milo (73). This can be partly explained because it is likely that the problem set of Roger is not complete yet, since only 7 children participated in the UT. Kanis and Arisz [13] show that it is possible to calculate how many participants should be included in a test to be able to detect 80% of all usability problems. It turned out that UT for Roger should have been done with 11 children (instead of 7), and then the number of uncovered problems would have been 61 (instead of 39).

The validity decreases (but not as much as the thoroughness increases) as the number of experts increases. This means that in terms of percentage the number of false positives increases faster than the number of hits. This is in line with Sears' findings [20].

Compared to values from other studies SEEM's scores are still promising. Cockton et al. [7] found a validity of 0.31 for the Heuristic Evaluation, which in a later version increased to 0.50. This means that SEEM scores about equally good as their first version, but improvements are still desirable.

4.3 Appropriateness of SEEM's Questions

Table 5 shows the numbers and percentages of both the correctly and incorrectly applied questions.

Table 5. SEEM's (in-)correctly applied questions by experts

	Correct		Incorrect	
	Number	Percentage	Number	Percentage
SEEM Milo	203	75.2	67	24.8
SEEM Roger	86	70.5	36	29.5
SEEM	289	73.7	103	26.3

The results show that almost 74% of the problems were related to a correct question. Cockton et al. [7] found percentages of 31% and 57% in their experiments with the Heuristic Evaluation, meaning that the appropriate application of SEEM's questions is very good.

5 Improving SEEM

5.1 Increasing the Thoroughness

Although the thoroughness is quite high, there were quite a few misses. A possible explanation for not predicting some of the actual problems is that the experts did encounter some problems but they did not write them down. It is possible that this happened because the experts realized faster than children what went wrong and what they were supposed to do. This hypothesis was confirmed by observing some experts during their evaluations and by statements from experts made after the evaluation. By stressing the importance of immediately writing down the problem predictions and by urging experts to constantly imagine how children would use the game; the number of missed problems due to not writing them down could be decreased.

5.2 Increasing the Validity

The results show that validity is somewhat disappointing. This is due to the many false positives predicted by the experts. This number is relatively high and should decrease. Probable causes for the list of false positives were determined to see how SEEM could be improved. Two possible causes were frequently found:

1. Under- or overestimation of children. Many experts made wrong assumptions about the (cognitive) level of understanding from children.
2. Incorrect assumptions about the game. Most of these predictions occurred because there are multiple solutions to reach a sub goal in the games. For example it is almost never necessary to click at exactly one small hotspot; clicking beside the hotspot will also make the feature work. It is possible that experts did not realize this and therefore still reported this as a problem.

To decrease the number of false positives caused by incorrect assumptions of the game a suggestion is to make sure that experts evaluate sub games more than once. That way it is possible for experts to try more things in the game and get to know the game better. This could also decrease the relatively high number of Understanding Feedback problems. Giving experts more specific information about children and their cognitive level of understanding could decrease the number of false positives caused by over- or underestimating children.

Finally, the questions were analyzed and some of them were changed. Some problem predictions were described at such a high level that it was hard to judge their realness. An example is: ‘The flow of the game is not logical’ with no further information given. Problem predictions that were too general were mainly due to the format of the global questions. It would be better to let experts have a look at these aspects (if still necessary) at each screen to make it possible to write down detailed and complete problem predictions. Furthermore, questions that were often used in the wrong manner or that generated many false positives had to be changed. Based on these considerations and comments from the experts, the new questions are the following:

1. Goal: Can children perceive the goal? Do children understand the goal? Do children think the goal is fun?
2. Planning and translation into actions: Can children perceive and understand the actions they have to execute in order to reach the goal? Do children think the actions they have to execute in order to reach are fun?
3. Physical actions: Are children able to perform the physical actions easily?
4. Feedback (after correct and wrong actions): Is the feedback (if any) perceivable? Do children understand the feedback? Is the feedback motivating?
5. Continuation: Is the goal getting closer fast enough? Is the reward in line with the effort children have to do in order to reach the goal?
6. Navigation: Are the navigation possibilities and the exits from a (sub) game clear?
7. Are there other (e.g. technical) problems?

These new questions will be tested in a follow-up study.

