Development and evaluation of the problem identification picture cards method

Wolmet Barendregt · Mathilde M. Bekker · Ester Baauw

Abstract In this paper the development and assessment of a new formative evaluation method called the problem identification picture cards (PIPC) method is described. This method enables young children to express both usability and fun problems while playing a computer game. The method combines the traditional thinking-aloud method with picture cards that children can place in a box to indicate that there is a certain type of problem. An experiment to assess this method shows that children may express more problems (verbally, or with a picture card, or with a combination of a picture card and a verbalisation) with the PIPC method than without this method (in which they can only indicate problems verbally). Children in the experiment did not just replace verbalisations by using the provided picture cards and some children preferred to use the PIPC method during the test instead of the standard thinking-aloud method. The PIPC method or some aspects of the method could be a good instrument to increase the amount of information expressed by young children during an evaluation.

W. Barendregt (✉) · M. M. Bekker · E. Baauw
Department of Industrial Design,
Eindhoven University of Technology, Den Dolech 2,
5600 MB Eindhoven, The Netherlands
e-mail: wolmetb@hotmail.com

M. M. Bekker
e-mail: m.m.bekker@tue.nl

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

1 Getting information from children

Nowadays, computers are used by almost everyone in the developed world, and also children are exposed to computers and technology at an increasingly early age. For example, there are educational CD-ROMs for children as young as 18–36 months, and special keyboards have been developed for babies and small children, e.g. by Génération5 (2005), Berchet (2005), Ergocube (2005). However, for most children the first contact with the computer is through some sort of (educational) game, and children play computer games very often. In The Netherlands, in 2005, 61% of the children under 15 played computer games every day (ANP 2005). It is therefore important that computer games for children are well designed for the intended age group.

One of the most commonly used design philosophies to create high quality products for users is the User-centred design (UCD) approach (Norman and Draper 1986; Rubin 1994; Nielsen 1993). UCD refers to the philosophy that the intended user of a product should always be at the centre of the design process throughout all phases of the design. Druin (1999) gives a classification of the different roles children can play during the design process; children can be users, testers, informants, or design partners. Although, the levels of engagement are different for the different roles, they all include evaluations with child participants as evaluators. This means that that products should be evaluated by having children use an actual implementation of the product in some form in a representative way. While Hanna, Neapolitan, and Risden (2004) focus on the evaluation of different game concepts, which belongs to the earlier stages of the design cycle, this article will focus on the evaluation of games at later stages in the design cycle in which the children can play with a version of the game in order to detect usability and fun problems.

One way to classify evaluation methods is the following: *inquiry methods, observational evaluation methods*, and *analytical evaluation methods. Inquiry methods* focus on users' likes and dislikes, needs, and understanding of a product by asking users to answer questions verbally or in written form. Inquiry methods tend to identify broad usability problems or opinions about a product as a whole. Examples of inquiry methods are User Satisfaction Questionnaires (Reiterer and Oppermann 1993) and Focus Groups (Zirkler and Ballman 1994). An example of a specific questionnaire about fun in computer games for children is the Fun-questionnaire for Kids developed by Stienstra and Hoonhout (2002).

Evaluation methods that collect data by observing users' experiences with a product are called *observational evaluation methods*. Some types of observational evaluation methods are the usability test (Lewis 1982), the user performance test (Nielsen 1993), and cooperative evaluation (Wright and Monk 1991). Methods that do not collect data from users' experiences but rely on the opinion of experts are called *inspection* or *analytical evaluation methods*. Examples of analytical evaluation methods are Heuristic Evaluation (Nielsen and Molich 1990) and the Cognitive Walkthrough (Lewis et al. 1990). This article focuses on observational evaluations in which children participate in tests with the products. However, we try to get more information out of the children during this observation by giving them a task, which includes some aspects of inquiry methods.

Often, observational usability evaluations are performed to determine quantitative measures like efficiency, effectiveness, and satisfaction (ISO 1998). These measures can be used to compare or assess the level of usability of an entire product. Evaluations in order to determine these measures are called *summative* evaluations (Hartson et al. 2001). However, another common goal is to identify as many aspects as possible of a product that cause users trouble (Hertzum and Jacobsen 2001) for the purpose of improving the product by fixing these problems. This type of evaluation is often called *formative* evaluation (Barnum 2002; Hartson et al. 2001). By involving children in formative evaluations of computer games it is possible to improve the games based on their input.

