Guidelines for user testing with children

Wolmet Barendregt, Mathilde M. Bekker
Department of Industrial Design, Eindhoven University of Technology
P.O. Box 513, Den Dolech 2
5600 MB, Eindhoven, The Netherlands
+31 40 247 5714
{w.barendregt, m.m.bekker}@tue.nl

ABSTRACT

This article gives an overview for practitioners of how to conduct user tests (of computer games) with children of about five to seven years old. The advice is based on the experiences of user tests with many children of this age group in the usability lab for children at Eindhoven University of Technology and at schools. Issues that are discussed are the preparation of the test, how to behave towards the children during the test, and the number of evaluators required to analyze the test.

Keywords

User testing, children, computer games, guidelines

INTRODUCTION

Although many books have been written on how to conduct user tests with adults [11, 13] it is only recently that attention has also been given to user testing with child participants. Hanna, Risden, and Alexander [7] created a document with guidelines for usability testing with children and this is still one of the very few and most often cited articles on this subject. In this article some additional guidelines for usability testing with children are given, based on our research of optimizing user testing with children [3] [1, 2, 4]and on personal experiences in the lab and at schools. The guidelines in this article are organized in the same way as those of Hanna, Risden, and Alexander, to make it possible to combine them easily. First the set-up and planning is discussed, followed by the way to make introductions and how to conduct the test itself. Second, rounding off the test is considered. The article ends with a short discussion of the number of evaluators necessary to analyze the videotapes made during the test sessions.

Set-up and planning

The number of children

One of the first decisions when planning a user test with children is how many participants are needed to detect a sufficient percentage of problems. Several researchers have shown that the first three to five test participants are enough to find 80% of the usability problems [12]. The formula$1-(1-p)^n$, where n is the number of test participants and p is the detection rate of a given problem is used to calculate this number [14]. This means that the average detection rate p of a problem should be as high as 0.42 with only three test participants, and that it should be 0.28 with five participants in a test.

However, detection rates are often much lower [4, 9] and in that case many more test participants are needed to uncover 80% of all problems. This is also our experience with children as test participants in user tests of computer games. The average detection rate ranged from 0.12 to 0.14 in our experiments, which means that eleven to thirteen children would have been needed as test participants to detect 80% of the problems.

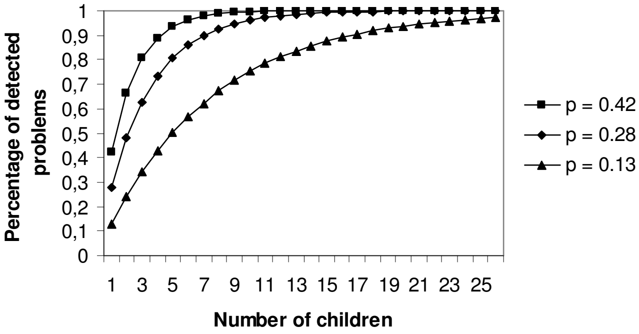

Figure 1 Percentages of detected problems with increasing number of participants for three different detection rates p.

Even when dismissing all problems detected by only one child, the average detection rate was 0.19, which means that eight children would have been needed to detect 80% of the problems.

At least five and preferably more than five children will uncover much higher percentages of problems. Furthermore, using five or more participants also gives a much clearer impression of the severity of problems. However, even one child is always better than no children at all.

Selecting children

Especially when the budget for testing is small and therefore not many children can be included in the user test, it is crucial to get as much information out of each child as possible. Including only children who will experience many problems and verbalize them as well is a good way to optimize children's input.

Our research [2] showed that scores on a limited set of personality characteristics can be used to predict which children will experience many problems and which children will verbalize many problems. The commonly used set of personality characteristic consists of five main personality characteristics divided into eighteen sub characteristics. The sub characteristic 'Curiosity' of the main characteristic 'Intelligence' is a good indicator for the number of problems, while the combination of the main personality characteristics 'Extraversion' and 'Friendliness' is a good indicator for the percentage of verbalized problems.

