Cogn Tech Work (2008) 10:129–140
DOI 10.1007/s10111-007-0068-x

ORIGINAL ARTICLE

A comparison of two analytical evaluation methods for educational computer games for young children

Mathilde M. Bekker · Ester Baauw · Wolmet Barendregt

Received: 3 July 2006 / Accepted: 4 March 2007 / Published online: 3 April 2007

Abstract In this paper we describe a comparison of two analytical methods for educational computer games for young children. The methods compared in the study are the Structured Expert Evaluation Method (SEEM) and the Combined Heuristic Evaluation (HE) (based on a combination of Nielsen's HE and the fun-related concepts from Malone and Lepper) with both usability and fun heuristics for children's computer games. To verify SEEM's relative quality, a study was set up in which adult evaluators predicted problems in computer games. Outcomes based on thoroughness (whether the analytical method finds all problems), validity (whether the analytical method uncovers problems that are likely to be true) and appropriateness (whether the method is applied correctly) are compared. The results show that both the thoroughness and validity of SEEM are higher than the thoroughness and validity of the Combined HE. The appropriateness scores indicate that SEEM gives evaluators more guidance when predicting problems than the Combined HE does.

Keywords Analytical evaluation methods · Children · Computer games · (Combined) heuristic evaluation · SEEM

M. M. Bekker (✉) · E. Baauw · W. Barendregt
Department of Industrial Design,
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven,
The Netherlands
e-mail: M.M.Bekker@tue.nl

E. Baauw
e-mail: esterbaauw@hotmail.com

W. Barendregt
e-mail: wolmetb@hotmail.com

1 Introduction

Analytical or inspection-based methods rely on evaluators assessing (usability-related) aspects of a user interface. These methods are popular because they often require less formal training, take little time to apply, can be used both early and late in the development process and do not require test users (Sears 1997). Furthermore, they are complementary to empirical testing approaches, such as usability testing with users (Baauw et al. 2005; Chattratichart and Brodie 2004).

This paper presents a comparative study of two analytical evaluation methods (AEMs) for evaluating children's educational computer games from the adventure genre. It compares a new AEM, called Structured Expert Evaluation Method (SEEM), with a Combined Heuristic Evaluation (Combined HE) both intended to assess usability and fun. The initial development of SEEM and its assessment compared to User Testing (UT)[1] was described separately in an earlier paper (Baauw et al. 2005). Eighteen experts participated in the first study. They were experienced in at least one of the following areas: children, usability and/or usability testing methods and computer games. The experts evaluated two different educational computer games for young children (aged 5–7). They predicted 76% of the problems uncovered with UT of the same game. Because the experts also predicted many problems that were not found during UT, an improved version of SEEM was created.

[1] Even though we believe that the term Usability Testing touches the essence of the test better than the term User Testing, in this paper the term User Testing will be used to make the distinction with the analytical evaluation methods clearer.

1.1 Usability and fun

Analytical evaluation methods for work-related products are usually aimed at finding usability problems. When a usability problem is encountered this means that a user is not able to reach a goal in an efficient, effective or satisfactory way. However when developing computer games, the most important evaluation criterion is whether the game provides a fun experience. Therefore, it is not sufficient to focus on usability alone. Pagulayan et al. (2003) wrote '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 the quality of a computer game depends on both usability and fun. Furthermore, when testing either fun or usability it is very likely that problems of the other type will be encountered as well. As a consequence, a list of both fun and usability problems and causes to be fixed by the developers should be created (Barendregt et al. 2003).

Structured Expert Evaluation Method is an analytical method that has been developed to assess usability and fun problems of young children's computer games. SEEM has been developed to evaluate adventure games, which is a very common game genre for children between 5 and 7 years old. Most adventure games focus on exploration; usually involve item gathering and simple puzzle solving. They have roughly the same structure; they contain several sub games that have to be played in order to reach an overall goal. Among these sub games there are usually motor-skill games and logical games. Sub games often have some educational value. Children usually have some freedom in deciding the order in which they want to play the sub games.

1.2 Related design and evaluation methods

To put SEEM's scope in perspective, this sections mentions a limited number of related design and evaluation methods. Many well-known analytical methods, such as the HE (Nielsen 1994) and the Cognitive Walkthrough (CW) (Wharton et al. 1994) focus on the usability of products in general. Most analytical methods developed specifically for evaluating computer games focus mostly only on fun and on adult games like, e.g. Desurvire et al.'s set of 43 heuristics for playability (2004) and Federoff's set of 40 heuristics for fun and usability (2002). Fabricatore et al. (2002) provide a design model on playability, with a very large set of detailed design recommendation and prescriptions, which is intended for action video games for adults. Another very large set of design guidelines, which was developed specifically for children's products, focuses on websites for children, but not on computer games (Gilutz and Nielsen 2002). Finally, the one set of heuristics which was developed specifically for children's computer games focuses mostly on fun heuristics, and is intended for design and not for evaluation purposes (Malone 1980, Malone and Lepper 1987). The part of the set that focuses on individual, as opposed to interpersonal motivations, consists of four main and 12 related and more detailed heuristics. In summary, many design guideline and heuristic sets exist, but they vary on their intended scope; for children or for adults, their intended purpose; for design or for evaluation and also on the amount of items in the respective sets and their level of detail; ranging from very abstract, e.g. games should be easy to learn, to very detailed, e.g. the appearance should always transmit some information about what the entity is wearing or using.

1.3 The Structured Expert Evaluation Method

The Structured Expert Evaluation Method consists of a checklist with questions, originally based on Norman's theory of action model (Norman 1998) and on Malone's concepts of fun (Malone 1980). Norman's model allows a systematic analysis of user–product interaction. The model consists of two main phases of user product introduction: first, the Execution phase, that covers planning the actions, translating the plans into actions and executing the actions on a product, and second, the Evaluation phase, 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. This kind of behaviour is also applicable for both children and computer games from the adventure genre. 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). SEEM's questions focus on the various phases of Norman's action cycle, complemented with questions based on the fun-related aspects from Malone. The questions deal with the goal, the planning and translation into actions, the physical actions, the feedback and the continuation in the game. So, e.g. evaluators have to determine whether the goal can be perceived, understood and whether the goal will be fun (see Fig. 1).