2 Thinking-aloud with children

The 'thinking-aloud' technique, commonly used for formative evaluations of products with adult participants (Nielsen 1993), has the disadvantage that young children can have difficulty verbalizing their thoughts (Boren and Ramey 2000). Because they often forget to think aloud, they need to be prompted to keep talking. However,

prompting could result in children mentioning problems in order to please the experimenter, leading to non-problems being reported (Donker and Reitsma 2004; Nisbett and Wilson 1977). Therefore, some of our experiments (Barendregt et al. 2005), as well as experiments by other researchers (Donker and Reitsma 2004) relied on a combination of self-initiated spoken output complemented with observations of children's behaviour. In this paper this self-initiated spoken output will be referred to as the results of the thinking-aloud method. Unfortunately, the amount of self-initiated spoken output in the thinking-aloud method is often limited. For example, in the study by Donker and Reitsma (2004) only 28 out of 70 children made any remarks at all. Still, verbalisations or other clear signals from the child are very valuable because they may indicate problems that are likely to go undetected when relying on observations alone. For example, when a child thinks something is strange or silly, this is often difficult to detect unless the child says something about it. Furthermore, when an observable problem is accompanied by verbalisations or other explicit indications of a problem, the number of breakdown indications per problem increases, making it more likely that a problem will be detected by multiple evaluators (Vermeeren et al. 2002). Therefore, the reliability of a method that encourages children to express their thoughts while playing the game will be higher.

In this paper a new method that could help children between 5 and 7 years to express more of their thoughts than the thinking-aloud method is described and evaluated. First, the development and rationale of this new method is described. Subsequently, an experiment to test whether this method really encourages children to express more problems explicitly than the thinking-aloud method is described.

3 Development of the method

The first attempt to develop a new method to make children express more problems assumed that children might be too shy to verbalise their thoughts in front of an unfamiliar facilitator. Based on literature about interviewing strategies in child assessment (Kanfer et al. 1983) it was hypothesized that children may talk more to someone they feel closer to than to the adult facilitator. The description of the Berkeley Puppet Interview method to assess children's self-perception by Measelle et al. (1998) gave rise to the idea to equip the facilitator with a hand puppet. The hand puppet would try to build rapport with the child, and hopes were that children would try to engage the puppet in the game by talking to it about the game. In several pilot tests with children this idea was investigated with a cute hand puppet representing a fox (see Fig. 1).

Fig. 1 The fox hand puppet that was used in the unsuccessful pilot study

However, it appeared that the method was hard to apply and would probably not give the expected results. There were several reasons for this failure:

- To give children the feeling that the hand puppet is real and engage them in a conversation the facilitator must be a rather good puppeteer. This makes it less suitable as a general method for facilitators.
- The children actually appeared to be very comfortable with the facilitator. Therefore, they kept addressing the facilitator even when the hand puppet was present. The conversational situation contained therefore three participants which made it complex and unnatural for the facilitator to keep track on using the hand puppet as a mediator.

On the advice of a play therapist it was decided to develop a method with picture cards that children can place in a box to express different kinds of problems either verbally or non-verbally. There are several reasons why these picture cards would help children to express more problems explicitly than when the facilitator just asks the child to verbalize as much as possible about anything:

- During the introduction, the facilitator can use the picture cards to explain not only verbally but also visually what kind of information he/she is interested in. This combination of auditory and visual information adheres to the principles of multiple resources and redundancy gain (Wickens et al. 2004) and may make it easier for children to understand the explanation.
- During the test the picture cards serve as memory aids for the things the evaluator is interested in, thereby putting 'knowledge in the world' (Norman 1998)

instead of 'in the head' and thus relying less on long-term memory.
- Some children are able to verbalize what they think or feel, while others may be less verbally capable. With the picture cards method less verbally capable children can express themselves explicitly without having to verbalize. This is a similar approach as several interviewing techniques for young children (Measelle et al. 1998; Greca 1983).

3.1 Choosing the pictures

In order not to overload the children with too many different concepts to remember, it was decided to use a maximum of eight pictures. These pictures had to cover the feelings children may have when encountering different kinds of problems or when they really enjoy the game. For usability problems, a distinction was made between problems related to perception, cognition, and action (Norman and Draper 1986). For fun problems, a distinction was made based on the taxonomy of Malone (1980) and Malone and Lepper (1987) for what makes computer games fun. For each usability and fun problem type one or more possible expressions or feelings of children were determined that could be represented by a picture card. The decision about which expressions and feelings were going to be used was based on the combination of verbalisations of children made during earlier evaluations, and on the Fun-questionnaire by Stienstra and Hoonhout (2002). Some pictures could be used for different kinds of problems, and this was also necessary to limit the number of cards. For example, when something is hard to see or hear children may say that it is difficult. Children may also say that it is difficult when something is hard to click because it is very small. Although, these problems are not of the same type we just used one picture card for 'difficult' because it was reasoned that the available context of the game would help to determine the meaning.