Although in our research a questionnaire called Blikvanger [6] was used to assess these user characteristics it may be enough to ask the parents whether their child is curious, extravert, and not overtly friendly. These characteristics should be explained in terms of behaviour of the children in order to make clear to the parents what is meant. For example the term 'unfriendly' may have a negative connotation for parents. It means that children should be critical and willing to express their criticism to a facilitator, which sounds much less negative.

Allowing children to practice before the actual test

The numbers and types of problems that are uncovered can depend on the amount of practice children have with the tested game before they participate in the test.

Another of our experiments showed that knowledge and judgment problems occur most often during first use. These problems are related to knowing what to do and understanding the feedback. Control problems occur more often when children have practiced with a game. These problems are related to whether explanations and feedback take too long and cannot be interrupted. Furthermore, problems caused by too high a challenge level also occur more often during first use.

Because many knowledge problems are quite serious and require someone helping the child to overcome them, it is more important to test during first use than after the child has practiced with the game. Control problems are merely annoyances which could probably also be avoided by using heuristics like 'Make stories, explanations, and feedback interruptible, unless it is the first time that they are given' and 'Keep stories, explanations and feedback as short as possible'. Therefore, testing after children have gained some experience with the game to detect control problems does probably not justify the costs.

Challenge is something that can make or break a game; there is a very delicate balance between the right challenge level and either a too high or too low challenge level. To determine whether the challenge level of specific parts of the game is appropriate it may be advisable to give children some opportunity to practice and to test those parts again after the practice.

Using the Picture Cards Method

The aim of a user test with children is to detect as many problems as possible in the tested product with as much explanation as possible from the children.

We proposed the Picture Cards Method [1] to make children express more problems explicitly during the user test. It is essentially a box with picture cards symbolizing different types of problems that children can encounter while playing a game. These picture cards were used to explain the purpose of the test and served as a reminder during the test. Finally, the picture cards could be used by less articulate children to express problems clearly in a non-verbal way by putting a picture card in the box. An experiment showed that children indicated more problems explicitly with the picture cards than without the picture cards, without decreasing the number of verbalized problems. Furthermore the children liked to use the picture cards.

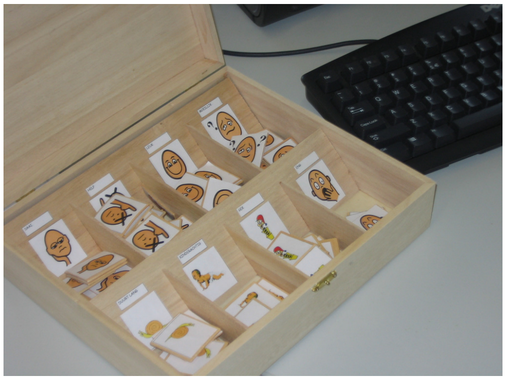

Figure 2 Picture Cards box, with pictures indicating different types of problems children can encounter during a test.

In practice we would certainly recommend using pictures to explain what information you would like to get from the children because it is much easier to keep children's attention during the explanation. The facilitator can engage the children more during the introduction and it is a good opportunity to establish rapport. Furthermore, we would also strongly recommend keeping the pictures within children's focus of attention during the test to serve as a reminder of the concepts. However, making the children put the picture cards in a box is probably not very beneficial because it takes too much time to pick up a picture card and put it in the box, and is therefore too distracting. Other less intrusive means to select pictures should be provided.

Introductions

Using a protocol

Although it helps to put some essential things about how the test will proceed on paper for the test facilitator, it is not advisable to refer too strictly to this protocol when introducing the children to the test. It should be something that you use in a very natural way. During the first few sessions of a series of tests the first author often tried to adhere strictly to the protocol but it always seemed that the children could pick up her nervousness and unnaturalness and would become very quiet, which is something you don't want. Only after several tests would she get comfortable and less concerned about whether she had said everything in the right order, which almost always had a positive effect on how open the children were towards her.