Based on previous pilot studies a separate question about the navigation, which is quite an important aspect of computer games, was added. Navigation between different screens and sub games in these kinds of games is often realized by clicking at the edge of the screen or by clicking an arrow-like button. Navigation issues are treated in a separate question because it makes applying the walkthrough analysis easier in the context of screen-based interactions.

Since both SEEM and the CW are based on Norman's action cycle, the two methods are very similar in their

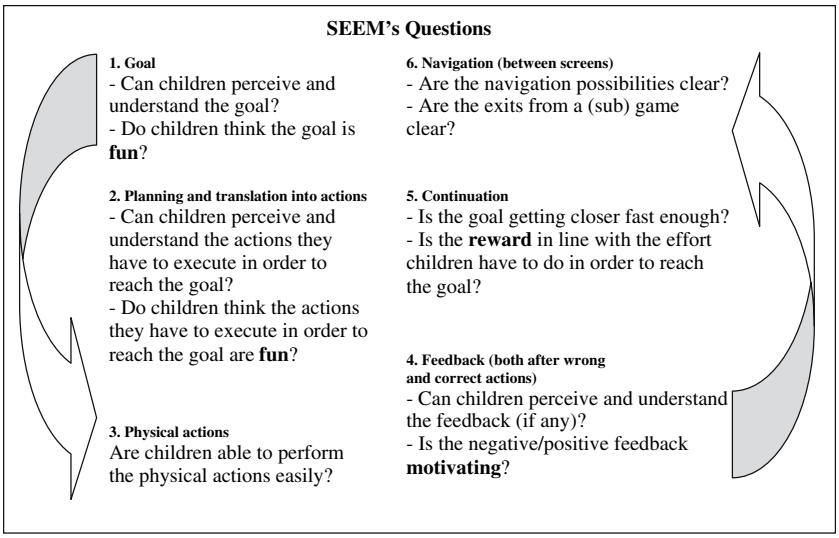

Fig. 1 The task and fun-related questions of SEEM, following the *action cycle*, which have to be checked at each screen of a computer game

evaluation approach. In contrast with CW, experts applying SEEM do not have to write success and failure stories for each question. They have to fill in a problem report template when they feel the answers to one of the questions is 'no'.

Furthermore, SEEM combines the task-based approach for the basic structure of the questions, with fun-based heuristics. This integrated fun and usability question-based structure is the result of various studies in which experts reasoned about usability and fun problems in children's computer games (Barendregt et al. 2003, Barendregt and Bekker 2004, Bekker et al. 2004).

1.4 Predicting problems with SEEM

To predict problems evaluators, also called experts, receive the game to be evaluated, a manual explaining the use of SEEM, a description of the game and a description of the focus of the evaluation, e.g. what screens and sub games to evaluate and the children's age range to be taken into account. Subsequently, they will evaluate the game, while exploring and playing the game themselves. Evaluators are expected to check all questions for each screen or sub game, and in case the answer to a question is 'no', record the predicted problem in a problem report template. When they are finished, the experts hand in their problem reports. The individual problem reports of separate experts are subsequently combined into an overall usability and fun problem report, including a list of the most important usability and fun problems found. For the purpose of a comparison study such as described in this paper, the emphasis is on the number of evaluators, who found each problem. When the report is written for the developer of the game the emphasis would be more on what design decisions to change and why.

1.5 Assessing the quality of SEEM

In a previous study to assess a first version of SEEM 18 experts participated (Baauw et al. 2005). They were experienced in at least one of the following areas: children, usability and/or usability testing methods and computer games. The experts evaluated two different educational computer games for young children (aged 5–7). The results showed that the experts predicted 76% of the problems uncovered in UT of the same game. The problems in the UT were determined using a coding scheme with breakdown indication types for usability and fun problems (Barendregt and Bekker 2006). Unfortunately, experts also predicted many problems that were not found during UT. Based on these findings an improved version of SEEM was created. The main change consisted of integrating the questions related to fun issues into the action cycle, instead of adding them as a separate set of questions in addition to questions about the action cycle.

This paper describes a second study in which the improved version of SEEM was compared to an alternative AEM. While in the first study the predictions using SEEM were compared to the outcome of UT, the second study also examines how SEEM compares to another predictive method. Since none of the existing analytical methods have exactly the same scope as SEEM, i.e. predicting both fun and usability problems of children's computer games, two existing methods were combined for the comparison study. Most of the methods are heuristic-based and contain a fairly large set of heuristics. Despite the fact that some of the sets with fun-related guidelines are more recent than the one developed by Malone and Lepper (1987), we chose their set because the items still capture many of the most relevant issues (e.g. a game should have a clear goal). Also the set is manageable and developed specifically for

children's games. This set of four heuristics was combined with Nielsen's set of ten usability heuristics (Nielsen 1994) to ensure that the combined set of 14 heuristics would focus both on usability and fun. Nielsen's set was chosen since these usability heuristics have undergone extensive testing and several iterations of design. Other heuristics have not been used as often as Nielsen's heuristics have and may need further work before they are ideal for use in a HE. For the purpose of this paper the alternative method is called the Combined HE to indicate the difference with, e.g. Nielsen's HE (Nielsen 1994).

2 Performance metrics

Globally, existing analytical methods follow two different approaches. The first approach requires evaluators to assess whether a design complies with a set of 'rules of thumb'. The second approach takes a walkthrough form, in which evaluators follow a step-wise process to uncover possible user problems. SEEM follows the walkthrough approach, while the Combined HE follows the 'rule-comparison' approach. Two frequently compared analytical methods are the CW and the HE like, e.g. Sears does in his study (Sears 1997). Since no comparative studies have been conducted on analytical methods for assessing children's computer games, no assumptions about the direction of the research questions were derived from existing comparative studies.