3.1.1 Usability

Perception

- To be able to use a game, a child first needs to perceive the information given by the game. When a child encounters a *perception problem* he/she may say it is *difficult* to hear or see something clearly.

Cognition

- When a child encounters a *usability problem related to knowing what to do, how to do it*, or understanding the

feedback he/she may say that he/she does *not understand* what to do, or what has happened.

Action

- When a child encounters *a usability problem related to performing the physical actions* he/she may find it *difficult* to use the mouse in order to click objects.

For these problems, we reasoned that children would not make a difference between something being hard to perceive or hard to activate. Therefore, there are no separate cards for Perception and Action problems.

3.1.2 Fun

The taxonomy of Malone and Lepper provides heuristics for what makes games fun. Based on these heuristics, four types of fun problems can be distinguished: Challenge problems, Fantasy problems, Curiosity problems, and Control problems.

- When a child encounters a fantasy problem because the game is aimed at older children he/she may find it *scary*.
- When a child encounters a fantasy problem because the game is aimed at younger children he/she may find it *childish*.
- When a child encounters a fantasy that is incongruent with the story or with his/her experiences he/she may find it *silly/strange*.
- When a child experiences a problem related to a *too high challenge level* he/she may find it too *difficult*.
- When a child experiences a problem related to a *too low challenge level* he/she may find it *boring*.
- When a child experiences a *control problem* he/she may think it takes too long. It is possible that a child would also experience a control problem when something goes too fast, but this reaction was never experienced during earlier evaluations so no card was included for this.
- When a child experiences a *curiosity problem* he/she may find it *boring*.

We re-used the 'difficult'-card for perception/action problems to also indicate challenge-problems. Furthermore, we used the 'boring'-card for both control and curiosity problems.

To make it clear to the children that the evaluation of a game is of course also about fun, one last concept 'Fun' was added.

In the first version of the picture cards, small icons were chosen from different online libraries to represent the different concepts. These icons were glued to wooden cards of about 2×2 cm. This first version was tested with two

children in their home. Although, the children did put some pictures in the box they had trouble picking up the small cards. Therefore, it was decided to make bigger cards of about 4×4 cm. For these bigger cards, clearer pictures were selected to represent the concepts. The final pictures were chosen from two on-line picture libraries that are also recommended for the PECS (Picture Exchange Communication System)-method. This PECS-method developed by Bondy and Frost (1994) is used to teach non-verbal autistic children to express themselves by exchanging picture cards. The libraries that were used to select pictures from are:

http://www.childrenwithspecialneeds.com/pecs/pecsindex.html

http://www.pdictionary.com/english/

The pictures chosen for the problem identification picture cards (PIPC) method are given in Fig. 2.

Each picture was glued to both sides of a wooden card. A wooden box with eight compartments was created in which children could place one of the cards when they encountered a problem that they wanted to express to the evaluator (see Fig. 3).

4 The problem identification picture cards method

When using the problem identification picture cards (PIPC) method, children get an explanation of each picture and the kind of situation for which they can use it before the test session. During the test, the box and numerous picture cards for each problem category are placed on the table next to the computer on which the game is played. Children can place as many picture cards in the box as they like. The children can always ask for an explanation of a card if they happen to forget it. It does not matter whether they use the correct picture card for a particular problem. If the facilitator does not understand why a certain card is used he/she can ask the child for an explanation. Finally, the behaviour of the child with the game together with the picture cards is
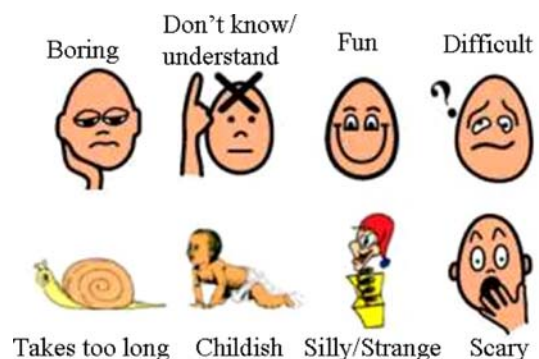

Fig. 2 The eight pictures used for the picture cards

Fig. 3 The box with compartments for the picture cards. Above each compartment of the box, the concept represented by the picture card is printed

used to do the actual analysis of the test session. The picture cards are not meant to be used without doing further observations of the children's behaviour.

4.1 Experiment: evaluation of the PIPC method

The aim of the method was that children would express more problems, either verbally and/or using the picture cards than when they would just have been asked to verbalise as much as possible. To test whether the PIPC method would serve this aim, an experiment was set up to compare the two methods. The hypotheses concerning the differences between the PIPC method and the method solely relying on self-initiated spoken output are discussed in the next subsections.

