Usability Evaluation with Children

Benedikte S. Als[1], Janne J. Jensen[2], Mikael B. Skov[1]

[1]Department of Computer Science
Aalborg University, Denmark
+45 96 35 80 80
{als, dubois}@cs.aau.dk

[2]Department of Communication Technology
Aalborg University, Denmark
+45 96 35 80 80
jjje@kom.aau.dk

ABSTRACT

Emerging technologies for children often require the involvement of children as test subjects in software development projects. Previous research studies have indicated that children behave differently than adults when involved in usability test sessions. Thus, children impose new opportunities and limitations in the evaluation and we still need to investigate proper and fruitful ways of involving children.

We present two studies on usability evaluation with children. The first study involved eight children in a development project on a mobile educational device. The children evaluated a number of different prototypes. The second study involved 60 children that participated in the evaluation of an existing mobile technology where the children applied either the think-aloud protocol or constructive interaction.

The results from the first study showed revealed that evaluation with a high-tech prototype does not necessarily provide more useful information, compared with an evaluation of a low-tech prototype. Our results from the second study revealed that constructive interaction provided the identification of more usability problems compared to think-aloud when the pair composition in constructive interaction is acquainted dyads.

Keywords

Children, usability testing and evaluation, think-aloud, constructive interaction informants

1. INTRODUCTION

The design and evaluation of children's technologies have received increased attention during the last several years [9]. Children should be considered individuals with strong opinions, needs, likes, and dislikes, and they should be treated as such [7]. Druin [6] provides a classification of involvement where children play the roles of users, testers, informants, or design partners. The four roles encompass different levels of engagement and impose different opportunities and limitations. All roles apply usability tests where children participate as subjects.

Usability testing has been studied extensively and is generally acknowledged to identify some of the key interaction problems in user interfaces [12]. Emerging technologies for children have produced the need for involving children as test subjects in software development projects. However, usability testing may be challenging with children situation since they are typically less organized [5]. Hanna et al. [10] propose adjusted guidelines for usability testing with children such as reflection on common target age ranges and how different age groups can verbalize their thoughts and feelings during a test.

Going through CHI Proceedings for the last ten years, we found several studies in which children participated as test subjects in usability tests applying e.g. think-aloud [1, 4, 6], constructive interaction [11] or both approaches [12]. Typically, the children are involved in the testing of existing technologies or high-tech prototypes [2, 3, 4, 8, 11, 12] whereas they are less involved in the testing or evaluation of low-tech or paper prototypes [12].

In this paper, we report from two different studies on usability evaluation with children. The first study involved eight children in a development project on a mobile educational device. These children evaluated a number of different prototypes from the development project. The second study involved 60 children that participated in the evaluation of an existing mobile technology where the children applied either the think-aloud protocol or constructive interaction.

2. STUDY A

This study used a revised form of a four-phased method used in the [12]. The revised form consisted of a method were all expert adult influence were removed from the first three phases. In the fourth phase we excluded all but the adult experts, who should comment on the correctness of the information provided by the children. The aim of the study was: How much information can we derive from children alone? And how valid is the information?

2.1 Participants

8 children (3 girls and 5 boys) at the age of 10 to 12 years old (*mean*=11.5, *SD*=0.76) participated. Five of these children had been living with diabetes for an average of 2,87 years. The adult participants had an average of 30 years working as a nurses; one had worked with diabetic children for 21 years the other for 1½ year. The children were not aware of that they would receive a small gift for their involvement in the study.

2.2 Procedure

The usability test sessions were conducted in a specialized usability laboratory. The laboratory integrated two rooms; an observation room in which the evaluations took place and a control room where one of the researchers would handle electronic equipment for recording the sessions. The two rooms were separated with a one-way mirror allowing people in the control room to see what was going on in the observation room.

The first usability evaluation was conducted with a low-tech prototype; it was made out of colored paper and plastic slides fitted to the same size as the screen of the PDA, so the children could get an idea of the size of the screen. For the second evaluation we used a high-tech prototype running on the PAD, it was developed in eMbedded Visual Basic 3.0.

2.3 Tasks

The children in the first session were given two tasks, tell us about diabetes, and tell us about your likes and dislikes regarding computer games and mobile phones. During the second session, with the same children, the children were presented with a paper prototype of the system, they commented on the idea and the design. We also presented them with 26 questions that could be implemented in a system, they were asked to answer the questions (pick the right answer from the four options) and comment on the questions.

In the fourth session the children were asked to solve the same tasks as the children in the second session. The only difference were that the prototype they were presented for were a running prototype. Afterwards the two adult experts solved the same tasks, thereby giving the team knowledge about the correctness of the information given by the children.

2.4 System

The system used in the experiment was a an edutainment system, the users were presented with a paper prototype of the system and a prototype which ran on a Compaq PocketPC

The target group of the system was children who had been diagnosed with diabetes and their friends. The system should teach the children specific information about diabetes, as well as entertaining the children, the reason for this being that it is easier to capture children's attention through edutainment systems, than it is through purely educational systems.(Note) The inspiration for the system came from the game "who want's to be a millionaire", which has one right and three wrong answers for each question. Thereby avoiding that the children should write an answer for each question.

