

# Assessing the Applicability of the Structured Expert Evaluation Method (SEEM) for a wider Age Group

Ester Baauw, Mathilde M. Bekker and Panos Markopoulos

Eindhoven University of Technology

Department of Industrial Design

P.O. Box 513

5600 MB Eindhoven, The Netherlands

+31 40 247 5239

esterbaauw@hotmail.com, M.M.Bekker@tue.nl, P.Markopoulos@tue.nl

## ABSTRACT

This paper describes a study which examines whether a predictive evaluation method (SEEM) is suitable to assess products for an older age group than for which the evaluation method was originally developed. SEEM stands for Structured Expert Evaluation Method and is an analytical evaluation method especially developed for assessing the fun and usability of young children's educational computer games (children from 5 to 7 years old). In the present study SEEM was applied to assess educational computer games for children between 9 and 11 years old. Outcomes on scores for thoroughness (whether SEEM finds all problems), validity (whether SEEM makes predictions that are likely to be true) and appropriateness (whether SEEM is applied correctly) were compared. The results show that the trends for the thoroughness and the validity are the same for the two different age groups; however SEEM scores a bit better for the oldest age group. The appropriateness scores are about the same for the two age groups. The results indicate that SEEM can also be applied for assessing educational computer games for children between 9 and 11 years old.

## Keywords

Analytical evaluation methods, children, educational computer games, fun, SEEM, usability

## INTRODUCTION

This paper describes an experimental study where an analytical evaluation method called SEEM is assessed by comparing it to empirical evaluation through usability testing of a game for children.

SEEM stands for Structured Expert Evaluation Method, meaning that it is a structured method that helps experts identify interaction design problems that affect the usability and the fun that can be experienced in the use of computer

games by young children. SEEM has been especially developed for the evaluation of educational computer games for children aged 5 to 7 [5].

SEEM requires that experts, also called evaluators, examine subparts of the computer game, e.g. different screens, addressing a checklist with questions. The questions are originally based on Norman's theory of action model [17] and on Malone's concepts of fun [14]. Each of SEEM's questions corresponds to the various phases of Norman's action cycle (see Figure 1), e.g. asking the evaluator to determine whether the goal can be perceived, understood, and whether the goal will be fun. The fun-aspect was added to the action cycle since this is an important aspect of children's computer games. Fun or enjoyment is one of the major motivations for children to interact with technology [12]. A separate question about navigational issues (transition from one screen to the next) was added (see Figure 1). Evaluators go through the checklist and when they expect a problem to occur, they record the predicted problem in a problem report template.

