

Assessing the effectiveness of usability evaluation methods for children

Afke Donker

FACULTY OF PSYCHOLOGY
UNIVERSITY OF MAASTRICHT
Universiteitssingel 40, Maastricht, The Netherlands
Afkedonker@hotmail.com

Panos Markopoulos

IPO, CENTER FOR USER-SYSTEM INTERACTION,
EINDHOVEN UNIVERSITY OF TECHNOLOGY
Den Dolech 2, Eindhoven, The Netherlands
P.Markopoulos@tue.nl

ABSTRACT

The comparison of three Usability Evaluation Methods (UEM's) is reported with respect to the number of problems uncovered on software products, by children participants. The UEM's compared require different levels of verbalisation of the child that is performing the evaluation. Age of the children, gender, verbal competence, and extroversion level are things that could influence which UEM works best. The outcome of the experiment was that a UEM that requires the children to verbalise a lot is the most effective, and that girls find more problems in the program than boys.

KEYWORDS: Usability evaluation, talk aloud protocol, cooperative evaluation, comparing methods, methodologies.

INTRODUCTION

Several related studies have attempted to assess which Usability Evaluation Method (UEM) works best for adult users (see a seminal assessment of such studies in [3]). It is very well conceivable that for children, different methods will work better than for adults. Hereby we report the comparison of three usability evaluation methods or, more accurately, three protocols for the behaviour of children during usability testing : (a) *think aloud* [1] (b) *structured interview* with the evaluator during interaction as in cooperative evaluation [6], and (c) a *written questionnaire* which gets completed during the evaluation session.

DESIGN OF THE EXPERIMENT

The UEM's mentioned above were adapted in such a way that the only difference between them was the level of verbalisation they required of the children. On other aspects, like the number of children involved, tester intervention during the evaluation, etc., the methods are the same. All evaluations were done at the schools of the children, the experimenter was present and interfered to the same extent in all conditions and the children were asked to perform the same tasks on the same software product (Bio-Mania, a semi-educational game made by Transposia).

The effectiveness of UEM's was measured by counting the number of problems the children found. We determined what the children perceived as problems by asking them the following questions: (a) Did you need help solving the task? (b) Did the computer make it easy for you to do the task? (c) What, if anything, happened that you did not expect or want? Everything the children could not do without help was counted as a problem. The answers to the second question were subsequently disregarded because a lot of children gave a very vague and ambiguous answer to this question ("kind of easy "). What we did include as problems were the things that the children did not want or expect the system to do.

The verbal competence of the children was measured using four parts of the Wechsler Intelligence Scale for Children (WISC-R) [2] (information, understanding, similarities and vocabulary). These sub-tests test various capabilities, one of which is productive language proficiency that we were interested in. The extroversion of the children was measured by letting them complete the extroversion items of the *Amsterdamse biografische vragenlijst voor kinderen* (ABV-K: the Amsterdam biographical questionnaire for children). This test consists of yes/ no items of which we used twenty that related to extroversion.

PARTICIPANTS

Fourty five children of five different grades in three schools were asked to participate. The children were aged 8 to 14, with a mean age of 10 years and 2 months. Seven children were later excluded from the study. One was ill and with six there were technical difficulties to such an extend that the data became useless.

PROCEDURE

The experimenter was introduced to the class by the teacher and nine children from each class were selected (by the teacher or randomly). These children were first asked to complete the 4 sub-tests of the WISC-R (a paper and pencil version) and the extroversion items of the ABV-K. After that, they were randomly assigned to one of the three UEM's and they were asked to

evaluate the program individually. The other children were asked to return to the classroom until they were sent for. To counteract a possible social desirability bias [5] we stressed that we wanted to find as much awkward and illogical things in the program as possible. After the introduction, the experimenter demonstrated what children were supposed to do and the children had a chance to try it. They got some short feedback and then started the real evaluation. Children were thanked for their participation. Most of them enjoyed the experiment.

RESULTS

A linear regression analysis was performed. Only the gender of subjects and the high verbalisation UEM (talk aloud) had an effect. The regression model is shown in table 1.

Model	The number of problems found		
	B	SE B	β
1	Constant	7.000*	.813
	Talk aloud	7.500*	1.447 .654
2	Constant	8.484*	.913
	Talk aloud	7.478*	1.326 .652
	Gender	-3.509*	1.248 -.325

Table 1: Summary of Stepwise Regression Analysis for variables predicting the number of problems children found in the product (N= 38). Note, $R^2 = .427$ for Model 1; $\Delta R^2 = .106$ for model 2 ($p < .01$).

DISCUSSION

The 'talk aloud' protocol uncovered most problems than the other two methods. The number of problems found with those latter UEM's did not differ significantly. We had expected the results to be different for children of different abilities (different verbal understanding IQ and extroversion), but this was not the case. It could have been, however, that the children that found it difficult to talk aloud tried very hard to overcome their difficulties to make the experimenter happy, another example of the social desirability bias. However, the effort children invested does not explain why more problems are found in the UEM that requires children to verbalise a lot. This could be because talking more leads to reporting more problems, but it could also be because the children that used the high verbalisation UEM reported problems as they occurred, while the children in the medium and low condition waited with reporting the problems until the task was finished.

Boys on average found fewer problems than girls. This could be because girls have less experience on computers and find using one more difficult. Whitley

[7] reports that this is in fact the case, but the difference was not very large in the present study. Another explanation could be that girls do not like this kind of game. Whitley [7] found no evidence for the hypothesis that boys like computers better than girls, but it could still be that boys just like this particular game more and that they therefore find less problems.

CONCLUSION

Usability evaluation methods seem to be more effective when children are required to talk more. Hanna, Risdén, & Alexander [4] say that children have more trouble than adults in verbalising their thoughts. However we found that this apparent limitation is compensated by their motivation. Even when they find it difficult to do something they will try their best to do it anyway. Our study suggests that it might be better to use girls to test a product, as they found on average three and a half problems more than boys. Given the limited scale of the study, it is not yet wise to generalise from these conclusions to different products and different contexts of interactive system evaluation.

BIBLIOGRAPHY

1. Boren, T. and Ramey, J. (2000) Thinking aloud : reconciling theory and practice. In *IEEE transactions on professional communication*, 43(3).
2. De Bruyn, E.E.J., Vander Steene, G., and van Haasen, P.P. (1974). *Wechsler Intelligence Scale for Children - Revised, Nederlandse uitgave, Verantwoording*. Lisse: Swets & Zeitlinger B.V.
3. Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. In *Human-Computer Interaction*, 13(3), 203-261.
4. Hanna, L., Risdén, K., & Alexander K.J. (1997). Guidelines for usability testing with children. *Interactions September + October*, 9-14.
5. Manstead, A.S.R. & Semin, G.R. (1996). *Methodology in Social Psychology*. In: Hewstone, M., Stroebe, W., & Stephenson, G.M. Introduction to Social Psychology. Oxford: Basil Blackwell.
6. Monk, A., Wright, P., Haber, J., & Davenport, L. (1993). *Improving your Human-Computer Interface: A Practical Technique*. New York: Prentice Hall.
7. Whitley, B.E.(1997) Gender Differences in Computer-Related Attitudes and Behavior: A Meta-Analysis. *Computers in Human Behavior Vol 13(1)*, 1-22.