To assess the relative quality of AEM's, the problem predictions should ideally be compared to a total list of real problems in the product (see Fig. 2a). However, since it is impossible to determine the total real problem set, UT was used to generate a standard problem set as is state-of-the-art procedure in method comparison studies (e.g. Hartson et al. 2001; Cockton et al. 2003). To make the outcome of the user test (the bottom-left circle in Fig. 2a) to be as close to the real problem set as possible (the top circle in Fig. 2a) many children (26) were involved in the user test and a rigorous analysis procedure was followed (see Sect. 2.1). These minimized the risk of missing problems or including 'non-real' problems by misinterpreting users' behaviour. The standard problem set based on UT (see Fig. 2b) is used as a benchmark to determine whether the analytical methods find all problems (thoroughness), and whether the analytical methods make predictions that are likely to be true (validity) (Hartson et al. 2001). These scores are determined for all possible groups of 2 up to 9 evaluators, since the value of these scores are highly dependent on the number of evaluators. The values for thoroughness tend to increase with more experts, whereas the values for validity tend to decrease with more experts.

Furthermore, the extent to which the methods were applied as intended (appropriateness) will be compared to determine to whether the application of the method itself contributed to the predictions made, as opposed to the expertise of the evaluators themselves (Cockton and Woolrych 2001). Even though the appropriateness score may be low, this measure does not influence the thoroughness and validity score of the findings of the experts as such, it shows the relative contribution of the application of the method to the amount of hits and false positives found.

2.1 User testing

Twenty-six children participated in the UT of a computer game. The computer game for children used in this study was Milo and the magical stones (2002), an educational adventure game in which children have to help mice find magical stones on an island. More details of this game are described in Sect. 3.1.

All participating children were between 5 and 7 years old. Children participants played the computer game as they liked in a 30-min session, as previous research has shown that providing tasks severely influences how children play computer games (Barendregt et al. 2003). Observer Pro (Noldus 2002), a software package for observational research, was used for coding the video data in order to detect problems. With this software observations can be logged in the digital video data by clicking the appropriate behavioural category as defined in a coding scheme for breakdown indications. The result of this stage of analysis is a list of time stamps combined

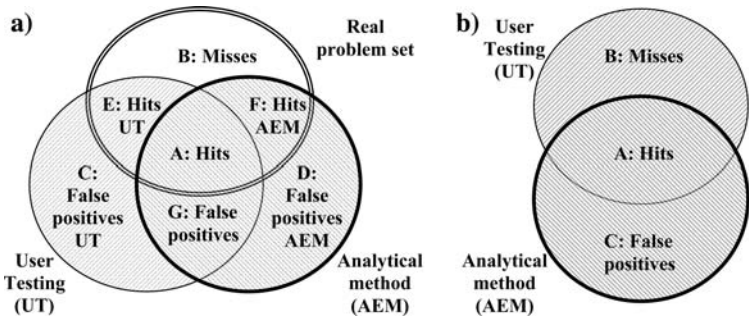

Fig. 2 **a** Overview of the terminology used when comparing outcomes of different evaluation methods, such as User Testing (UT) and an Analytical Evaluation Method (AEM), to a real problem set, and **b** when comparing the outcome of an AEM to UT

with a behavioural category, the breakdown indications. An example of a breakdown indication is (0.00.13.36, Quit), meaning that at 13 s and 36 ms the child quit the sub game. This can be a breakdown indication for a usability problem, e.g. when the child quit because the intended goal is unclear, or an indication of a fun problem, e.g. when she quits because the challenge of the game is too high. Breakdown indications were subsequently grouped and interpreted. Because of concern for the evaluator effect, eight out of the 26 user tests were analysed by two of the authors. Furthermore, the two evaluators discussed all problems extensively. The data analysis of the user test resulted in a list of 49 problems. For more information about the study set-up and data analysis, see Barendregt et al. (in press).

2.2 Thoroughness

The first performance metric, thoroughness, measures the proportion of real problems identified by an analytical method (Hartson et al. 2001; Sears 1997). Real problems are the problems approximated in our study to those found by UT.

$$\text{Thoroughness} = \frac{\text{number of real problems predicted}}{\text{number of real problems}} \quad (1)$$

In other words: thoroughness is the relation between the number of Hits (problems uncovered with UT that have been predicted by evaluators) and the total standard problem set (problems identified with UT).

2.3 Validity

Validity is the number of Hits divided by the total number of predictions from an evaluator. Validity measures the proportion of the problems identified by an analytical method that are real problems (Hartson et al. 2001; Sears 1997).

$$\text{Validity} = \frac{\text{number of real problems predicted}}{\text{number of problems predicted}} \quad (2)$$

2.4 Appropriateness

Analytical methods should support evaluators to predict problems by asking relevant questions or providing relevant guidelines. To measure to what extent the heuristics in the Combined HE and the questions in SEEM supported the evaluators in predicting problems, the appropriateness of both methods will be compared (Cockton and Woolrych 2001). The appropriateness is the percentage of correctly applied questions or heuristics. While thoroughness and validity do not take into account whether the problems were uncovered through the appropriate application of the method, the appropriateness measure provides an indication of the evaluators' understanding of the method.

However, evaluators are allowed to give more than one question or heuristic as an explanation of a problem. Evaluators can assign multiple questions or heuristics that are all correct, they can assign multiple questions or heuristics that are all incorrect, or they can assign multiple questions or heuristics that are partly correct and partly incorrect. Therefore, we will determine the appropriateness in two different ways. First, the percentages of correctly applied questions and heuristics are calculated by approving only questions and heuristics which are all applied correctly. Second, the percentages are calculated by approving sets of questions and heuristics, respectively that are partly correct.