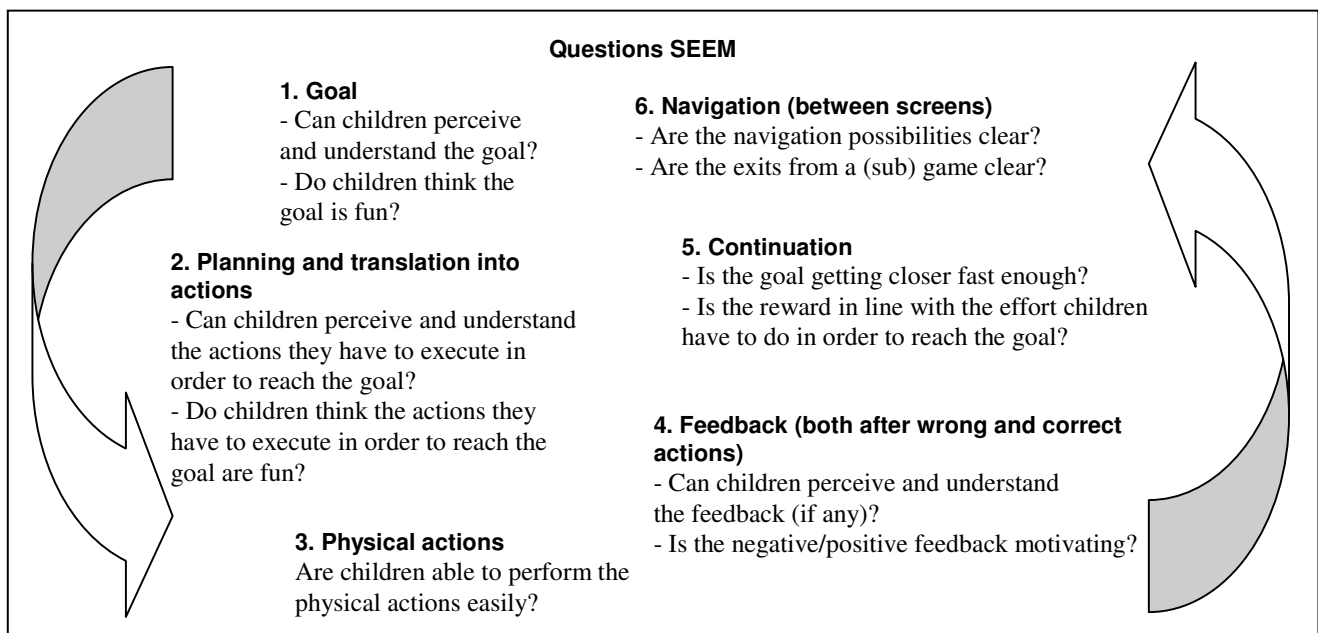
## Assessing the Quality of SEEM

Two previous studies have been conducted to assess SEEM's quality. Eighteen experts participated in the first study [5]. 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 to 7). They predicted 76% of the problems uncovered with User Testing (UT) of the same game.<sup>1</sup> Unfortunately, experts also predicted many problems that were not found during UT. Based on these findings an improved version of SEEM was created.

The second study involved a comparison between SEEM and another analytical evaluation method [4]. To make a valid comparison the alternative method had to address

---

<sup>1</sup> The term User Testing is used in this paper instead of Usability Testing to avoid any confusion between the analytical method (SEEM) and the empirical evaluation method (UT)



**Figure 1** The questions of SEEM, which have to be checked at each screen of a computer game

both usability and fun (since SEEM also covers both aspects). Therefore we compared SEEM to Nielsen's Heuristic Evaluation [16], where Nielsen's set of heuristics was extended with the fun-related guidelines by Malone and Lepper [15]. We called this method the Combined Heuristic Evaluation (Combined HE). In this study nine students from our department evaluated an educational computer game for young children (aged 5 to 7) with SEEM. The computer game in question was the Dutch version of Milo and the Magical Stones [2]. Ten other students evaluated the computer game with the Combined HE. The results showed that the overlap with UT was higher for SEEM than for the Combined HE. Also the proportion of problems that are likely to be true (related to all problem predictions from evaluators) was higher for SEEM than it was for the Combined HE. The correct use of the questions or the heuristics was slightly higher for the Combined HE than it was for SEEM. However evaluators who used the Combined HE filled in more different heuristics next to a problem prediction than evaluators that applied SEEM filled in different questions next to a problem prediction. This indicates that SEEM gave evaluators more guidance when predicting problems. Overall, the study has shown SEEM's walkthrough approach compared favorably to the heuristic-based approach on the aspects described above.

The present paper describes a third study we conducted to determine whether SEEM can also be applied for assessing educational computer games for another age group. The target age group used in the first studies with SEEM was 5 to 7 years, which is the intuitive phase (the upper half) from what Piaget calls the preoperational period [18]. For this study we wanted to test SEEM on computer games for Piaget's next developmental stage, the concrete operational

stage. Children in this stage vary from 7 to 11 years. To be consistent again the upper part of this stage was chosen (children from 9 to 11). The set-up of the study was comparable to the setup from the second study [4] to allow us to compare the application of SEEM for educational computer games targeting these two different age groups. The evaluators that applied SEEM in the second study are the first group for the present comparative study; so part of the data from the former study is reused for the present study.