4.1.1 Hypothesis 1

Each picture card shows one of the pictures of Fig. 2. These pictures represent the feelings children may have when they experience a problem (except for the Fun picture, which expresses enjoyment). Through the use of the picture cards children will probably have a clearer understanding of the feelings that they can communicate to the facilitator that indicate a problem. Furthermore, the picture cards may serve as a visual reminder of these feelings. Finally, children who are not so verbally capable can also express their feelings non-verbally by using a picture card. Therefore, the first hypothesis is that children will express more problems when they use the PIPC method than when they have only been asked to verbalise as much as possible.

To test this, hypothesis regression analyses will be performed to decide whether the difference in expressed problems between the methods can be explained by the game with which the methods are used, or by the order in which the methods are used, or in a combination of order and game. If this is not the case, a Wilcoxon signed ranks test will be performed to determine whether there is a significant positive effect of the PIPC method in the numbers of expressed problems.

4.1.2 Hypothesis 2

Although, the picture cards give a clear indication of a problem, verbalisations can also give valuable information to an evaluator. Therefore, the picture cards should be an addition to thinking-aloud; children should not just substitute verbal indications of problems with picture cards. The hypothesis is that this is not the case; the number of verbalised problems will not be lower for the PIPC method than for the thinking-aloud method.

To test this hypothesis, a Wilcoxon signed ranks test will be performed on the numbers of verbalised problems with both methods.

4.1.3 Hypothesis 3

It is not always easy to find children who are willing and able to participate in a user test. Therefore, a user test should be a pleasurable experience to the children who participate so they like to participate again. The hypothesis is that children will like the PIPC method at least as much as the usual think-aloud method.

4.2 Method

#### 4.2.1 Test participants

To test the hypotheses, an experiment was set up with 23 children of four groups two (second year kindergarten) of the Wethouder van Eupen school, an elementary school in Eindhoven, The Netherlands. This school is situated in a neighbourhood that is mainly inhabited by people who received higher education and earn more than minimum wage. All children were 5 or 6 years old, twelve girls and eleven boys. They were recruited by a letter to the parents asking for their cooperation.

4.2.2 Experimental set-up

The results of one of our other experiments (Barendregt et al. 2005b) indicated that there are very large individual differences in how much of the experienced problems children will verbalise. This difference can largely be predicted by certain personality characteristics. It was decided that the experiment to test the effect of the PIPC

Fig. 4 Temporal representation of how the test procedure was implemented (time in minutes)
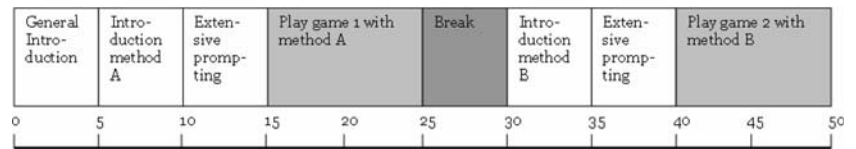
method should be a within-subject experiment in order to lower the effect of these individual differences.

Because it is also likely that the types of problems that children experience change when they become more experienced with a game (Barendregt et al. 2005a), it was decided that children should play a different game for each method. These games should be of similar difficulty but different in the types of sub games that can be played.

It can be expected that children will learn from performing the first method and will thus perform better on the second method. To compensate for the order in which the children used the different methods, each method should be used equally often as the first and as the second method.

Because the games should be different in the types of sub games that can be played, it is not expected that the children will learn how to play the second game from playing the first game. Altogether, there were four different conditions and 23 children in the experiment. The children were randomly assigned to one of the conditions:

4.2.3 Test material

The 23 children in the experiment were asked to participate in a user test of two computer games 'Milo and the red fruit' (MediaMix 2004b), and 'Little Polar Bear, Do you know the way?' (MediaMix 2004a). These games are intended for children between 4 and 8 years old and are good representatives of software products for children of the chosen age group of children between 5 and 7 years old. They were new to the market at the time of the experiment; therefore, children would probably be unfamiliar with them. Furthermore, large numbers of problems were anticipated for children playing these games alone because even the adult researchers had some problems playing them. This would make the games quite suitable for the experiment.