3. STUDY B

Our experiment utilized a setup for comparison of think-aloud and constructive interaction for usability testing with children. In particular, we wanted to

1) Measure think-aloud and constructive interaction on identification of usability problems

2) Explore the impact of different compositions of pairs in constructive interaction

3) Analyze children's perception of the testing situations using think-aloud and constructive interaction.

We designed the experiment as a 3x2 matrix consisting of three types of sessions: individual testers using think-aloud, acquainted dyads (pairs) using constructive interaction, and non-acquainted dyads using constructive interaction. Furthermore, we configured the usability test sessions with same-sex dyads having sessions with girls and boys for each of the three setups.

3.1 Participants

60 children (30 girls and 30 boys) at the age of 13 and 14 years old (*mean*=13.35, *SD*=0.48) participated as test subjects in the experiment. The children were all 7th grade pupils from five different elementary schools in the greater Aalborg area. The children did not receive compensation for their involvement in the experiment.

3.2 Procedure

The purpose of the evaluation was explained in detail to the children and they were shown the facilities of the usability lab. Test subjects intended for roles as non-acquainted dyads were kept separate before the test sessions. The children received questionnaires on which they had to provide answers to such as age, name, school, and mobile phone experience.

The usability test sessions were conducted in the same usability laboratory as Study. All sessions were recorded on videotapes for later analyses including perspectives of the children and of their interactions with the mobile phone.

3.3 Tasks

The children were asked to solve twelve tasks one at a time addressing standard and advanced functionalities in the inno-100 mobile phone. This included making a phone call, sending a short text message, adjusting the volume of ring tones, and editing entries in the address book. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks. All children were able to solve all specified tasks. On average, the children spent 26:45 minutes (*SD*=06:39) on the twelve tasks.

3.4 System

The selected system for our experiment was an inno-100 mobile phone by innostream. This particular mobile phone was selected since it had not been released on the European market at the time of our experiment. Thus, all children would have to learn to use the mobile phone.

The inno-100 integrates a range of standard mobile phone features, such as making and receiving phone calls and short text messages, and more advanced features, including speed dial functions and options for creating personalized ring tones. The inno-100 has two separate screens with a main 128x144 pixel 16 bit color screen and 64x80 pixel sub screen on the cover. The navigation is primarily based on icons in the two upper menu levels. The lower levels are textual based including choice menus for setting values. Furthermore, the inno-100 integrates a number of games.

4. RESULTS

4.1 Study A

Our results indicate that the overall comments from the children testing a low-tech prototype are almost identical to the comments we provided by the children testing the high-tech prototype. The only difference we could find were that the second evaluation gave information about functionality errors.

From the usability evaluation of the low-tech prototype, we found that the children were capable of imagining a real system while looking at a paper prototype. Both the boys and the girls commented on what they thought could be funny features. We got suggestions as animation of face, changing the needles gender, music, reading the text out loud, giving the needle arms etc. The children understood most of the functionality that would be incorporated in the buttons, the girls had some trouble understanding and "quit" they suggested that quit should be replaced with a Danish word. The girls had a hard time imagining what could be search for in the game, and therefore they didn't understand the “search” button, after seeing the search page, they understood the function. When presented with the search page none of the children thought that it would be a good idea if the player should write the word themselves, since a person without diabetes wouldn’t know the words. Furthermore one of the boys suggested that a historical anecdote could be added to each topic, and one of the girls suggested the addition of pictures or small clips of film. As for the game, all the children liked that they had four possible answers to choose between, since it would be easier than of they had to write it themselves. The boys liked the idea of sticking the needle in the right answer; they suggested that the background had skin colored texture. All the children suggested that the game could be played as a multiplayer game, and the boys suggested that it should be possible to race the clock. During the talk over amount of questions one of the girls noted that if the computer would pick questions randomly, some of the most important information regarding diabetes might be missed.

The information provided by the children in the second evaluation was almost identically as the results from the first evaluation. The children however stressed that it was important that the voice of the needle was the voice of a child, since they didn’t want it to be an adult. All the children in this session had doubts on whether the “help” function would give them help to the quiz or how to play the game. The boys suggested that it should be possible to click on the needle during the quiz, to get some help if needed. The two children with diabetes didn’t like to stick the needle into the right answer, whereas the boys who didn’t have diabetes thought that the needle could be helpful to children who were afraid of needles since he looks so nice. The two boys without diabetes liked the search function, which was found a bit boring, by the two children with diabetes. During the quiz the girl accidentally hit the “next” button twice, and thereby answering the next question with the second tab. Additionally one of the children suggested that it should be possible to race each other with two linked PDA’s online. The two boys also suggested an idea for a game where they could control the needle and maybe shoot unhealthy food.