## RESEARCH METHOD

### The educational Computer Games

The computer game for the younger children was Milo and the Magical Stones [2], from now on referred to as Milo. Milo is an adventure game in which children have to help Milo and two mice friends to find magical stones on an island. These magical stones keep the mice warm during the winter, and the other mice say that Milo and his friends are their only hope. The game is developed for children aged 4 to 8. Milo contains ten sub games, two navigational screens, three story screens, one help-screen and one stop-screen. Among the sub games there are e.g. motor-skill games, logical games and creative screens.

The computer game for older children used in this study was Bio Mania [1]. Bio Mania is an educational adventure game that teaches children biological aspects in a playful manner. The computer from professor Nuvoloni broke down and therefore his experiments are likely to fail, unless children find all missing data on the island. Children have to find seven objects (e.g. a big flying insect), take one picture (small mammals) and record two sounds on tape (e.g. quacking amphibian) in order to complete the first level. To get information from the islanders, the children

have to answer questions. The answers to the questions can be found in the encyclopedia, which is included in the computer game. The game is developed for children from 9 to 14 years old. The first level of Bio Mania contains ten sub games, two navigational screens, an encyclopedia, one screen with the materials children have to collect and one screen for stopping the game. The sub games contain two motor-skill games, four logical games and four screens where children have to collect items and answer questions.

Many problems were anticipated for children playing both computer games, because even adult researchers had some difficulties while playing the games. This makes the computer games suitable for the experiment.

### Participants

In sum twenty Industrial Design students from our department participated in the test. Group 1, consisting of nine students, evaluated Milo [2] and group 2 (eleven students) evaluated Bio Mania [1]. In Table 1 an overview of the conditions with the corresponding number of participants, computer game from the training and computer game for the actual evaluation can be seen.

**Table 1 Overview of the two conditions**

	Group 1	Group 2
Number of participants	9	11
Computer game training	Rainbow (3-7 years)	Milo (4-8 years)
Computer game actual evaluation	Milo (4-8 years)	Bio Mania (9-14 years)

### Procedure

All participants were given an introductory lecture on analytical evaluation methods. They received a written manual about SEEM. The manual contained an explanation of SEEM's questions with many corresponding examples of problems that children had encountered when playing other educational computer games. The manual also explained the format for the Inspection Problem Report (IPR), which is based on Lavery et al. [13]. The IPR requires evaluators to fill in the screen number, the number of SEEM's question the problem refers to, a short problem description, the expected causes of the problem and the expected outcomes of the problem. The IPR was made to assist evaluators in reporting problems accurately and thoroughly but also to enable the analysis of their predictions.

All participants were given a training of an hour and a half with a computer game for young children; the first group practiced with the Dutch version of the educational computer game 'Rainbow the most beautiful fish of the ocean' [3]; an adventure game developed for children from 3 to 7 years old. Their real evaluation was with Milo [2]; the adventure game for children between 4 and 8 years old (see Table 1 for an overview of the conditions). This means that the computer game they practiced with is targeted for

slightly younger children than the computer game from the actual evaluation is. The second group practiced with Milo and their real evaluation was with Bio Mania [1]; developed for children between 9 and 14 years old. So the computer game of the training from the second group differed with the computer game from the first group. If we would have let the evaluators from the second group practice SEEM with the computer game Rainbow, we think the gap between the training and the actual evaluation would be bigger than it was for the evaluators from the other condition (concerning the target age). In an ideal situation the training would have been performed with another educational computer game for older children (about 8 to 12 years old). Since we did not have any data from UT of such a game, we decided to let the evaluators practice with Milo as it was the most closely related game to Bio Mania in terms of the target group with available data from UT.

The training contained a class demonstration of the method with a comparison to the problems uncovered with UT of the respective computer game. Subsequently, participants were instructed to go through the questions at least once per screen. Then participants had to evaluate two sub games of their respective computer game in pairs. The reported problems were discussed in class and compared to the problem list from these sub games obtained from UT.