4.2.4 Procedure

Each individual child was taken from the classroom for 50 min to perform two user test sessions; one for each method with a different game. First, the test facilitator explained the general purpose of the test session, the procedure for the first method: either the think aloud method or the picture cards method. The child then played the game for 15 min. As a training session the facilitator prompted the child extensively to talk aloud and/or use the

Table 1 Description of the four conditions in the experiment and the number of children in each condition

Condition	Part 1 of the test	Part 2 of the test	No. of children
1	PIPC for game 1	Thinking aloud for game 2	5
2	PIPC for game 2	Thinking aloud for game 1	6
3	Thinking aloud for game 1	PIPC for game 2	6
4	Thinking aloud for game 2	PIPC for game 1	6

cards during the first 5 min. During the subsequent 10 min, the child could play the game as he or she liked without any specific tasks or prompting from the facilitator. When a child asked for help the first time, the test facilitator would only encourage the child to keep on trying. The second time a child asked for help the test facilitator would give a hint and only after the third time a child asked for help the facilitator would explain the solution in detail. After finishing the first test session, the child would get a short break of at most 5 min in which the facilitator started up the next game. After that the facilitator explained the next method for 5 min, and then prompted the child while playing the game extensively for 5 min. Finally, the child played the second game with the next method without prompting for 10 min. Each test session was videotaped, recording a split-screen shot of the face of the child and the on-screen actions. A graphical representation of the procedure is given in Fig. 4.

At the end of the test session, the child was asked to fill in a very short questionnaire. In this questionnaire the child had to mark with a cross which game he/she preferred and whether he/she preferred to do another evaluation in the future with or without the picture cards. The order of the possible answers was randomly changed to ensure that a preference for one of the games or with/without the picture cards were not due to the presentation of the answers. (Table 1)

5 Analysis

For each child the recorded video material was used to transcribe the protocols of both conditions for the 10 min

Table 2 Example of a transcribed protocol

Problem count	Verbalisation or picture card of the child	Facilitator
1	(Places picture 'don't understand' in box)	
		Yes, do you want me to help you?
	Yes	
		Well done! Look, here is one of the games, and here is another one, and here is another one. So now you can choose which one you would like to do
	That one	
		See, then there is a yellow light
2	(Places picture 'This takes too long' in box)	
		Yes, does it take too long? Do you want me to tell you what to do?

In the leftmost column, it is indicated whether a verbalisation and/or a picture card will be considered as an indication of a problem. Indications that belong to the same problem are given the same problem number

that they played the game without much interference from the facilitator (the two light grey boxes in Fig. 4). For the picture cards it was also noted in the protocol when a child placed a picture card in one of the boxes (see example in Table 2). These protocols were used to count the number of unique problems (meaning that if a child experienced the same problem more than once, there was still only one problem counted) that were indicated verbally, with a picture card, or with a combination of a verbalisation and a picture card.

A second evaluator checked these numbers by looking at the protocols and asking critical questions about why certain verbalisations were or were not taken into account, and whether certain verbalisations should be grouped or split. This review led to some minor changes in the final problem counts; for the PIPC method two problems expressed by one child were combined into one, and for four children a verbalisation was no longer counted as a problem, for the thinking-aloud method one verbalisation was removed as a problem and one was added.

6 Results

The results of the analysis of all protocols are given in Table 3.

For testing the hypotheses in this experiment, a significance level of 0.05 was chosen.

Table 3 Numbers of expressed problems per child during the test with the think-aloud method and the PIPC method, and the number of verbalized problems during the PIPC method

Child	Expressed problems thinking aloud	Expressed problems PIPC	Verbalised problems PIPC
1	0	3	2
2	3	0	0
3	4	5	5
4	5	4	4
5	5	11	4
6	3	7	6
7	4	3	3
8	1	4	1
9	2	3	3
10	7	7	7
11	4	10	7
12	4	5	5
13	1	0	0
14	2	1	1
15	4	7	7
16	5	6	5
17	6	7	7
18	6	6	6
19	0	4	4
20	0	0	0
21	3	3	3
22	8	8	7
23	6	5	5
Mean (and SD	3.6 (2.3)	4.7 (3.0)	4 (2.4)

(1) None of the regression analyses for the games, the order of the methods, or the combination of these two factors on the difference between the numbers of expressed problems with both methods was significant ($p > 0.05$). The Wilcoxon signed ranks test showed that there was a significant positive difference between the number of problems expressed with the PIPC method and the thinking-aloud method ($Z = -2.024$ based on negative ranks, $p < 0.05$).

(2) The Wilcoxon signed ranks test showed that there was no significant difference ($p > 0.05$) between the number of verbalised problems with the PIPC method and the number of verbalised problems with the thinking-aloud method. This was in line with our hypothesis that the PIPC method would not prevent children from thinking-aloud naturally.