3 The comparative study

3.1 The computer game

The computer game for children used in this study was the Dutch version of Milo and the magical stones (2002), from now on referred to as Milo. At several places in the game the children can find magical stones, but they cannot take them without playing a game (they have to earn the magical stones by completing different sub games).

An example of a screen shot of Milo can be seen in Fig. 3. The purpose of this sub game is that children find and click two crabs that make the same noise. Other relevant navigational elements on the screen are the stone in

Fig. 3 Screen shot from one of the sub games from Milo and the Magical Stones (©MediaMix Benelux)

the bottom-left corner (represents a map for navigation), the butterfly in the right-hand corner (for quitting the game) and the two arrow-sticks (one in the center and one at the left side of the screen).

The computer game is intended for children aged four to eight. Milo contains ten sub games, two navigational screens, three story screens, one help-screen and one screen for stopping the game. Among the sub games there are motor-skill games, logical games and creative screens. Many problems were anticipated for children playing the computer game, because even adult researchers had some difficulties while playing the games. This makes the computer game suitable for the experiment.

3.2 Participants

In sum 19 students from our department participated in the test. Eight students were freshman, six were second year students and five were third year students. A between-subjects design was used. Nine students evaluated Milo with the help of SEEM and ten students evaluated Milo with the help of the Combined HE. To determine whether the students are good enough experts, the outcome of the predictions will be compared to how well experts in a previous study predicted.

3.3 Procedure

All participants were given an introductory lecture to the field of inspection-based evaluation methods. They received a written manual about either SEEM or the Combined HE depending on which method they had to use, so each participant was taught only one method in detail. The SEEM manual explained the theoretical basis of the method. Furthermore it contained an explanation of SEEM's questions and a description of the procedure for applying the questions. Evaluators applying SEEM were asked to systematically check SEEM's questions for all possible actions on the screens to be evaluated. The Combined HE manual contained an explanation of Nielsen's usability heuristics (Nielsen 1994) and Malone and Lepper's concepts of fun (Malone and Lepper 1987) and a description of the procedure for applying the two heuristic sets. The evaluators applying the Combined HE approach were asked to consider both sets of heuristics when playing the sub games on the screens to be evaluated. In comparison to the SEEM approach the Combined HE approach allowed a more free-style exploration of each screen to uncover possible violations of the heuristics. The manuals contained many corresponding examples of problems that children had encountered during game play of other computer games and it also explained the problem report format. The format for the Inspection Problem Report (IPR) is based on Lavery et al. (1997). Evaluators had to fill in the screen number, the heuristic or SEEM's question the problem referred to, a short problem description, expected causes of the problem and expected outcomes of the problem. The reports from UT of Milo were written according to the same structure. By teaching and constraining evaluators to use the IPR format, the comparison of their predictions to the problems uncovered with UT became easier. Nevertheless, the main purpose of the IPR was to assist evaluators in reporting problems accurately and thoroughly.

The participants were given a training of an hour and a half in which they evaluated another computer game for young children. This game was the Dutch version of Rainbow, the most beautiful fish of the ocean (2002), an adventure game developed for children from 3 to 7 years old. The training contained a class demonstration of the method with examples of the problems uncovered with UT of Rainbow. Participants were instructed to go through the questions or heuristics at least once per screen. Then the participants had to evaluate two sub games of Rainbow in pairs. The reported problems were discussed in class and compared to the problem list from these sub games obtained from UT.

One week later participants were provided feedback on their IPR's from the training session in a session of 30 min. This feedback addressed how evaluators filled in the IPR, e.g. whether consistent mistakes were made concerning the use of the questions or the description of the problem predictions. After receiving the feedback all participants evaluated Milo for an hour and a half. Participants were not given any tasks, however there were some sub games that they were obliged to evaluate. These sub games were marked on an overview with screenshots of the sub games that was handed out to the participants. The authors were all present during the evaluation, so if they noticed a participant was spending time on a sub game that was not obligatory, they mentioned to the person it would be best if he or she would evaluate another part of the game. The obligatory sub games were selected because about 50% of the children played these sub games during UT and because these sub games contained many uncovered problems. Participants had to evaluate at least all six obligatory sub games during the evaluation.

3.4 Analysis of the data

Two of the authors collaboratively judged whether a prediction from an evaluator matched with a problem obtained from UT, resulting in a Hit. An example of a Hit in Milo is the following: in one sub game Milo has to cross a lake by jumping on water lilies to get to a toad. The toad is only letting Milo pass if he catches some flies for the toad. The

explanation for the children of how they should catch the flies is: 'Stand behind the flies and jump when they are down'. For 13 children this explanation was not enough, they did not understand how to catch the flies (they had to jump on a water lily and when a fly was in front of them they had to jump to the next water lily). This problem was predicted by five evaluators that evaluated Milo with SEEM and by eight evaluators that applied the Combined HE.

A Miss is a problem uncovered with UT which has not been predicted by evaluators. An example of a Miss in the same sub game of Milo is a problem that two children experienced. Once these children had caught a fly, they did not know how to give the fly to the toad. They were supposed to jump to the shore where the toad was and then wait until the toad grabbed the fly from Milo. These children tried to drag the fly to the toad as soon as they jumped on the shore.

Predictions from evaluators that could not be matched to one of the problems from the standard problem set were classified as False Positives. An example of a False Positive in the same sub game is the following: a challenge children have to overcome is that there are fish that should be clicked away because the fish are intended to make Milo fall. Two evaluators who used SEEM and three evaluators who used the Combined HE predicted that it was not clear to children that the fish could be chased away by clicking at them. Since none of the children explicitly indicated this during UT and some children actually did click at the fish in order to chase them away, this problem prediction was judged to be a False Positive.

The same two people judged whether the predictions from evaluators were correctly linked to either one of the questions or the heuristics. The combination of a question or heuristic and a problem prediction was correct when the number of a question or heuristic was the right fit according to the judges. Also when evaluators filled in a number that was not corresponding to our first choice but very well possible in relation to the problem, this was counted as correct. The combination was counted as incorrect, when either the wrong question or heuristic was used or in some rare cases when the evaluator had not filled in any number at all.