Before the actual evaluation started (one week after the training) participants were provided feedback on the IPR's they filled in during the training session. Aspects such as the incorrect interpretation of SEEM's questions were explained once more.

The participants of both groups evaluated their respective games for an hour and a half. An overview of the different screens was handed out to the participants before the evaluation. This overview contained screenshots of the different sub games they had to evaluate and a flow chart of those sub games.

Group 1 assessing Milo had to evaluate five sub games and one navigational screen. For the second group there were nine different screens from Bio Mania participants had to evaluate; four of them were sub games while the others were navigational screens. These screens were selected since (many of) the children that participated in the UT visited those screens. This means that the number of obligatory screens is different for the two conditions. Due to the linearity of Bio Mania it was not possible to let evaluators assess the same number of sub games (the sub games then had to be about the same size, because there is quite some variation in the effort required to complete the different sub games). So the sub games from Bio Mania were analyzed by the first author on their size and compared to the total size of the obligatory screens from Milo. After this analysis, we decided to let evaluators from group 2 (Bio Mania) assess four sub games (they were all shorter than the sub games from Milo) and five navigational screens to outweigh the differences in number of obligatory

screens. Also the numbers of problems obtained from UT for both computer games were comparable for the obligatory screens; 49 problems for Milo versus 50 problems for Bio Mania.

### PERFORMANCE METRICS

To assess the effectiveness of SEEM, the problem predictions with SEEM are compared to the problem list obtained from UT. The set of problems identified through UT is used as a benchmark to determine whether SEEM finds all problems (thoroughness), and whether SEEM makes predictions that are likely to be true (validity). Furthermore, the correct use of the questions (appropriateness) was assessed.

#### The User Tests

Twenty-six children participated in the UT of Milo. All children were between 5 and 7 years old. Children participants played Milo as they liked in a 30 minute session. Analysis of the data led to a list of 49 problems for those sub games that evaluators were obliged to assess. For more information about the study set-up and data analysis, see Barendregt et al. [7].

Twenty-four children participated in the UT of the computer game Bio Mania [6]. All children were between 9 and 11 years old. The UT was originally set up to compare two strategies for obtaining verbalization data during UT with children. The methods were think aloud (with active prompting) and post-task interview. Twelve children evaluated Bio Mania while thinking aloud and the other twelve children evaluated the computer game with post-task interview. For more details about the UT see Baauw and Markopoulos [6]. The data from the UT was reanalyzed due to more experience and renewed insights into the field of problem analysis from the first author. This analysis led to a revised list of 50 problems obtained from UT.

#### Determining the Thoroughness

The first performance metric, thoroughness, measures the proportion of real problems identified by an analytical evaluation method [11] [19]. Real problems are the problems approximated in our study to those identified by UT.

$$\text{Thoroughness} = \frac{\text{Number of Real Problems predicted}}{\text{Number of Real Problems}}$$

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).

#### Determining the 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 [11] [19].

$$\text{Validity} = \frac{\text{Number of Real Problems predicted}}{\text{Number of Problems predicted}}$$

#### Determining the Appropriateness

To measure to what extent SEEM's questions helped evaluators in predicting problems the appropriateness of SEEM will be calculated [10]. The appropriateness is the percentage of correctly applied questions. This measure is an indication of the evaluators' understanding of the method.

#### ANALYSIS OF THE DATA

All predictions from evaluators were assessed by persons involved in the research. If a prediction from an evaluator matched with a problem obtained from UT; this resulted 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 thirteen 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 5 evaluators that evaluated Milo.

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 UT were classified as False Positives. An example: 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. Four evaluators predicted that it would be more fun when children could use a tactic to solve this sub game. During UT this was not detected as a real problem.