6 Discussion

6.1 Generalizability

To increase the generalizability of our results experts had to evaluate two computer games. Although both games are adventures, one (Roger) focuses more on education than the other (Milo). Together they cover a wide range of activities that are often presented in adventure games for children, like motor skill games and cognitive challenges. Therefore the combination of these two games is a good representative of the genre. The results on thoroughness, validity and the understanding of SEEM show similar trends for the two games. Thus, SEEM is likely to perform similarly for other computer games as well.

6.2 Comparing Evaluation Methods

There are several issues related to conducting quality assessments of evaluation methods, among other things the number of experts participating in the study and the assumption that UT uncovers all actual problems. The large number of experts may have influenced the validity and thoroughness scores of SEEM. If every subsequent expert finds relatively more false positives than extra hits, then the validity score will decrease with more experts. In contrast, the thoroughness score is likely to increase with more experts, since the number of hits is likely to increase. This is one of the reasons why it is difficult to compare validity and thoroughness scores of different predictive evaluation methods when the studies are conducted with different numbers of experts. Therefore, SEEM's quality scores were compared to those of the Heuristic Evaluation as determined by Cockton et al. [7]. This has been done because our assessment approach is similar to theirs in terms of judging the realness of problems, the definition of the quality criteria and the number of experts (in their study groups).

We applied a similar approach as used by Hartson [12] and Cockton et al. [7] to assess the quality of SEEM. This assessment approach assumes that UT is capable of finding all problems in a game. However, various researchers have argued that predictive methods can predict actual problems not found in UT. As Gray and Salzman [11] already stated: 'It is a sure bet that no usability evaluation method (both empirical and analytical) is perfect; any method will detect some problems while missing others'. Nielsen [17] argues that predictions not found in UT were not false positives for the Heuristic Evaluation but were due to the characteristics of the users who were involved in UT. Finally, Chattratichart and Brody [5] introduce the term *false negatives* to describe the real problems uncovered by a predictive method, which are not uncovered by the UT. Thus, SEEM may have been able to uncover additional real problems that were not uncovered by UT. An example of such a problem is the following: in Milo children have to click at two crabs that make the same sound. However, these crabs walk around and all look alike, so it is impossible to follow any tactic. Children just clicked the crabs randomly until they clicked the right ones. The experts predicted that it would be more fun when children could use a tactic to solve this sub game. However, none of the children explicitly indicated this. Thus, while this problem was not found in UT, it could very well be true.

Because the validity measure is based on all false positives including those that may be real problems, it is a conservative estimate of SEEM's validity. To investigate this effect, we re-analysed the set of false positives. Two researchers independently determined whether a false positive was a true false positive or a false negative. Of the 73 false positives for Milo 45 were judged to be *false negatives*, and 77 of the 124 false positives were judged to be *false negatives* for Roger. As a consequence the validity score of SEEM would increase because the number of false positives would decrease.

6.3 Number of Experts

In this study many more experts were involved than e.g. the 5 that are normally advised for conducting a Heuristic Evaluation [17]. This was done because the aim was to test and further develop SEEM, and not to test the computer games. Another reason for including a large number of experts in this study was to decrease the influence of experts performing either very well or very poorly.

6.4 Type and Amount of Expertise

Nielsen [17] states that usability specialists were better than non-specialists at performing a Heuristic Evaluation. Nielsen distinguishes three levels of expertise; novice, single expert or double expert. In our study none of the participants can be determined as novice. All the 18 experts in this study were experienced in at least one of the following areas: children, usability and/or usability testing methods and computer games. A preliminary analysis of the results in relation to the expertise gives the impression that the scores for thoroughness and validity do not differ much as the expertise increases. A possible explanation could be that the experts do not differ greatly in their level of expertise and therefore there are no clearly marked differences in thoroughness and validity.

However, none of the experts can be categorized as novices. It is possible that the promising results of SEEM are only practicable with experts and not with novices. The differences between experts and novices regarding the thoroughness, validity and appropriateness of SEEM's questions will be investigated further in a follow-up study with novices.

7 Conclusion

The study shows that SEEM predicts actual problems quite well, the thoroughness of SEEM is 0.76. The problems from UT that are not predicted by experts can mainly be assigned to the Understanding Feedback question. Only very few severe problems from UT were not predicted by experts while using SEEM. Unfortunately the number of false positives is also rather high, resulting in a fairly low validity. The results show that experts understood SEEM quite well; almost 74% of the problem predictions were correctly related to a predictive question. Based on the analyses of the missed problems, the appropriateness of the questions, the causes of the false positives, and the useful comments from experts, the method was improved. Because the findings are very promising, we intend to conduct another study to evaluate the quality of the new version of SEEM.