4 Results

The analyses are based on the data related to the obligatory sub games since evaluators had to give these screens the most attention. Figure 4 shows an overview of the numbers of all predicted problems per evaluation method. The number of problems uncovered with UT of these sub games is 49. Thirty of these problems were predicted by both SEEM and the Combined HE. The students applying SEEM uncovered 70 problems, did not find 11 problems uncovered by UT (misses), and predicted 32 problems not found by UT. The students applying the Combined HE approach uncovered 68 problems, did not find 14 problems uncovered by UT, and predicted 33 problems not found in UT. Thirteen problems, not found through UT, were uncovered both by students applying SEEM and those applying Combined HE.

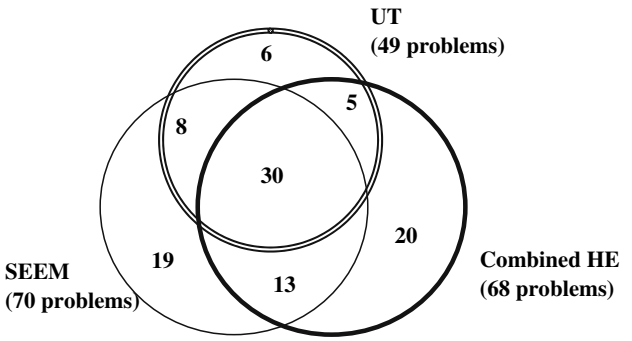

Fig. 4 Overview of number of problems predicted per evaluation method, with the *overlap in the centre* indicating the 30 problems found by all three methods, and the *three separate circles* indicating the number of problems found by UT, SEEM and Combined HE

4.1 Assessment of the thoroughness

The three performance metrics were determined as described in Sect. 2.2–2.4. Figure 2 shows the thoroughness of SEEM and the Combined HE. Statistical analyses were performed on the average thoroughness scores for all possible groups of two, three and four up to nine evaluators. The numbers of groups possible for one, two, three, up to nine evaluators, based on the number of evaluators that participated in the study are shown in Table 1.

The thoroughness of SEEM ranges from 0.26 for one evaluator (SD = 0.10) to 0.78 for all nine evaluators (the standard deviation cannot be calculated for nine evaluators because there were only nine evaluators that evaluated the computer game with SEEM, so there is only one thoroughness score). The thoroughness of the Combined HE

Table 1 Number of possible groups based on the total number of evaluators per method and the number of evaluators per group

Number of evaluators	SEEM	HE
1	9	10
2	36	45
3	84	120
4	126	210
5	126	252
6	84	210
7	36	120
8	9	45
9	1	10

starts at about the same value (0.25 for one evaluator, SD = 0.11) but ends at a lower value (0.70 for groups of nine evaluators, SD = 0.02). This is illustrated in Table 2.

This table shows that the thoroughness of SEEM is higher than the thoroughness of the Combined HE (with groups of three evaluators the difference is already statistically significant with two-tailed testing; $p < 0.01$), and as more evaluators are included the difference increases.

4.2 Assessment of the validity

The values of the validity of SEEM and the Combined HE were computed for all possible groups of two to nine evaluators (see Table 3 for an overview).

The validity of SEEM begins with an average value of 0.68 for one evaluator (SD = 0.13) while the validity of the Combined HE is slightly lower, it begins with a validity of 0.59 for one evaluator (SD = 0.10). When comparing the validity of groups of nine evaluators, SEEM's validity is 0.54 (once again the standard deviation cannot be determined because there were only nine evaluators that evaluated the computer game with SEEM meaning that there is only one validity score when looking at nine evaluators) and the validity of the Combined HE is 0.52 (SD = 0.01). So in terms of validity, SEEM scores higher than the Combined HE. From groups of two evaluators up to seven evaluators according to a two-tailed test the difference is statistically significant ($p \leq 0.001$), and for eight evaluators it is significantly different at a $p \leq 0.01$.

4.3 Assessment of the appropriateness

The appropriateness is calculated as an average score of all overlapping predictions from evaluators (both Hits and False Positives) to make a more valid comparison. The first appropriateness measure in which all questions and heuristics given by the evaluator should be correct was 75% for SEEM and 35.9% for the Combined HE. The second appropriateness measure in which only a part of the questions or heuristics given by the evaluator should be correct was 90% for SEEM and 39% for the Combined HE. Evaluators using SEEM were thus better able to assign a correct question to a problem than the evaluators using Combined HE were able to assign a correct heuristic to a problem.

Table 2 Thoroughness scores and *t*-test outcome of group sizes 1 to 9 of experts applying SEEM and Combined HE (C-HE)

Group size	SEEM	Mean (st.dev, *N*)	C-HE	Mean (st.dev *N*)	*df*	*t*
1	0.26	(0.10, 9)	0.25	(0.11, 10)	17	246
2	0.41	(0.10, 36)	0.39	(0.09, 45)	79	1,152
3	0.51	(0.09, 84)	0.48	(0.08, 120)	202	2,830**
4	0.58	(0.08, 126)	0.54	(0.07, 210)	334	5,036***
5	0.64	(0.07, 126)	0.59	(0.05, 252)	376	7,258***
6	0.68	(0.06, 84)	0.62	(0.04, 210)	292	8,729***
7	0.72	(0.05, 36)	0.65	(0.03, 120)	154	8,768***
8	0.75	(0.03, 9)	0.68	(0.02, 45)	52	7,260***
9	0.78		0.70	(0.02, 10)	9	4,593**

** $p < 0.01$, two-tailed

*** $p < 0.001$, two-tailed

Table 3 Validity scores and *t*-test outcome for group sizes 1 to 9 of experts applying SEEM and Combined HE