Predictions from evaluators were also judged on the fact whether they were correctly linked to one of the questions from SEEM. The combination of a question and a problem prediction was correct when the number of a question was the right fit. Also when evaluators filled in a number that was not corresponding to our first choice but which was very well possible in relation to the problem, this was counted as correct. The combination was counted as incorrect, when either the wrong question was used or in

some rare cases when the evaluator had not filled in any number.

**EXPECTATIONS**

Concerning the thoroughness (all overlapping predictions with UT) the hypothesis is that it is easier to predict problems that older children will find in a computer game than to predict problems that young children will uncover. The reason for this assumption is that the older the children are, the easier it is for the evaluators to place themselves in their position because the target group resembles the evaluators more closely. Therefore the hypothesis is that the thoroughness score for SEEM on computer games for children aged 9 to 11 will be higher (and therefore better; more problems from UT will be predicted) than the thoroughness will be for SEEM on computer games for younger children.

For the validity the same assumption has been made: the older the children, the fewer mistakes or wrong assumptions evaluators make about the children’s abilities and behavior. Therefore the number of False Positives (predictions from evaluators that could not be matched to a problem obtained from UT) will be lower for SEEM for the older age group than it is for SEEM applied on computer games for children aged 5 to 7. So the number of real problems found will be higher for SEEM Bio Mania (children 9 to 11 years old) than it will be for SEEM Milo (5 to 7 years old); however the number of False Positives will be lower for SEEM in combination with the oldest children. Looking at the formula for the calculation of the validity, this shows that the hypothesis is that the validity of SEEM for computer games for children aged 9 to 11 is higher than the validity of SEEM for computer games for children aged 5 to 7.

Concerning the appropriateness the assumption is made that there will be no difference. The procedure of the training and the evaluation of Bio Mania with SEEM was kept as similar as possible compared to the procedure and training of the evaluators that assessed Milo with SEEM. Therefore our hypothesis is that there will be no difference in understanding of the method between the two different age groups.

**RESULTS**

Because thoroughness and validity scores are influenced by the number of evaluators we calculated these scores for all possible combinations of one to eight evaluators out of the total sets of evaluators in the two studies (respectively nine and eleven). This approach is adopted from Sears [19]. The number of possible combinations of one to eight evaluators out of the total of nine and eleven evaluators, respectively can be seen in Table 2.

**Table 2 Number of possible combinations of a sub-set of evaluators from the total set of evaluators per study**

Number of evaluators	SEEM Milo (5-7 years)	SEEM Bio Mania (9-11 years)
1	9	11
2	36	55
3	84	165
4	126	330
5	126	462
6	84	462
7	36	330
8	9	165

The three performance metrics were computed with the formulae given previously.

**Thoroughness**

The mean thoroughness scores (or overlap with UT) with the corresponding standard deviation for all groups of evaluators (up to groups of eight evaluators) can be seen in Table 3. The final column shows the levels of significance.

**Table 3 Mean thoroughness scores (with standard deviation)**

Number of evaluators	SEEM Milo (5-7 years)	SEEM Bio Mania (9-11 years)	Sig. (1-tailed)
1	0.26 (0.10)	0.28 (0.07)	.267
2	0.41 (0.10)	0.45 (0.07)	.019
3	0.51 (0.09)	0.56 (0.07)	.000*
4	0.58 (0.07)	0.64 (0.07)	.000*
5	0.64 (0.06)	0.69 (0.06)	.000*
6	0.68 (0.06)	0.73 (0.05)	.000*
7	0.72 (0.05)	0.76 (0.04)	.000*
8	0.75 (0.03)	0.78 (0.04)	.004

\*  $p < 0.001$

This table shows that despite the fact that the trend is the same for the two studies; SEEM scores a bit better for the oldest age group. This difference is statistically significant starting at groups of two evaluators with 1-tailed testing. At groups of eight evaluators the difference is still statistically significant ( $p = 0.004$ ).

**Validity**

The validity of SEEM or the proportion of problem predictions that are real problems can be seen in Table 4.