(3) For the third hypothesis, a Chi-Square test was performed on the expected and actual numbers of children who liked to perform another test with or without the PIPC method were compared. The

number of children who would rather perform another user test with a new game with the PIPC method (14 of the 23 children = 61%) was not significantly lower ($X^2 = 1.087$, df = 1, $p > 0.05$) than the number of children who would rather perform another user test with a new game with the thinking-aloud method (9 out of the 23 children = 39%). Our hypothesis that the children would like the new method at least as much as the standard thinking-aloud method was thus conformed.

The lowest number of cards used by a single child was 0. The highest number of cards used by a single child was 9. Most children did not use more than two cards. The card that was used most frequently was the 'Don't know/understand' card. This is not surprising since both games contained many sub games that the children could not understand from the given explanation. By using this card the children were able to get help from the facilitator to continue playing the game.

7 Discussion

7.1 Games

The PIPC method was tested with two different adventure games. However, it is unclear whether the method also works for other game genres. Especially with very fast-paced games, children may not be able to give attention to the picture cards during the user test. An example of such a game for children in the chosen age group is 'Freddi Fish, Silly Maze' (Transposia 2000). Further research is needed to determine which games could be evaluated with this method.

7.2 Procedural issues

It is quite hard to find teachers who are willing to let their pupils participate in an experiment at school. Therefore, we had to restrict the experiment to children aged 5–6 in their second year at school, because they do not yet have to perform so many educational tasks. Children in third year of school, who are typically 7 years old, were not allowed to participate since they have to reach specific end terms, and many teachers we contacted were concerned that the experiment would interfere too much with their school work. To minimise the intrusion on the normal routine in the school classes, the experiment also had to be restricted to one test per child within 1 week, making the sample size rather small. It would be very useful to repeat the experiment with larger and other groups of children.

Furthermore, because young children have short attention spans, the maximum testing time had to be about 30 min to 1 h (Hanna et al. 1997). Altogether, this meant that the children had to perform the two test sessions in a short time. Consequently, the (training) time in which the facilitator could prompt for verbalisations or picture cards was also very short (5 min). In these 5 min children did not encounter many problems for which the facilitator had the opportunity to prompt, so the children were not very well trained with each method.

When performing an evaluation of one game, practitioners will usually have more time to train the test participants. This holds for both the PIPC method and the thinking-aloud method. Because no detrimental effects of the PIPC method on the number of verbalisations was found it is likely that the PIPC method will still give a higher number of expressed problems than the standard thinking-aloud method, even with better training. However, the effect of a longer training time on the number of expressed problems should be examined further.

7.3 Gender differences

We did not look specifically at gender differences during the evaluation of this method. The main reason is that we were interested in a general method that would work for both boys and girls. During the tests there were both boys and girls who seemed to respond well to the method. Probably their reaction to the method is also related to other factors such as personality. This was also the reason why the experiment was set up as a within-subject experiment. It would be interesting to see whether other personality characteristics than the ones we identified as being good predictors for the verbalisation of problems (Barendregt et al. 2005b), influence how children react to methods such as the PIPC method.

7.4 Possible improvements to the PIPC method

In this study, the picture cards in their present form were an effective addition to the thinking-aloud method that relies solely on self-initiated spoken output. Still, several changes to the picture cards method are possible, but it has to be tested whether they would really be improvements.

Firstly, because the box in which to put the cards has to be put alongside the computer, children have to shift their attention from the screen to the box to place a picture card in one of the compartments. When the game initiates the interaction, for example when explaining something, children have to divide their attention between the game and the picture cards. Shifting attention from one display location to another requires effort (Wickens et al. 2004).

Placing the pictures within closer proximity to the computer screen may make it easier for children to use the pictures in combination with the game.

Secondly, it was striking that only few children used a picture card without verbalizing anything. It seemed that the picture cards functioned much more as an aid to remember the things of interest than as an aid to help children who have difficulty verbalizing express their thoughts in a non-verbal way. This impression was also corroborated by the fact that some children just looked at the picture cards and then started verbalizing their thoughts. Maybe, it is therefore not even necessary to ask children to put a picture in the box, but just to point to it. It could be interesting to compare this pointing-method to the presented PIPC method and the thinking-aloud method.

Thirdly, the pictures used from the PECS-libraries were chosen because they were thought to express the feelings children would have when they encountered the different types of problems. The actual words associated with the pictures in the libraries were not always the same as the feelings or thoughts they had to represent. For example, the 'jack-in-the-box' picture was used to express 'this is silly'. It is uncertain whether the pictures used were the best pictures for the different types of problems children can encounter when playing a game. However, children were not obliged to be able to remember the meaning of the pictures perfectly. When they forgot the meaning of a picture they could ask the facilitator. Therefore, it was concluded that the pictures used were sufficient to remind the children of the concepts, even when they were not perfect. Further research is needed to determine whether other pictures may be superior in expressing these concepts better.