Group size	SEEM	Mean (st.dev, N)	C-HE	Mean (St. dev N)	*df*	*t*
1	0.68	(0.13, 9)	0.59	(0,10, 10)	17	1,764
2	0.64	(0.09, 36)	0.58	(0.05, 45)	79	3,638***
3	0.62	(0.07, 84)	0.58	(0.04, 120)	202	5,606***
4	0.60	(0.06, 126)	0.57	(0.03, 210)	334	6,973***
5	0.58	(0.05, 126)	0.56	(0.03, 252)	376	7,387***
6	0.57	(0.04, 84)	0.55	(0.02, 210)	292	6,752***
7	0.56	(0.03, 36)	0.54	(0.02, 120)	154	5,288***
8	0.55	(0.02, 9)	0.53	(0.02, 45)	52	3,470**
9	0.54		0.52	(0.01, 10)	9	1,668

** $p < 0.01$, two-tailed

*** $p < .001$, two-tailed

4.4 Student participants as experts

The suitability of using students as participants is determined by comparing the thoroughness and validity scores of the students applying SEEM in this study with the scores of the experts applying SEEM in one of our previous studies (Baauw et al. 2005). The average thoroughness of the experts and the students is 0.82 and 0.78, respectively. The average validity of the experts and the students is 0.56 and 0.54, respectively. This comparison can only lead to tentative conclusions because of the differences between the two studies, e.g. the experts used an earlier version of SEEM. Based on these results we assume that the students were knowledgeable enough to function as participants in our study to assess SEEM.

5 Discussion

5.1 Comparison with other studies

It is difficult to compare outcomes of different method comparison studies, because they vary on so many aspects of the study set-up and analysis approach. However, to give

a global impression of the relative quality of our findings, the results from this study will be compared to the results of other studies. Of the various studies that have compared CWs and HE (e.g. Cuomo and Bowen 1994; Jeffries et al. 1991; Sears 1997) we discuss the study by Sears (1997) because the comparison is based on the evaluation by a larger number of evaluators per evaluation method and the outcome is compared to the outcome of UT. We discuss our results on appropriateness in relation to the study by Cockton et al. (2003). Sears compared three AEMs (HE, CW and Heuristic Walkthrough); only the results concerning HE and CW will be discussed here (because these methods most resemble the methods compared in our study). Sears found that CW was less thorough than HE. Our results show that the thoroughness of SEEM is slightly higher than the thoroughness of the Combined HE, which is not in line with Sears' result. This could be due to the many differences between Sears' study and our study, e.g. evaluating computer games (end product) for children versus evaluating a paper design document of a product for adults (early in the interface development process). The lower number of problems found with CW in Sears' study was due to the limited number of less serious problems found by evaluators applying CW. Sears explained this result by stating that the HE approach allows evaluators to explore the interface freely to look for additional problems while the CW approach does not allow a free-form evaluation. In our study the opportunities for free exploration were fairly limited, because the experts focused on a limited set of screens, thus limiting the chance of evaluators applying the Combined HE to uncover extra problems. Another explanation for the difference between our and Sears' results, is that SEEM's questions, have been specially developed for children's computer games, so they are supposed to cover most aspects of computer games that cause problems including the less serious ones, whereas CW is based on a more general model applicable to a wider range of products.

Sears found that HE scored lower on validity than CW did, due to the number of False Positives. False Positives are the predictions from evaluators that were not uncovered in the User Test. Evaluators using the HE may focus their attention on issues of less importance to users (because HE provides less guidance than CW), resulting in an increased number of False Positives for the HE. In our study SEEM scores higher than Combined HE in terms of validity, which is in line with the results found by Sears.

Cockton et al. (2003) conducted a HE with many participants (thirty-one analysts divided over ten groups in the latest study compared to 96 analysts divided over 16 groups in an earlier version of the study). They found appropriateness scores for the HE of 31 and 57%. In the present study, the appropriateness for the Combined HE varies from 35.9 to 39%, which indicates that the understanding of the Combined HE was similar to the understanding of HE in the first study by Cockton et al. (2003), but worse than in their second study (taking note of the difficulty of comparing evaluation methods assessed in different studies, this can only be seen as an indication). The appropriateness score for SEEM's questions in an earlier study was 74% (Baauw et al. 2005). The appropriateness in the present study was higher for SEEM (90%), which indicates that the understanding of SEEM has improved. This might be due to the fact that some of SEEM's questions were changed after the first study.

Note that the descriptions of heuristics are relatively vague and thus have a wider scope, as compared to the descriptions of SEEM's questions. Because of the differences in the way the heuristics and SEEM's questions are described, usually many more heuristics can be matched to one problem description, whereas usually only one question can be matched to a problem description. Thus, it is usually much clearer, whether evaluators have applied SEEM correctly, than whether the Combined HE approach was applied correctly.

5.2 Reassessment of the False Positives

As mentioned earlier in this article, analytical and empirical methods are complementary (Baauw et al. 2005; Chattratichart and Brodie 2004). This means that some problems that were predicted by evaluators and were not found during UT might very well be true. Thus, problems coded as False Positives (sector C in Fig. 2b) can in fact be either hits by the AEM only (sector F in Fig. 2a; now to be called Complementary Hits) or False Positives of an AEM only (sector D in Fig. 2a; now to be called True False Alarms). False Positives that were clearly no problems for any of the children in the UT and problem descriptions based on incorrect assumptions by the expert about the game are judged to be True False Alarms. False Positives related to problems that children are not inclined to verbalize and where children's behaviour is difficult to interpret and related to suggestions for improvements to the game were judged to be Complementary Hits.

The same researchers that determined whether predictions from experts matched with problems uncovered with UT reassessed the False Positives to determine the likelihood of them being real problems, i.e. Complementary Hits or in other words a real problem that has not been found during UT. An example of a Complementary Hit is the following: in one sub game of 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. Many evaluators predicted

that it would be more fun when children could use a tactic to solve this sub game (12 evaluators predicted this Complementary Hit, four evaluated the computer game with SEEM and eight with the Combined HE). However, none of the children explicitly indicated this. Thus, while this problem was not obtained from UT, it could very well be true.