4.2 Study B

Our results indicated that constructive interaction provided the identification of a higher number of usability problems compared to think-aloud, but the differences were mostly not significant. However, we found significant influence of the pair composition in constructive interaction as the non-acquainted dyads identified significantly less problems than the acquainted dyads. The acquainted dyads identified more total numbers of problems and serious problems. However, this did not seem to increase level of frustration for the acquainted dyads. We further found that the girls identified more problems in constructive interaction as acquainted dyads compared both girls applying think-aloud and non-acquainted girls. No similar differences were found for the boys.

Specifically, our study resulted in the identification of 81 different usability problems. Based on a classification scheme, we classified 32 of these 81 usability problems as critical problems, 13 as serious problems, and 36 as cosmetic problems.

Our results showed that the sessions with acquainted dyads identified the highest number of usability problems of the three setups. The 12 acquainted dyads identified a total of 65 of the 81 usability problems whereas the non-acquainted dyads identified only 51 of the 81 usability problems and this difference was significant according to a two-tailed Chi-square test (χ^2=5.131, *df*=1, *p*=0.0235). The individual testers identified 56 of the 81 usability, but the difference between the individual testers and acquainted dyads was not significant (χ^2=2.090, *df*=1, *p*=0.1483) nor is the difference between the individual testers and acquainted dyads (χ^2=0.440, *df*=1, *p*=0.5069).

Looking at problem severity, we found that the acquainted dyad sessions identified nearly all critical problems (28 of the 32 critical problems), but this was not significant compared to the individual testers or the non-acquainted dyads according to a Chi-square test (χ^2=0.439, *df*=1, *p*=0.5076) (χ^2=2.286, *df*=1, *p*=0.1306). However, we found that the acquainted dyads identified significantly more serious problems than the non-acquainted dyads (χ^2=4.514, *df*=1, *p*=0.0336). Alternatively, no significant differences were found between acquainted dyads and individual testers on serious problems (χ^2=1.950, *df*=1, *p*=0.1626) nor between individual testers and non-acquainted dyads (χ^2=0.155, *df*=1, *p*=0.6940). We found no significant differences for the cosmetic problems.

5. CONCLUSION

This paper has reported from two different studies on usability evaluation with children. The first study involved eight children evaluating a number of different low-tech and high-tech prototypes. The second study involved 60 children in the evaluation of an existing mobile technology where the children applied either the think-aloud protocol or constructive interaction.

The first study revealed that evaluation with a high-tech prototype does not necessarily provide more useful information, compared with an evaluation of a low-tech prototype.

The second study revealed that constructive interaction provided the identification of more usability problems compared to think-aloud when the pair composition in constructive interaction is acquainted dyads.

ACKNOWLEDGMENTS

We would like to thank the participating children and their parents from the two studies..

REFERENCES

1 Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K. T., Stanton, D., Sundblad, Y., and Taxén, G. (2000) Designing storytelling technologies to encouraging collaboration between young children. In Proceedings of the Human Factors and Computing Systems CHI’00, ACM Press, pp. 556 - 563

2 Bers, M. U., Ackermann, E., Cassell, J., Donegan, B., Gonzalez-Heydrich, J., DeMaso, D. R., Strohecker, C., Lualdi, S., Bromley, D., and Karlin, J. (1998) Interactive Storytelling Environments: Coping with Cardiac Illness at Boston's Children's Hospital. In Proceedings of the Human Factors and Computing Systems CHI’98, ACM Press, pp. 603 - 610

3 Bruckman, A. and Edwards, E. (1999) Should We Lev-erage Natural-language Knowledge? An Analysis of User Errors in a Natural-language-style Programming Language. In Proceedings of the Human Factors and Computing Systems CHI’99, ACM, pp. 207 - 214

4 Danesh, A., Inkpen, K. M., Lau, F., Shu, K., Booth, K. S. (2001) Geney: Designing a collaborative activity for the Palm handheld computer. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 388 - 395

5 Druin, A. and Solomon, C. (1996) Designing Multimedia Environments for Children. Wiley & Sons, New York

6 Druin, A. (1999) The Role of Children in the Design of New Technology. HCIL Technical Report No. 99-23, University of Maryland, USA

7 Druin, A. (1999) The Design of Children's Technology. Morgan Kaufmann Publishers, Inc., San Francisco, CA

8 Ellis, J. B. and Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 474 - 481

9 Gorriz, C. M. and Medina, C. (2000) Engaging Girls with Computers through Software Games. Communications of the ACM, vol. 43, No. 1, pp. 42 – 49

10 Hanna, L., Risden, K., and Alexander, K. J. (1997) Guidelines for Usability Testing with Children. In interactions, September + October, pp. 9 – 14

11 Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W., and D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. In Proceedings of the Human Factors and Computing Systems CHI'02, ACM Press, pp. 299 - 306

12 Nielsen, J. (1993) Usability Engineering. Academic Press

13 Scaife, M., Rogers, Y., Aldrich, F., and Davies, M. (1997) Designing for or Designing with? Informant Design for Interactive Learning Environments. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 343 - 350