During the test

Thinking aloud and using a protocol

The thinking aloud technique is often used with adults. During the test the participant is asked to verbalize his/her thoughts while using the product. Young children are often not very good in thinking aloud. One of the reasons is that it is unnatural to talk to no-one in particular. During our tests some children when asked to think aloud responded with: 'But to whom should I be talking, to you?' This is a clear example of Boren and Ramey's [5] position that people always need a conversational partner to be able to think aloud. The facilitator should therefore respond naturally to remarks of the children without biasing them. For example, children will often be very enthusiastic when reaching a sub goal, and they will say things like 'I did this very well, didn't I?'. Usually, we do respond to these remarks but try to keep it short in order to keep the child playing, e.g. 'Hm hm, very well'.

Before beginning the test the facilitator should write down appropriate ways to respond to children in order to avoid bias and different treatments of children. However, testing with young children is very unpredictable. Sometimes children will start to cry, or talk about their health or ask questions about other children. The facilitator should be prepared for this and should be able to improvise. Therefore, strictly following a protocol will often not be possible and is not advisable. For example, in most tests for our research the protocol was that children would not get help if they did not ask for it repeatedly. However, one girl in our pilot test first stared at the screen for a long time, when the facilitator finally asked her whether she knew what to do she started crying and indicated that she had no idea what to do. To make her feel more comfortable the facilitator told her that they would play the game together and took over the mouse. When she started to feel more comfortable again and was telling what they should do the facilitator gave her back the mouse and she continued playing the game on her own. In this case the facilitator did not stick to the protocol but we think the remainder of the test was still very valuable.

Furthermore, prompting children to keep talking when they keep silent is often not very useful. Sometimes children will respond to your request but it is very doubtful whether their response is valid because it seems that they are just making something up to say to you. After a single response most children remain quiet again.

Giving help

Children can sometimes become really sad when they don't understand how to proceed or what has happened. It is therefore very difficult for a facilitator to give no help to children when they get stuck and ask for help. Altogether it is almost inescapable that the facilitator will give help when children ask for it. However, the facilitator should first encourage the children to try a bit longer. This will make it easier to determine the severity of a problem. Furthermore, the facilitator should make sure that help is given at the right level. This means for example that the facilitator should assure him/herself first that children understand the goal of a game before he/she explains what actions should be taken to reach the goal or what the feedback means.

Tasks or free play

One of our experiments showed that to detect usability and fun problems in a computer game in a realistic situation, it is necessary that children are allowed to play the game freely for at least part of the test session, as opposed to working on tasks [3]. For specific functional parts it can be useful to add some small tasks, e.g. to test whether children know how to turn the volume down. However, a risk of giving tasks is that they can give away information about the game, which the children otherwise may not have found. For example, in Milo and the Magical Stones [10] there is a map to navigate more easily from one part of the game to another without having to repeat already finished games. Most children did not notice this functionality although it was explained in the introduction, resulting in frustration about having to repeat games to go to previously visited screens. When one of the tasks would have been to find the map and use it, the children would not have shown this frustration afterwards. This task should therefore only be given after the period of free play.

Reading

Children in this age group are starting to learn to read. Although the facilitator might sometimes like to skip some things by reading it for the children because it is faster, it is very frustrating for the children if you don't let them try to read for themselves. Accept that things like this will make the useful parts of your session much shorter than the duration of the session itself.

Finishing up

Making the children stop

At the end of the session the child has to stop playing the game for example to answer some questions or because another child is already waiting. Especially with games, it is sometimes hard to make children stop playing. When the session is almost over make sure you warn the child that he/she has to stop in a couple of minutes. Be firm and say something like: 'We are going to stop in several minutes.'

Sometimes it can be helpful to say that the child can play until a certain (sub) goal is reached.