However other problem predictions that were originally determined to be False Positives were not judged to be Complementary Hits, but as incorrect predictions (True False Alarms). For example, one expert applying SEEM had insufficient knowledge about the game and mentioned that it was a problem that one sub game had to played, before being able to continue the game. However, since this is not the case, this problem is a True False Alarm.

Table 4 shows the number of False Positives, Complementary Hits and True False Alarms of all evaluators. This table shows that most False Positives are Complementary Hits. Overall, the two methods are very similar in the sense that they each predict a similar amount of new problems or Complementary Hits (respectively 25 problems with SEEM and 24 with the Combined HE) and only a few problems that are unlikely to be experienced by real users (True False Alarms). These results are tentative since the approach for coding problems into these categories needs to be refined further.

5.3 Study set-up

Another topic for discussion is the influence of using students as experts on the outcome in a study such as this. Our experience in previous studies has been that people become better able to predict problems correctly, if they have a basic understanding of user–system interaction issues, if they learn to work systematically and if they receive specific feedback on a first attempt at predicting problems. A comparison was made of the quality of the predictions of the students participating in this study with the experts participating in the previous study (see Baauw et al. 2005). The comparison showed that the scores for validity and thoroughness were very similar for the students and the experts, indicating that the students had enough usability expertise and were trained well enough for the purpose of this study to function as an ‘expert’.

Table 4 Outcome of the reassessment of the False Positives into Complementary Hits and True False Alarms

	False Positives	Complementary Hits	True False Alarms
SEEM	32	25	7
The Combined HE	33	24	9

5.4 Generalization

The results described in this paper should be looked at with a few aspects in mind. The first is that this comparison has been made with the help of only one educational adventure game. It cannot be said with certainty that the results will still hold if the experiment would be conducted again with another kind of game. However, we think the game we used in this study is a good representative of an adventure game. In the first study with SEEM (Baauw et al. 2005) two different educational adventure games were evaluated. One of the computer games was the same as the one used in this study (Milo). The other game, which was the Dutch version of Roger Rabbit: Fun in the Clouds (2003), focused more on education than Milo. Together they covered a wide range of activities that are often presented in educational adventure games for children, like motor-skill games and cognitive challenges. Therefore the combination of these two games was a good representative of the genre. The results on thoroughness, validity and the understanding of SEEM showed similar trends for the two games. Therefore it is likely that the results as described in this paper can be generalized to a variety of educational adventure games.

The second point regarding the generalization is the fact that all evaluations described in this paper deal with children between 5 and 7 years old. Therefore it cannot be said with certainty that the results will also apply for older children. Another study, in which SEEM was applied to computer games for older children led to similar results on thoroughness, validity and appropriateness scores (Baauw et al. 2006).

It is possible that another set of heuristics generates other results. The reason why we did not choose another, perhaps more recent, set of heuristics is that the sets we used have undergone extensive testing and several iterations of design. Other sets with heuristics have been used much less extensively and therefore may need further work before they are ideal for use in a (Combined) HE.

Another issue that might be pursued in the future is whether SEEM might be used by children evaluators. We initially developed SEEM for adult users, assuming that the people applying an analytical method need to have at least human-computer interaction expertise. Furthermore, we assumed that, especially young children, have not yet developed their cognitive abilities enough to be able to reflect on their behaviour to uncover and analyse the problems they might encounter. Other studies have made a start with examining whether children can apply the HE approach (MacFarlane and Pasiali 2005). In a similar vein, it might be possible to have older children (e.g. 12–14 year olds) evaluate younger children’s games, based on the assumption that it is easier for slightly older children to

reflect on the games and that they are more closely in tune with younger children than adult evaluators. This would require that SEEM's question and SEEM's manual be adapted to younger users. It will also be interesting to determine whether SEEM can also be applied to products for children other than computer games.

6 Practical advice

Based on the findings of this study we provide some advice for practitioners. Since the outcome of UT and AEMs is complementary, it is always good to combine these approaches during the development of computer games. Furthermore, the findings on thoroughness indicate that in our study nine experts find 70 and 78% applying Combined HE and SEEM, respectively. This figure is lower than the percentage of problems found by experts applying HE as provided by Nielsen (1994). As a consequence when evaluating computer games for children we would advice practitioners to involve at least eight to nine experts to predict problems using SEEM or combine HE. Such experts need to receive a thorough training on the method including at least one prediction session and providing them with feedback on the thoroughness, validity and appropriateness of their predictions.

7 Conclusion

We described a comparative study assessing SEEM's effectiveness and appropriateness compared to a Combined HE, which is a method based on a combination of Nielsen's HE (Nielsen 1994) and the fun-related concepts from Malone and Lepper (1987). The results show that the thoroughness of SEEM is higher than the thoroughness of the Combined HE. In other words the overlap with UT is higher for SEEM than for the Combined HE. Also the validity of SEEM is higher than the validity of the Combined HE, meaning that the proportion of problem predictions that are real problems is better for SEEM than it is for the Combined HE. The appropriateness, or correct use of the questions or the heuristics, is much higher for SEEM than it is for the Combined HE. This indicates that SEEM gave evaluators more guidance when predicting problems than the Combined HE did. Furthermore, the results show that UT finds problems not uncovered by the predictive approaches, and the predictive approaches find problems not uncovered by UT. This indicates that ideally UT and predictive approaches should be combined in practice.

Overall, the study has shown that SEEM's walkthrough approach compares favourably to the heuristic-based approach on the effectiveness scores of thoroughness, validity and appropriateness.

Acknowledgements We thank all students who participated in the study. We also thank Arnold Vermeeren from the Technical University of Delft (the Netherlands) 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. Even though we believe that the term Usability Testing touches the essence of the test better than the term UT, in this paper the term UT will be used to make the distinction with the AEMs clearer.