Finally, it is possible that some of the concepts depicted by the pictures are superfluous, or that additional pictures are necessary. For example, the picture cards with 'scary' and 'childish' were almost never used. Therefore, further research is needed to determine the optimal set of pictures for the cards.

7.5 How do problems detected with the PIPC method relate to predicted and observed problems?

Expert researchers in children's usability detect many both large and small issues as children play, even if the children themselves appear to be having fun. Furthermore, they can predict some problems based on their experience. So how do the problems indicated by the children themselves with the PIPC method relate to the problems that can be detected or predicted by usability practitioners? For this experiment, we can look at our predictions compared to what the children experienced, and we can look at which

problems the children tended to indicate compared to what could be observed.

Before testing the games with the children, we played them ourselves in order to see which problems we could expect and how we would be able to help children when they asked for help. It became clear that for both games we could expect many problems related to knowing what to do. Furthermore, we expected to see impatience since both games were rather slow paced, especially during introductions. During the tests at least one (and often many more than one) child experienced the problems that we had predicted. However, they also experienced other problems that we had not predicted. For example, in 'Little Polar Bear, Do you know the way?' a shadow of an object was displayed that was actually built up by placing different objects together and shining a light on this construction. The children had to find and click the objects in the rest of the screen that could form this shadow object. Many children did not see the difference between the objects to pick from and the shadow object that had to be created. We had not expected this, so it was very enlightening to observe this problem.

If we look at the problems that the children expressed and the problems they experienced that could be observed by the evaluators, we notice that children only indicate a very low percentage of the problems they experience. This is not so surprising since we already knew that children also verbalise only a very low percentage of all problems (Barendregt et al. 2005b). One of the nicest things about the picture cards was that we finally got a clear indication from one of the children that he was annoyed by long introductions that could not be interrupted. We had observed impatient behaviour such as repeatedly clicking before, but we had never seen any verbal indication of impatience or frustration. Now this child indicated his frustration explicitly by placing a card with the snail in the box. However, we do recommend using the picture cards only as an addition to observations not as a replacement, since there are so many problems that children do not indicate themselves.

7.6 An unpredicted benefit of the PIPC method

One of the main advantages of the PIPC method that was not anticipated was the fact that it was much easier for the facilitator to keep the attention of the children when explaining what the children were supposed to do. Although, the facilitator tried to explain this in both conditions, it was clear that many children could not keep their attention when the explanation was done only verbally. When using the picture cards it was much easier for the facilitator to explain the purpose of the test in a playful way by making the children guess the meaning of a certain

picture and talk about it. Therefore, the children could direct their attention to the explanation of the test situation while in the verbal condition their eyes were often drawn towards other things in the room. Because we had not expected this we could only do some tests afterwards. We used a post-hoc Wilcoxon signed ranks test to check whether the children really used more of the discussed concepts in their verbalisations with the PIPC method than with thinking-aloud. For each child, the number of times a concept from the picture cards was used in a verbalisation was compared for both conditions. The test showed that with the PIPC method children used the concepts explained with picture cards significantly more often than the same concepts explained verbally for thinking-aloud ($Z = -3.26$, $p < 0.01$). However, it is very possible that this was also caused by the fact that the cards served as a reminder during the test.

8 Conclusions

The problem identifying picture cards can be a good addition to the thinking-aloud method based on self-initiated spoken output. When children can use these picture cards in addition to thinking-aloud, they may express more problems than with standard thinking-aloud. In the present experiment children did not just replace verbalisations by picture cards without any verbalisations and the children did not think that it was less pleasurable to use the picture cards than standard thinking-aloud. Whether other versions of the picture cards method (with more or fewer pictures, different pictures, and placement of the pictures closer to the computer screen, and with or without the tangible aspect) can further improve the outcome of a user test should still be investigated.

The PIPC method can be a good method to be used by practitioners because the pictures help to explain the different types of problems that children can experience more clearly than verbal explanations alone. This can also help the facilitator feel more at ease when explaining the purpose of the test. Furthermore, the picture cards serve as a memory aid during the test, and children are able to clearly express problems in a non-verbal way. Therefore, the number of explicitly indicated problems may be higher than with standard thinking-aloud.