**Table 4 Mean validity scores (with standard deviation)**

Number of evaluators	SEEM Milo (5-7 years)	SEEM Bio Mania (9-11 years)	Sig. (1-tailed)
1	0.68 (0.13)	0.76 (0.08)	.066
2	0.64 (0.09)	0.74 (0.07)	.000*
3	0.62 (0.07)	0.72 (0.06)	.000*
4	0.60 (0.05)	0.70 (0.06)	.000*
5	0.58 (0.05)	0.68 (0.05)	.000*
6	0.57 (0.04)	0.67 (0.05)	.000*
7	0.56 (0.03)	0.65 (0.04)	.000*
8	0.55 (0.02)	0.63 (0.03)	.000*

\*  $p < 0.001$

This table shows that once again the trends are quite similar. However SEEM scores higher and therefore better for the oldest age group than SEEM does for the younger children, which is in line with our expectation. The difference is statistically significant from groups of two to eight evaluators ( $p < 0.001$ ).

#### Appropriateness

The appropriateness is calculated over all Hits. Table 5 shows the different appropriateness scores.

**Table 5 Appropriateness scores**

	SEEM Milo (5-7 years)	SEEM Bio Mania (9-11 years)
Correct number	97	102
Wrong number	23	23
Appropriateness	0.81	0.82

The appropriateness scores are almost exactly the same for the two age groups, which corresponds to our expectations.

## DISCUSSION

### Generalization

The results of SEEM for educational computer games for children between 9 and 11 have only been tested on one computer game. Therefore it cannot be said with certainty that the results will still hold if the experiment was conducted with another educational computer game for children from the same target age. However the results show similar trends for the two different age groups compared in this study. Milo [2], the educational computer game for children from 5 to 7 years old, has been shown to be comparable to another educational computer game for young children [5]. Since the results show similar trends it is likely that the results can be generalized to other educational computer games for the age of 9 to 11 as well.

### Differences between the two Conditions

The procedure of the current experiment has been kept as similar as possible to the procedure of the experiment in which SEEM was compared with the Combined HE in

order to make the comparison between the two conditions possible and valid. However to put the results in perspective differences in the set up of the two studies should be discussed.

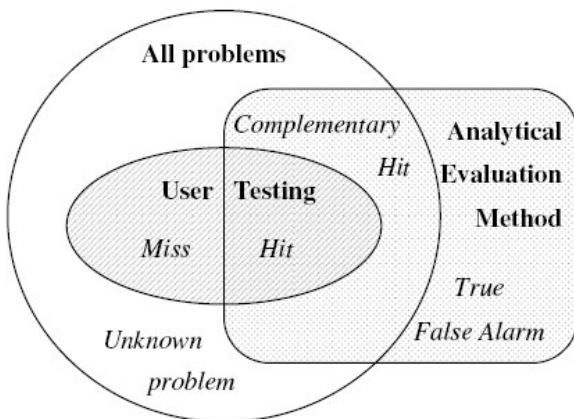
The first difference between the two studies or conditions is the way the two user tests were carried out. In the UT of Bio Mania the children were given small tasks they had to complete. In the UT of Milo children were free to play the game as they liked. Another difference concerning the procedures of the two user tests is the verbalization instructions given to the children. The young children (5 to 7 years old) were asked before the evaluation started to state as many aspects that they found problematic in the game as possible. The older children were distributed over two conditions: think aloud (with active intervention) and post-task interview. These differences cannot be denied, however we believe that this mostly makes a difference in the number of problems that have been verbalized (the older children have verbalized more problems); this does not mean that the young children found a smaller proportion of real problems. This is based on the assumption that problems will be found, whether they are verbalized or based on a wrong action made by a child. Earlier research has shown that the influence of tasks on the number of problems found is limited [8]. That study did not find any difference in verbal and non-verbal indications or number and kind of problems. So we think that the difference in procedures from the two user tests does not influence the results as they are presented in this paper.