Questionnaires

Often it may be necessary to ask children some questions after the test session, for example what they think of the game as a whole or how they liked to participate in the test. Unfortunately, asking children these additional questions after the session is difficult. Often children do not want to spend much time in the lab after you have made them stop playing the game. The questions should therefore be short and easy, and there should not be many questions.

Gifts

As a token of appreciation it is common to give the participating children a small gift. If you plan to do this, make sure you have some extra gifts ready. Often parents will bring siblings or friends to the lab and it would be disappointing if they would not get a gift.

If you want to give the children something to eat, get some alternatives for those children who cannot eat wheat, milk, chocolate, sugar, peanuts etc. due to allergies.

When testing at a school instead of a usability lab it is better to give a present to the whole group instead of the individual children. This way, the children who were not able or not allowed to participate by their parents will feel less disappointed.

Number of evaluators

As Jacobsen et al. [8] already described, no single evaluator will detect all problems when analyzing a videotaped usability test session. For the analysis of user tests with children this also holds, and this is the reason that at for most of our research the analysis is done by two evaluators. However, there is another reason to include more than one evaluator. Young children are often not very good in thinking aloud. Therefore, the evaluator has to interpret a lot of non-verbal behaviour and unfinished or vague sentences. When doing this alone the evaluator is often not able to determine the exact problem or alternative views. Discussing certain behaviours and verbalizations with another evaluator to create clearer problem reports, is definitely worthwhile.

ACKNOWLEDGEMENTS

This research was funded by the Innovation-Oriented Research Programme Human-Machine Interaction (IOP-MMI) of the Dutch government.

REFERENCES

1. Barendregt, W., Bekker, M. M., 2005. Development and Evaluation of the Picture Cards Method. *Workshop Interaction Design for Children, Interact 2005 Rome, Italy*,
2. Barendregt, W., Bekker, M. M., Bouwhuis, D. G., Baauw, E., 2005. Predicting effectiveness of children participants in user testing based on personality characteristics. Accepted for *Behaviour & Information Technology* (in press)
3. Barendregt, W., Bekker, M. M., Speerstra, M., 2003. Empirical evaluation of usability and fun in computer games for children. *Proceedings of Human-Computer Interaction INTERACT-03'*, 3 September 2003, IOS Press, Zürich, Switzerland, 705-708.
4. Bekker, M. M., Barendregt, W., Crombeen, S., Biesheuvel, M., 2004. Evaluating usability and fun during initial and extended use of children's computer games. *People and Computers XVIII- Design for Life*, Springer, Leeds, 331-345.
5. Boren, M. T., Ramey, J., 2000. Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication*, 43, 261-278.
6. BeoordelingsLijst Individuele verschillen tussen Kinderen (Blikvanger): persoonlijkheidsvragenlijst voor kinderen in de leeftijd van 3-13 jaar. (Assessment List Individual Differences between Children (Blikvanger): personality characteristics questionnaire for children between 3-13 years old), Version , 2002. [Computer software] Leiden: PITS.
7. Hanna, L., Risden, K., Alexander, K., 1997. Guidelines for usability testing with children. *Interactions*, 4, 9-14.
8. Jacobsen, N. E., Hertzum, M., John, B. E., 2003. The evaluator effect in usability studies: Problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, HFES, Santa Monica, CA, 1336-1340.
9. Lewis, J. R., 1994. Sample Size for Usability Studies: Additional Considerations. Human Factors, 36, 368-378.
10. Max en de toverstenen (Milo and the magical stones), 2002. [Computer software] MediaMix Benelux.
11. Nielsen, J., 1993. *Usability Engineering*, 1993. Academic Press Inc., Boston.
12. Nielsen, J., 1994. Estimating the number of subjects needed for a thinking aloud test. International Journal of Human-Computer Studies, 41, 385-397.
13. Rubin, J., 1994. *Handbook of usability testing: how to plan, design, and conduct effective tests*, 1994. Wiley, Chichester.
14. Virzi, R. A., 1992. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, 457-468.