References

ANP (2005) Algemeen Nederlands Persbureau, Helft kinderen zit iedere dag te gamen (Half of all children plays games every day)

Barendregt W, Bekker MM, Bouwhuis DG, Baauw E (2005a) Identifying usability and fun problems in a computer game during first use and after some practice. Int J Hum Comput Stud 64(9):830–846

Barendregt W, Bekker MM, Bouwhuis DG, Baauw E (2005b) Predicting effectiveness of children participants in user testing based on personality characteristics. Behav Inf Technol (in press)

Barnum CM (2002) Usability testing and research. Longman, London

Berchet (2005) http://www.berchet.com/index.php (retrieved on December 5 2005)

Bondy A, Frost L (1994) The picture exchange communication system. Focus Autistic Behav 9:1–19

Boren MT, Ramey J (2000) Thinking aloud: reconciling theory and practice. IEEE Trans Prof Commun 43:261–278

Donker A, Reitsma P (2004) Usability testing with young children. Interaction design and children, 1 June 2004, ACM, Maryland, pp 43–48

Druin A (1999) The design of children's technology, Morgan Kaufmann, San Francisco

Ergocube (2005) http://www.store.ergocube.com/ (retrieved on December 5 2005)

Génération 5 (2005) http://www.generation5.fr/(retrieved on December 5 2005)

Greca AM (1983) Interviewing and behavioral observations. In: Walker CE, Roberts MC (eds) Handbook of clinical child psychology. Wiley-Interscience, Chichester, pp 109–131

Hanna L, Risden K, Alexander K (1997) Guidelines for usability testing with children. Interactions 4:9–14

Hanna L, Neapolitan D, Risden K (2004) Evaluating computer game concepts with children. In: Proceedings of the 2004 conference on interaction design and children, pp 49–56

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

Hertzum M, Jacobsen NE (2001) The evaluator effect: a chilling fact about usability evaluation methods. Int J Hum Comput Interact 13:421–443

ISO (1998) Ergonomic requirements for office work with visual display terminals (VDTs) Part 11. Guidance on usability (Rep. No. ISO No. 9241-11). International organization for standardization, Geneva

Kanfer R, Eyberg SM, Krahn GL (1983) Interviewing strategies in child assessment. In: Walker CE, Roberts MC (eds) Handbook of clinical child psychology, Wiley-Interscience, New York, pp 95–108

Lewis C (1982) Using the ''thinking-aloud method'' in cognitive interface design (Rep. No. RC 9265 (#40713)). IBM Thomas J. Watson Research Center, Yorktown Heights

Lewis C, Polson P, Wharton C, Rieman J (1990) Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In: Proceedings of CHI 90, ACM, New York, pp 235–242

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, Lawrence Erlbaum, Hillsdale

Measelle JR, Ablow JC, Cowan PA, Cowan CP (1998) Assessing young children's views of their academic, social, and emotional

lives: an evaluation of the self-perception scales of the Berkeley puppet interview. Child Dev 69:1556–1576

MediaMix (2004a) Kleine IJsbeer weet jij de weg? (Little polar bear, do you know the way?) [Computer software] MediaMix Benelux, Overijse

MediaMix (2004b) Max en het rode fruit (Milo and the red fruit) [Computer software] MediaMix Benelux, Overijse

Nielsen J (1993) Usability engineering. Academic, Boston

Nielsen J, Molich R (1990). Heuristic evaluation of user interfaces. In: Proceedings of ACM CHI'90 conference 1–5 April, Seattle, pp 249–256

Nisbett R, Wilson T (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231–259

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

Norman DA, Draper SW (1986) User centered system design: new perspectives on human-computer interaction. Lawrence Erlbaum, Hillsdale

Reiterer H, Oppermann R (1993) Evaluation of user interfaces: EVADIS II—a comprehensive evaluation approach. Behav Inf Technol 12:137–148

Rubin J (1994) Handbook of usability testing: how to plan, design, and conduct effective tests. Wiley, Chichester

Stienstra M, Hoonhout J (2002) TOONS Toys. Interaction toys as a means to create a fun experience. In: Proceedings of the interaction design and children conference 2002, Eindhoven, Shaker Publishing BV, Maastricht, pp 199–210

Transposia (2000) Freddi Fish, Dolle Doolhof (Freddi Fish, Silly Maze) [Computer software] Transposia

Vermeeren APOS, den Bouwmeester K, Aasman J, de Ridder H (2002) DEVAN: a detailed video analysis of user test data. Behav Inf Technol 21:403–423

Wickens CD, Lee J, Liu Y, Becker SG (2004) An introduction to human factors engineering. Pearson Education International, Upper Saddle River

Wright P, Monk A (1991) A cost-effective evaluation method for use by designers. Int J Man Mach Stud 35(6):891–912

Zirkler D, Ballman DR (1994) Usability testing in a competitive market: lessons learned. Behav Inf Technol 13:191–197