References

Baauw E, Bekker MM, Barendregt W (2005) A structured expert evaluation method for the evaluation of children's computer games. Proceedings of human–computer interaction INTERACT 2005, Rome, 12–16 September 2005, pp 457–469

Baauw E, Bekker MM, Markopoulos P (2006) Assessing the applicability of the structured expert evaluation method (SEEM) for a wider age group. Proceedings of the the interaction design and children conference, Tampere, 6–9 June 2006, pp 73–80

Barendregt W, Bekker MM (2004) Towards a framework for design guidelines for young children's computer games. Proceedings of the 2004 ICEC conference, Springer, Eindhoven, September 2004, pp 365–376

Barendregt W, Bekker MM (2006) Developing a coding scheme for detecting usability and fun problems in computer games for young children. Behav Res Meth 38(3):382–389

Barendregt W, Bekker MM, Speerstra M (2003) Empirical evaluation of usability and fun in computer games for children. Proceedings of human–computer interaction INTERACT '03, IOP Press, Zürich, pp 705–708

Barendregt W, Bekker MM, Bouwhuis D, Baauw E (in press) Predicting effectiveness of children participants in user testing based on personality characteristics. Behav Inf Technol

Bekker MM, Barendregt W, Crombeen S, Biesheuvel M (2004) Evaluating usability and fun during initial and extended use of children's computer games. In: Fincher S, Markopoulos P, Moore D, Ruddle R (eds) Proceedings of the BCS-HCI, people and computers XVIII–design for life, Leeds, Springer, September 2004, pp 331–345

Chattratichart J, Brodie J (2004) Applying user testing data to UEM performance metrics. Late breaking results paper, CHI extended abstracts on human factors in computing systems 2004, Vienna, pp 1119–1122

Cockton G, Woolrych A (2001) Understanding inspection methods: lessons from an assessment of heuristic evaluation. In: Blandford A, Vanderdonckt J, Gray PD (eds) People and computers. Springer, Berlin, pp 171–192

Cockton G, Lavery D, Woolrych A (2003) Inspection-based evaluations. In: Jacko J, Sears A (eds) The human–computer interaction handbook: fundamentals, evolving technologies and emerging applications, Lawrence and Erlbaum Associates, Mahwah, pp 1118–1138

Cockton G, Woolrych A, Hall L, Hindmarch M (2003) Changing analysts' tunes: the surprising impact of a new instrument for usability inspection method assessment. In: Palanque P, Johnson P, O'Neill E (eds) People and computers, designing for society (Proceedings of HCI 2003). Springer, Berlin, pp 145–162

Cuomo DL, Bowen CD (1994) Understanding usability issues addressed by three user-system interface evaluation techniques. Interact Comput 6(1):86–108

Desurvire H, Caplan M, Toth JA (2004) Using heuristics to evaluate the playability of games. CHI extended abstracts on human factors in computing systems 2004, Vienna, pp 1509–1512

Fabricatore C, Nussbaum M, Rosas R (2002) Playability in action videogames: a qualitative design model. Hum Comput Interact 17(4):311–368

Federoff MA (2002) Heuristics and usability guidelines for the creation and evaluation of fun in video games. M.Sc. thesis, Department of Telecommunications of Indiana University

Gilutz S, Nielsen J (2002) Usability of websites for children: 70 design guidelines based on usability studies with kids. Nielsen-NormanGroup, Fremont

Hartson HR, Andre TS, Williges RC (2001) Criteria for evaluating usability evaluation methods. Int J Hum Comput Interact 13(4):373–410

Jeffries R, Miller JR, Wharton C, Uyeda KM (1991) User interface evaluation in the real world: a comparison of four techniques. In: Proceedings of CHI 1991, pp 119–124

Lavery D, Cockton G, Atkinson MP (1997) Comparison of evaluation methods using structured usability problem reports. Behav Inf Technol 16(4):246–266

MacFarlane S, Pasiali A (2005) Adapting the heuristic evaluation method for use with children. In: Workshop on child computer interaction: methodological research, Interact 2005, Rome, pp 28–31

Malone TW (1980) What makes things fun to learn? A study of intrinsically motivating computer games. Technical Report CIS-7, Xerox PARC, Palo Alto

Malone TW, Lepper MR (1987) Making learning fun: a taxonomy of intrinsic motivations for learning. In: Snow RE, Farr MJ (eds) Aptitude, learning and interaction III cognitive and affective process analysis. Erlbaum, Hillsdale

Milo and the magical stones (2002) (in Dutch: Max en de toverstenen), Computer software, MediaMix Benelux (2002)

Nielsen J (1994) Heuristic evaluation. In: Nielsen J, Mack RL (eds) Usability inspection methods. Wiley, New York, pp 25–62

Noldus (2002) The observer PRO, Computer software, Noldus (2002)

Norman DA (1998) The design of everyday things. MIT Press, London

Pagulayan RJ, Keeker K, Wixon D, Romero R, Fuller T (2003) User-centered design in games. In: Jacko J, Sears A (eds) Handbook for human–computer interaction in interactive systems. Lawrence Erlbaum Associates, Mahwah, pp 883–906

Rainbow, the most beautiful fish in the ocean (in Dutch: Regenboog, de mooiste vis van de zee), Computer software, MediaMix Benelux (2002)

Roger Rabbit, Group 3: Fun in the Clouds (in Dutch: Robbie Konijn, Groep 3: Pret in de Wolken), Computer software, Mindscape (2003)

Sears A (1997) Heuristic walkthroughs: finding the problems without the noise. Int J Hum Comput Interact 9(3):213–234

Wharton C, Rieman J, Lewis C, Polson P (1994) The cognitive walkthrough: a practitioner's guide. In: Nielsen J, Mack RL (eds) Usability inspection methods. Wiley, New York, pp 105–140