

Are non-expert usability evaluations valuable?

Raila Äijö and Jussi Mantere

Helsinki University of Technology, P.O. Box 2300, 02015 Espoo

raila.aijo@hut.fi, jussi.mantere@hut.fi

Abstract

The purpose of the study was to evaluate the efficiency of having non-expert evaluators evaluating an interactive web service independently in a remote location without recording or control conducted by the researchers. The results show that the non-experts were capable of reporting usability problems on the interface but had difficulties in reporting usability problems related to interactivity with the system. The non-expert evaluators give designers valuable information about users' individual reactions to the service while working in their own environment and they could be used as an additional resource in usability evaluations.

Key words: usability, evaluation methods, non-expert evaluation

1. Introduction

There is a need to evaluate the efficiency as well as the cost-effectiveness of different usability evaluation methods in order to be able to use efficiently limited time and resources in fast product development cycles. Previous studies have shown that although the aim of different usability evaluation methods (e.g. heuristic evaluation, user test) is the same, the actual results produced by them are different. (Karat 1997; Doubleday & al. 1997.) In resource-constrained situations when a human factors specialist is not available the role of non-expert evaluators can be valuable. Non-expert evaluators are defined as evaluators who don't have formal training in usability but might be experts on other fields. (Virzi 1997). Studies involving non-expert evaluators in usability evaluations have concentrated on using non-expert evaluators as part of an evaluation team (Wright & Wonk 1991), on highly structured and monitored remote non-expert evaluations (Castillo & al 1998) and on comparing evaluation methods in software development (Smilowitz & al 1994).

Previous studies have shown that the non-expert evaluations have been relatively inefficient and that if non-experts are used the number of evaluators should be increased (Virzi 1997). According to Nielsen (1993), novice users without training in usability are overall poor evaluators, HCI experts are 1,8 times as good and evaluators with expertise on the domain of the interface and on HCI are 2,7 times as good. In the study conducted by Smilowitz & al. (1994) the independent non-expert evaluators reported an equal amount of unique problems compared to the findings of user test but the reported problems differed qualitatively. Their main finding was that the independent non-expert evaluators were not able to report the high severity usability problems. However Smilowitz & al. (1994) conclude that independent non-expert evaluators may uncover unforeseen problems introduced in a real setting use and could be used as a cost-effective resource late in the development cycle. The study by Castillo & al. (1998) showed that if the non-expert participants are given minimal training in identifying critical incidents they are capable of reporting high, medium and low severity critical

incidents. In their study the evaluators reported the critical incidents by using a structured report which was generated by an automatic system when the user pressed 'report incident' button.

The purpose of this study was to evaluate the efficiency of having non-expert evaluators evaluating an interactive web service independently in a remote location without any control conducted by the researchers (e.g. recording). The aim was to understand what kind of usability problems the non-expert evaluators notice and how they are capable of reporting the problems they have experienced. In this report the results of non-expert evaluation are compared to the results of expert evaluation and a user test of the same interactive web site.

2. Methodology

This study is based on usability evaluations, which were performed on the interactive web service (<http://www.posti.fi/yksityis/paketti>) of Finland Post Ltd's parcel delivery services. The evaluated site included three interactive services:

- price calculator, to automatically calculate the price of a parcel of specific size
- delivery time calculator, to automatically calculate how long it will take to deliver the parcel
- tracking service, for the user to track down where the parcel is going

The site was evaluated by using simultaneously three different methods: remote non-expert evaluation, user test and expert evaluation. At the time of the evaluations (spring 2000) the service was already published but not widely in use. None of the participants had used it beforehand. What follows is a description of the evaluation methods which were used.

The non-expert evaluation: The non-expert evaluation was carried out by nine potential users of the evaluated service. They were either students of computer science or students of educational science. The non-expert evaluators had no training in usability but were active users of the Internet. They were instructed to report what kind of things they would like to have modified on the site in order to make it easier to use. The participants were also asked to think about how other users would feel on the site. The evaluators were not provided heuristics or scenarios to use as a structure while getting acquainted with the site. However both the non-expert evaluators as well as the expert evaluators had parcel codes available in order to be able to try the parcel tracking service. They worked independently in their normal working environment at their own pace and returned the evaluation in the way they preferred to. In previous studies the remote evaluation processes of evaluators have been commonly recorded or controlled by the researchers (e.g. Castillo & al. 1998, Hartson & al. 1996). In this study no recording was done by the researchers since the goal was to get the evaluators' own reaction to the service and their own reflection of the usability problems. Neither questionnaires nor interviews were applied during the evaluation process.

