

A comparison of think-aloud and post-task interview for usability testing with children

Ester Baauw and Panos Markopoulos

Industrial Design, Eindhoven University of Technology

PO Box 513, 5600 MB Eindhoven, The Netherlands

+31 40 2475247

E.Baauw@tue.nl, P.Markopoulos@tue.nl

ABSTRACT

We describe an experimental study of different strategies for obtaining verbalization data, when conducting a usability test with children. Two software products were evaluated by 25 children of ages 9-11, at their school, using think aloud and post-task interviews. We have found confirmatory evidence, in support of earlier tentative results, that children reported more usability problems with think aloud rather than with post-task interviews and girls reported more problems than boys. However, when we counted also the usability problems identified through observation, we found no significant difference between the two methods or between the two genders.

Keywords

Evaluation, children, think-aloud, post-task interview.

INTRODUCTION

A minimal but essential part of designing interactive products for children is that they should test prototypes and early product versions. Tests can uncover usage problems and opportunities to improve interaction design. Usability-testing methods have been developed with adult testers in mind. Recently there has been a growing interest to adapt such methods for children or to develop usability-testing methods especially for children users. Involving children as usability-test-participants needs to accommodate for their developing capabilities at a cognitive and social level. One major issue in setting up a usability test with children is to decide what is the most appropriate way for obtaining verbalization data during the usability test (see [3]). Verbalization data will include utterances or self-report by the child. It complements observations by explaining the thoughts, the motivations and the problems encountered by the test-participant during a usability test.

A popular method for obtaining verbalizations from adult test-participants is the think-aloud protocol, where subjects verbalize their thoughts concurrently to the performance of

test-tasks. Think-aloud has two important drawbacks. The child-tester is expected to expend cognitive resources for both using the system under test as well as providing verbalizations. Second, the child is put in a socially awkward and uncomfortable situation, where she has to explain her interactions with the product under test, performing something akin to a monologue in the presence of an adult evaluator. The advantage of think-aloud is that test-participants do not have to recall their thought processes long after the usability test. An alternative to think-aloud is post-task interview where brief sets of questions are asked to the child after each task.

Donker and Markopoulos [1] found that the think-aloud technique originally developed for adults could work well also for children between ages 8-14. By comparing think aloud to post-task interview and written questionnaires, they found that think aloud helps uncover most problems. Further, it appeared that girls would help identify more problems than boys, though that study couldn't attribute this difference of performance to gender only. Here we report an experiment that aimed to test and extend these results. The next sections summarize this experiment and our initial conclusions from it.

METHOD

The two products tested

The first product tested was an educational game that helps children learn facts about biology. This was the same game tested in [1] to enable us to relate the two studies. The second product tested the children's section of the Dutch website of the World Wildlife Foundation. On the basis of the guidelines of [2] we anticipated several usability problems to be found with the website.

Tasks

In both cases, children were given a set of test-tasks to perform during the evaluation. Tasks were possible to execute within a few minutes and each task would touch on several potential usability problems.

PARTICIPANTS

11 boys and 14 girls, aged 9-11 participated in the study at their school in Best, the Netherlands. Permission was obtained from the parents. One girl was excluded from the analysis because of technical problems during the test.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IDC 2004, June 1-3, 2004, College Park, Maryland, USA

Experiment Design

We used a 2x2 within-subjects design. One within-subjects factor was the method used for obtaining verbalization data: think-aloud or post-task interview. In all other respects the evaluation method used was identical in the two conditions. The other factor was the computer application used (game or website). The order of the conditions was counterbalanced.

Dependent Variables

There were several measures taken; for reasons of space we discuss only the number of problems found.

Number of usability problems found. In order to compare which method works best, we counted the number of different problems that the two methods help identify. We considered as a problem any aspect of the system that makes it unpleasant, inefficient, difficult, confusing or impossible for the child to achieve their test-tasks. From the test sessions we obtained a mixed video stream of the on-screen activity and the child's face. The video footage was analyzed with the Observer software by Noldus. We coded this video using an adaptation of the DEVAN coding scheme [5], which includes codes for usability breakdown indications, based on both verbal and non-verbal behavior of the test-participants.

Procedure

Children performed the experiment at their school. They took part in the test individually. The first author acted as a facilitator. 6 boys and 6 girls tested first the website with think-aloud and then the game with post-task interview. 7 girls and 5 boys tested first the game with think-aloud and then the website with post-task interview.

RESULTS

We compared the average number of problems reported by a child (utterances) or instances where they needed help. A 1-tailed t-test, showed no statistical difference ($t(23)=-.92$, $p=.185$) between think-aloud and post-task interview. This was also the case when we included problems found by observation ($t(23)=.2$, $p=.422$). Similar results were found when we considered the website or the game separately.

This contrasts the findings in [1]. In that study though, utterances by the child in-between answering questions were not counted. By comparing the number of problems found on the game in think-aloud and when counting only answers to post-task interview questions (i.e., a strict execution of the post-task interview protocol), we found a significant difference by a 1-tailed t-testing ($t(22)=2.01$, $p=0.29$). This was also the case for the website ($t(22)=2.27$, $p=0.17$).

A 1-tailed t-test for independent samples, showed that girls, in general, report more problems than boys ($t(22)=2.63$, $p=.008$). However, when problems observed in the video analysis were included no significant difference was found ($t(22)=.13$, $p=.45$). Counting only problems that the girls reported (utterances) but not including cases where they

received help, we found that then girls indeed report more problems, but not significantly so ($t(22)=1.13$, $p=.136$).

These results show that girls received more help than boys in these sessions. One explanation could be an effect from the social interaction with the evaluator. A quick examination of the videotapes did not show an obvious difference in the conduct against girls and boys. Further analysis will be conducted to verify this. The second explanation relates to the differences in interaction patterns with girls and boys. Boys seemed to interact more opportunistically, trying out things to see what would happen, while the girls seemed to be trying to make sense of the interfaces and plan their actions. As a result boys seemed to be more efficient and to need less help. We believe this to be an interesting difference that we plan to study more thoroughly.

CONCLUSION

The analysis of our results is still under way. Our first results confirm our earlier findings [1], but also put them in perspective. In practice, evaluators will be inclined to use data from all possible sources (i.e. both observation and self-report) in which case both methods fare equally well. In such a case girls and boys seem as good at uncovering usability problems. However, it can be expected that a practitioner might not always be willing to invest time and effort to analyze tapes to obtain usability data as we did for our experiment. The post-task interview allows observation data and verbalization data to be obtained on the fly without analyzing tapes. Thus, post-task interviews can offer practical benefits at the cost of slightly longer evaluation sessions with children. Future work should examine these trade-offs and potential qualitative differences in the problems uncovered by each method and gender.

REFERENCES

1. Donker, A., and Markopoulos, P., A comparison of think-aloud, questionnaires and interviews for testing usability with children, In *Proceedings HCI 2002*, Springer, 305-316.
2. Gilutz, S. and Nielsen, J. *Usability of websites for children: 70 design guidelines*. Nielsen Norman Group: USA, 2002.
3. Markopoulos, P. and Bekker, M.M. On assessing usability testing methods for children. *Interacting with Computers*, 15 (3), Elsevier, 2003, 141-150.
4. Nielsen, J. *Usability engineering*. San Diego, CA: Kaufmann, 1994.
5. Vermeeren, A.P.O.S., Bouwmeester, K. den, Aasman, J. and de Ridder, H. DEVAN: a tool for detailed video analysis of user test data. *Behavior & Information Technology*, 21(6), 2002, 403-423.