The second difference between the two conditions was the computer game the evaluators practiced with during the training (see Table 1 for an overview of the conditions and their respective computer games). The computer game of the training for the second group (developed for children from 4 to 8 years) shows a big gap with the computer game from the actual evaluation (for children aged 9 to 14). This gap is bigger than the difference between the two computer games for the first group concerning their respective target ages (3 to 7 years and 4 to 8 years). This means that the training for the group that evaluated Bio Mania in the real evaluation (the second group) was less customized to the actual evaluation than it was for the first group. However, the results show that the second group scores even better on SEEM's scores for thoroughness and validity than the first group does. The differences in training game makes these results even more valuable; it is likely that SEEM would score even better for Bio Mania if the training would be better tuned to children between 9 and 11 years old.

All other aspects have been kept the same, e.g. evaluators got the same introduction to analytical evaluation methods, the same version of SEEM, the same manual, etc. So influences that follow from differences in the procedure have been minimized.

### Reassessment of the False Positives

Various researchers have argued that analytical evaluation methods can predict real problems not found in UT; so in that sense, analytical and empirical methods are complementary [5] [9]. This means that some problems that were predicted by evaluators and were not found during UT might very well be true. Such problems, which are judged to be False Positives when assuming UT finds all problems, will be called Complementary Hits (see Figure 2 for an overview of the terminology).



**Figure 2 Overview of terminology**

The first author, who also determined whether predictions from evaluators matched with problems uncovered with UT, reanalyzed the False Positives to determine the likelihood of them being real problems, i.e. Complementary Hits. An example of a Complementary Hit in Milo has already been given; the problem prediction described as an example of a False Positive in our opinion is a Complementary Hit. The prediction from four evaluators that it would be more fun when children could use a tactic to solve this sub game could very well be true, even though none of the children explicitly indicated this during UT. An example of a Complementary Hit in Bio Mania (to clarify further) is the prediction from three evaluators that one of the characters does not speak very audibly. Looking back at the data obtained with UT, we saw that some of the children were straining to understand the character while they did not seem to have any trouble listening to and understanding other characters. Since none of the children mentioned something about the audibility of the character nor did they ask the experimenter what the character was saying, this was not coded as a problem during the original data analysis. Thus, while this problem was not obtained from UT, it could very well be true.

True False Alarms are predictions from evaluators that are not likely to be true because they for example underestimated the children. In one sub game of Bio Mania children have to remember a route over a wooden bridge. The bridge is built of different logs (three logs in a row; a

total of eight rows); at the beginning the bridge is intact; then one by one some of the logs go down so there is a route of closed logs to follow. Then all logs close again and children have to click at the route they just saw. An example of a True False Alarm is the prediction from one evaluator that children would not know which logs made the path: the open logs or the closed ones. This was counted as a True False Alarm because all children that participated in the UT knew which logs formed the route they had to remember, e.g. none of them started with clicking at a log that opened earlier.

Table 6 shows the number of False Positives, Complementary Hits and True False Alarms of all evaluators.

**Table 6 Outcome of the reassessment of the False Positives (with the percentages)**

	SEEM Milo (5-7 years)	SEEM Bio Mania (9-11 years)
False Positives	32	23
Complementary Hit	25 (78.1%)	19 (82.6%)
True False Alarms	7 (21.9%)	4 (17.4%)

This table shows that quite a few predictions are Complementary Hits. Overall, the results are very similar in the sense that for each age group SEEM predicts a similar amount of new problems (Complementary Hits) and only a few problems that are unlikely to be experienced by real users (True False Alarms).

### CONCLUSION

In this paper we described a study in which the evaluation of SEEM as a method was extended to an older age group. The target age used in the first studies with SEEM was 5 to 7 years; in this study SEEM was applied to educational computer games for children between 9 and 11 years.

The outcome for the thoroughness (the overlapping predictions with UT) is very similar for the two conditions; however SEEM scores a bit better for the oldest age group. The scores for the validity (the proportion of the problem predictions that are real problems) are also better for the oldest age group. The appropriateness scores (indication of the understanding of SEEM) are about the same for the two age groups.