Expert evaluation: Expert evaluation of the site was carried out by four usability experts from Helsinki University of Technology. All the experts have several years of experience with usability evaluation, including special issues concerning interactive web sites. They were asked to evaluate the site and report the results in the way they wanted to. All of them had listed the problems and had also evaluated the severity of the problems. Two of them had used Nielsen's (1993) heuristics as a structure for the evaluation.

User test: Five users were invited to a user test of the site. According to Nielsen (1993), five users should be adequate to discover approximately 85% of the usability problems that can be discovered by user testing. The test users were selected as widely as possible from the potential users of the site. The test consisted of five test tasks which included the key services

of the site, i.e. finding out information about different kinds of parcels, how much it would cost to send a specific parcel and how soon the parcel would be delivered. The users were also asked to find out how a specific item should be packaged, and how they should mark the address on the parcel. The test sessions were videotaped.

3. Results

The data, which was captured by the three different methods, was analysed by using qualitative content analysis (Cohen & Manion, 1994). During the analysis no pre-defined structure was used. The usability problems identified and comments presented by the participants were classified in categories based on the content. The categories formed during the analysis are presented in Table 1.

Table 1. Classification categories formed during the analysis

Interface	Interactivity	Content	Other
<ul style="list-style-type: none"> - symbols - colours - fonts - overall layout 	<ul style="list-style-type: none"> - understanding input required from the user - understanding feedback received from the system - navigation 	<ul style="list-style-type: none"> - quantity of information - structure of information - terminology 	<ul style="list-style-type: none"> - error messages - download times

The focus during the analysis was on the qualitative aspects of the usability problems and not on the quantity because the frequencies and severity of the problems were reported differently by each of the group. However it can be concluded that based on the data the non-expert evaluators reported approximately 2/3 of the usability problems reported by the expert evaluators and captured during the user test. Overall, compared to the other methods the user test provided the most comprehensive analysis of usability problems on the site both quantitatively and qualitatively. In the following is a description of the differences between results captured by these three different methods.

There was a clear difference in the type of problems that the non-experts reported compared to the problems captured by the other evaluation methods. The non-expert evaluators had concentrated on evaluating the interface rather than on trying to evaluate the process of completing full tasks (e.g. sending a parcel, counting a price for a delivery). Most of the problems they had indicated were problems related either to the content and structure of the information, to the use of the symbols or to inconsistency in the use of links and back buttons. A clear finding is that the non-expert evaluators had reported usability problems only partly related to interactivity with the system although the main services on the site are based on interactivity. The usability problems in the interactive services were clear when the other evaluation methods were used. In the functions requiring interactivity the non-experts had reported problems in understanding input required from the user (e.g. problems in understanding how to fill in the details of a parcel) but not problems related to the feedback received (e.g. a understanding a description of delivery time) from the system. The low number of reported usability problems related to task completion is an indication that the non-expert evaluators were not able to imagine what it would be like to use the service in a real situation with real consequences. Some of the usability problems might have been missed because the evaluators didn't realise that they were in a problem situation. The evaluators might have also become frustrated and had stopped trying more easily than the participants in

the user test. During non-expert evaluation there was no possibility to ask the user to explain and clarify his feelings which could have been done during a user test.

While comparing the results of the non-expert evaluation to the results of the expert evaluation it was clear that the non-expert evaluators had evaluated the site only from their own perspective. They had reported problems that were problems to them but were not reflecting in any way that some things could be problems for other users. For example they had not considered elderly people as potential users of the site. An indication of this is that although the non-experts had clearly concentrated on usability problems related to the interface, only two of them had reported that the size of the fonts should be increased and that the colours did not support the use of the service. In addition, only two of them reported problems in understanding terminology used at the site. However, all these usability problems were indicated when the other evaluation methods were used. The small number of problems reported in relation to colours, fonts and terminology might also indicate that the non-experts had felt that not all the problems were valuable enough to report.

Compared to expert evaluation and to the user test a clear difference was that the non-experts were not able to describe in detail the usability problems they had found. The evaluation reports written by the non-experts did not include reflection of the possible causes of the experienced problems. Besides of that, the non-expert evaluators didn't provide any description of how the user proceeded to the situation he had reported as a problem. For example they might have sensed that something was not functioning properly in the parcel calculator but could not elaborate what was wrong and concluded by saying that it was not functioning.