Acknowledgements

First of all we thank all the experts for participating in our study. We also thank Professor Bouwhuis from the TU Eindhoven for his useful comments on an earlier version of this paper. Finally, we thank the Innovation-oriented Research Program Human-Machine Interaction (IOP-MMI) of the Dutch government that provided the grant that has made this research possible.

References

1. Milo and the magical stones (Max en de toverstenen). MediaMix Benelux (2002)
2. Roger Rabbit, Group 3: Fun in the Clouds (Robbie Konijn, Groep 3: Pret in de Wolken). Mindscape (2003)
3. Barendregt, W. and Bekker, M.M.: Towards a Framework for Design Guidelines for Young Children's Computer Games. Proceedings of the 2004 ICEC Conference, September 2004 The Netherlands, Eindhoven, Springer-Verlag (2004) 365-376
4. Barendregt, W., Bekker, M.M., Bouwhuis, D. & Baauw, E.: Predicting effectiveness of children participants in user testing based on personality characteristics. Submitted to Behaviour & Information Technology (Unpublished manuscript)
5. Chattratichart, J. and Brodie, J.: Applying User Testing Data to UEM Performance Metrics. Late Breaking Results Paper, 24 April 2004 Vienna, Austria (2004) 1119-1122
6. Cockton, G., Lavery, D. & Woolrych, A.: Inspection-based evaluations. In: Jacko, J. and Sears, A. (Eds.): The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Associates (2003) 1118-1138
7. Cockton, G., Woolrych, A., Hall, L. & Hindmarch, M.: Changing Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment. In: Palanque, P., Johnson, P. and O'Neill, E. (Eds.): People and Computers, Designing for Society (Proceedings of HCI 2003). Springer-Verlag (2003) 145-162
8. Cockton, G. and Woolrych, A.: Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation. In: Blandford, A., Vanderdonckt, J. and Gray, P.D. (Eds.): Springer-Verlag (2001) 171-192
9. Desurvire, H., Caplan, M. & Toth, J.A.: Using heuristics to evaluate the playability of games. CHI extended abstracts 2004, Vienna, Austria (2004) 1509-1512
10. Federoff, M.A.: Heuristics and usability guidelines for the creation and evaluation of fun in video games. Msc Department of Telecommunications of Indiana University (2002)
11. Gray, W.D. and Salzman, M.C.: Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction, 13(3) (1998) 203-261
12. Hartson, H.R., Andre, T.S. & Williges, R.C.: Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction: Special issue on Empirical Evaluation of Information Visualisations, 13(4) (2001) 373-410
13. Kanis, H. and Arisz, H.J.: How many participants: A simple means for concurrent monitoring. Proceedings of the IEA 2000/HFES 2000 Congress, (2000) 637-640

14. Lavery, D., Cockton, G. & Atkinson, M.P.: Comparison of Evaluation Methods Using Structured Usability Problem Reports. Behaviour and Information Technology, 16(4) (1997) 246-266
15. Malone, T.W.: What makes things fun to learn? A study of intrinsically motivating computer games. Technical Report CIS-7, Xerox PARC, Palo Alto (1980)
16. Nes, F. v.: On the validity of design guidelines and the role of standardisation. In: Nicolle, C. and Abascal, J. (Eds.): Inclusive Design Guidelines for HCI. London and New York, Taylor & Francis Group (2001) 61-70
17. Nielsen, J. and Mack, R.L.: Usability Inspection Methods. New York, John Wiley & Sons, Inc. (1994)
18. Norman, D.A.: The design of everyday things. London, MIT Press (1998)
19. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R. & Fuller, T.: User-centered design in games. In: Jacko, J. and Sears, A. (Eds.): Handbook for Human-Computer Interaction in Interactive Systems. Mahwah, NJ, Lawrence Erlbaum Associates (2003) 883-906
20. Sears, A.: Heuristic Walkthroughs: Finding the problems without the noise. International Journal of Human-Computer Interactions, 9(3) (1997) 213-234
21. Zapf, D., Maier, G.W. & Irmer, C.: Error Detection, Task Characteristics, and Some Consequences for Software Design. Applied Psychology: an international review, 43 (1994) 499-520