Overall SEEM has been shown to be a useful analytical evaluation method, both for educational computer games intended for 5 to 7 and 9 to 11 year old children. Furthermore, the results show that UT finds problems not uncovered by SEEM, and that SEEM find problems not uncovered by UT. This indicates that ideally UT and predictive approaches should be combined in practice.

### ACKNOWLEDGEMENTS

First of all, we would like to thank all students who participated in the two studies. Also many thanks to all children who participated in the two user tests, to their

parents for letting them participate and the primary schools for their cooperation. Finally, we would like to thank Wolmet Barendregt for her input on many aspects of this study.

## REFERENCES

1. Bio Mania - Een speelse toepassing op biologie (Bio Mania - a playful application to biology) [Computer software] Edison (1998)
2. Max en de toverstening (Milo and the magical stones) [Computer software] MediaMix Benelux (2002)
3. Regenboog, de mooiste vis van de zee (Rainbow, the most beautiful fish in the ocean) [Computer software] MediaMix Benelux (2002)
4. Baauw, E. and Bekker, M.M.: A comparison of two analytical evaluation methods for children's computer games. Position paper for workshop: Child Computer Interaction: Methodological research, Interact 2005, 12 September 2005 Rome, Italy (2005)
5. Baauw, E., Bekker, M.M., and Barendregt, W.: A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games. Proceedings of Human-Computer Interaction INTERACT 2005, 14 September 2005 Rome, Springer Verlag (2005)
6. Baauw, E. and Markopoulos, P.: A comparison of think aloud and post-task interview. 1 June 2004 Maryland, USA, ACM Press (2004) 115-117
7. Barendregt, W., Bekker, M.M., Bouwhuis, D., and Baauw, E.: Predicting effectiveness of children participants in user testing based on personality characteristics. Behaviour & Information Technology (in press)
8. Barendregt, W., Bekker, M.M., and Speerstra, M.: Empirical evaluation of usability and fun in computer games for children. Proceedings of Human-Computer Interaction INTERACT-03', 9 March 2003 Zürich, Switzerland, IOP Press (2003) 705-708
9. Chatratchart, J. and Brodie, J.: Applying User Testing Data to UEM Performance Metrics. Late Breaking Results Paper, 24 April 2004 Vienna, Austria (2004) 1119-1122
10. Cockton, G. and Woolrych, A.: Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation. In: Blandford, A., Vanderdonck, J., and Gray, P.D. (Eds.): Springer-Verlag (2001) 171-192
11. Hartson, H.R., Andre, T.S., and 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
12. Inkpen, K.: Three Important Research Agendas for Educational Multimedia: Learning, Children, and Gender. Proceedings of Educational MultiMedia '97, June 1997, Calgary (1997) 521-526
13. Lavery, D., Cockton, G., and Atkinson, M.P.: Comparison of Evaluation Methods Using Structured Usability Problem Reports. Behaviour and Information Technology, 16(4) (1997) 246-266
14. 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)
15. Malone, T.W. and Lepper, M.R.: Making learning fun: a taxonomy of intrinsic motivations for learning. In: Snow, R.E. and Farr, M.J. (Eds.): Aptitude, Learning and Interaction III Cognitive and Affective Process Analysis. Hillsdale, N.J., Erlbaum (1987)
16. Nielsen, J.: Heuristic Evaluation. In: Nielsen, J. and Mack, R.L. (Eds.): Usability Inspection Methods. New York, USA, John Wiley & Sons, Inc. (1994) 25-62
17. Norman, D.A.: The design of everyday things. London, MIT Press (1998)
18. Piaget, J.: Science of Education and the Psychology of the Child. New York, Orion Press (1970)
19. Sears, A.: Heuristic Walkthroughs: Finding the problems without the noise. International Journal of Human-Computer Interaction, 9(3) (1997) 213-234