In comparison to the user test the non-expert evaluation gives the designers feedback of the immediate problems (e.g. inconsistency in the use of symbols) experienced by the users in their own environment while using the service with different kinds of Internet connections. Previous studies have shown that in user testing the symptoms of the problems (i.e. an observed problem) are recorded whereas in heuristic testing the focus is more on identifying the causes (i.e. the reason for the observed problem) of the problems (Doubleday & al. 1997). This study showed clearly that the non-expert evaluators had focused even to a greater extent on the symptoms of the problems than the participants of the user test.

A clear finding was also that only the non-experts mentioned some positive findings of the evaluated site. Positive aspects were not stated by the experts and by the participants of the user test.

4. Discussion

Despite of the limitations the study shows that non-expert evaluators could be used as an additional resource while evaluating interactive web services. The non-expert evaluators give designers valuable information about users' individual reactions to the service while working in their own environment. In comparison to the other evaluation methods the focus of the non-expert evaluation is clearly different. While the focus of the experts is on concentrating on usability problems, the non-expert evaluation provides feedback which is based on users' experiences while freely exploring the service and reporting indications of problem situations. In comparison to the user test the non-expert evaluation provides valuable information of what kind of things on the site the users focus on without having special tasks to complete. This study indicates that the non-expert evaluators might be capable of providing feedback on problems in completing tasks that are important to them at the time of the evaluation but not on the services which they don't find relevant to them at the time of the evaluation. Involving

non-expert evaluators in usability evaluations could be more valuable if the group of evaluators had a real need to use the service at the time of evaluation. An indication of the narrower view of the non-expert evaluators is that none of them reported problems in compatibility between different services provided by the company (e.g. parcel service – letter service) which were indicated as potential problems in user test and in expert evaluation.

The results of this study indicate that the non-experts would have needed a structure to support their evaluation and encouragement in reporting all the ideas and problems they encountered, even minor ones. A specific scenario would have helped them to understand the usage situations (see also Smilowitz & al. 1994). The study shows that the non-experts were not able to imagine them in an actual usage situation while working alone. If structured more clearly the non-expert evaluation can be a valuable way to collect user feedback and user experiences as an addition to other methods. In this study the non-experts were not asked to evaluate the severity of the problems they found. That could have been helpful in order to understand the evaluators' personal evaluation scales. Overall it is more critical to combine the results of non-expert evaluations than the results of expert evaluations. The number of problems reported by each non-expert evaluator was low and there was clear variation in the type of problems they reported (see also Wright and Monk 1991). Nielsen (1993) has suggested that five expert evaluators can capture approximately 75% of the usability problems. Based on this study, if non-experts are used as evaluators the number of evaluators should be clearly higher than five.

References

- Castillo, J; Hartson, H. & Hix, D. 1998. *Remote usability evaluation: Can users report their own critical incidents*. ACM Proceedings of CHI '98 Conference on Human Factors on Computing Systems 1998, 253-354.
- Cohen, L. & Manion, R. 1994. *Research methods in education*. Routledge, New York.
- Doubleday, A., Ryan, M., Springlett, M & Sutcliffe, A. 1997. *A comparison of usability techniques for evaluating design*. ACM Proceedings of DIS '97 Designing Interactive Systems, 101-110.
- Hartson, R., Castillo, C. & al. 1996. *Remote evaluation: The network as an extension of the usability laboratory*. In Electronic Proceedings of CHI '96 Conference on Human Factors on Computing Systems, April 13-18 1996.
- Karat, J. 1997. *User-centred software evaluation methodologies*. In Helander (ed.) Handbook of Human-Computer Interaction. Second completely revised edition, 689-704.
- Nielsen, J. 1993. *Usability engineering*. Academic Press, San Diego.
- Smilowitz, E.; Darnel, M. & Benson, A. 1994. *Are we overlooking some usability testing methods? A comparison of lab, beta and forum test*. Behavior and Information Technology, vol. 13, no 1-2, 183-190.
- Virzi, R.A. 1997. *Usability inspection methods*. In Helander (ed.) Handbook of Human-Computer Interaction. Second completely revised edition, 705-715.
- Wright, P. and Monk, A. 1991. *A cost-effective evaluation method for use by designers*. International Journal of Man-Machine Studies, 35, 891-912.