International Remote Usability Evaluation: The Bliss of Not Being There

Mika P. Nieminen, Petri Mannonen, and Johanna Viitanen

Helsinki University of Technology, Department of Computer Science and Engineering,
P.O. Box 9210, 02015 TKK, Finland
{mika.nieminen,petri.mannonen,johanna.viitanen}@tkk.fi

Abstract. This paper describes the planning and implementation of a cross-border usability test that was to be executed in five European countries. The usability evaluation was designed by the Usability Group at Helsinki University of Technology who also performed the testing for the Finnish partner. In the other countries the usability tests were to be implemented by teams of subject matter specialists with very heterogeneous disciplines ranging from software engineering to social sciences, gender equality and vocational counselling. This paper describes the level of materials and training prepared for the remote usability testing and discusses its adequacy both via test personnel satisfaction and comments, and by comparing the found usability problems and observed phenomena in the test sessions between the test executed by the usability experts and the subject matter specialists.

Keywords: International usability testing, remote usability testing, localizing usability test materials, usability testing by non-expert evaluators.

1 Introduction

Usability testing has proven its worth as a crucial part of software engineering. Faster and wider communication mediums have made distribution of knowledge work, both geographically and temporary, an everyday practice. The change has created networked product development teams/communities and international cross-country organizations. Also usability engineering must be able to perform in these distributed surroundings. Most commonly applied method of usability engineering is conducting usability tests. It is often sighted that evaluating the usability of a product is very sensitive to the social, lingual or contextual environment where the testing is done. In many cases this has lead to increased costs, when international product testing must be executed in several countries by locally hired usability experts. The obvious alternative to boost the efficiency of international or multi-site usability testing has been to develop methods and procedures to conduct the usability tests remotely [1,2,3].

This paper describes the remote usability evaluation of an Internet portal for vocational counselling. The case project, funded by the European Commission's education and culture programme, produced a dynamic web site, which was to be

N. Aykin (Ed.): Usability and Internationalization, Part I, HCII 2007, LNCS 4559, pp. 388 – 397, 2007.

localized to each of the five partner countries. The development project had identified the need for good usability and its development tasks included usability evaluation of the portal's first functional version. Due to the fact that the portal was to have five different language versions with local content, the usability testing had to be performed locally at each partner country. The challenge was that only one of the project partners had usability expertise at their disposal. This lead to a situation where the tests were planned and coordinated by the members of the Usability Group at Helsinki University of Technology, while half of the actual user test sessions where executed in a distributed fashion by the local subject matter specialists.

In the literature remote usability testing is categorized based whether it happens in real time (synchronous or asynchronous) [3] and further whether the test data is collected automatically or presented by the users themselves [2]. Synchronous remote testing refers to monitoring the test via video connection to the test site or when possible by sharing the to-be-tested application via a broadband network connection using collaboration suites such as Microsoft Netmeeting™, Lotus Sametime™ or ShowMe™ from Sun Microsystems. The above classifications are combined in Fig. 1 with some examples of evaluation tools [4,5,6,7,8].

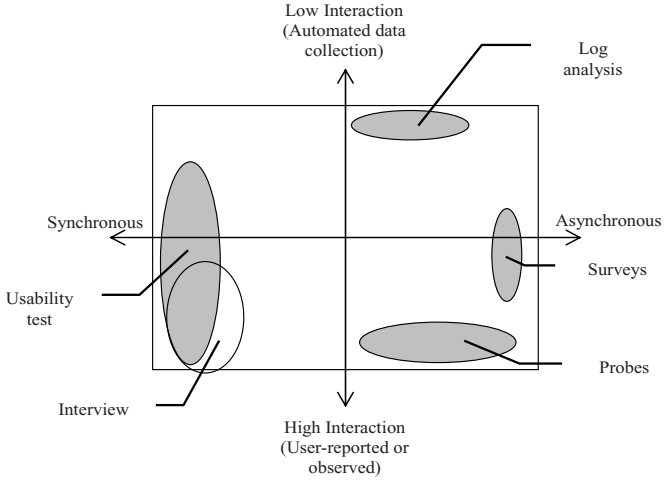

Fig. 1. Categorization of User – Evaluator interaction in remote usability evaluation activities

In this paper we describe an additional category of remote usability testing that is not only distributed geographically, but is also asynchronous. In our variation some of the user test sessions where organized and moderated by persons not fluent in usability testing based on a tailored test manual and a few hours of usability training. Our hypothesis was that with well-targeted instructions and minimal training the local personnel can manage the test sessions and with enough accuracy report the critical event they observe during the tests as shown by experiments conducted by Castillo et al [9,10]. Also by using the local subject matter experts (i.e. people fluent with the project's domain instead of usability engineering) our goal was to distill the cultural

anomalies that would have been incorporated to the analysis if we would have observed the user tests via a translator.

2 Evaluation Procedure

As stated earlier the evaluated system was an Internet portal for vocational counseling. Thus, the identified user groups were job-seekers (primary target group) and vocational counselors (secondary target group). Even thou, none of the project partners were native English speakers it was used as official language and the product development was done in English. During the development the portal was to be localized from the English development version into five different European languages: German, Danish, Slovenian, Romanian and Finnish.

The overall planning of the used usability evaluation methodology and the extent of the user tests were done in co-operation with the Austrian project coordinator and the members of the Usability Group at Helsinki University of Technology (later in this paper referred as "we"). Three user groups to be involved in testing were selected to be: low skilled job-seekers, medium skilled job-seekers and (high-skilled) counselors.

All together 17 users were to be involved in the distributed user tests. Table 1 illustrates the project partners and their planned number of test users.

Table 1. Breakdown of the participants to be included to the user tests

Target groups **Partners**	Low skilled	Medium skilled	Counselors
Project coordinator (Austria)			2
Partner 1 (Austria)	2		
Partner 2 (Romania)			2
Partner 3 (Denmark)			2
Partner 4 (Slovenia)	2		1
Partner 5 (Finland)	2	2	2

The complete usability evaluation procedure was to include the following seven stages:

1. Expert evaluations in Finland
2. Planning of the usability tests
3. Usability tests in Finland
4. Preparation of the test manual, including additional guidelines and checklists
5. Training sessions for the non-expert evaluators about practicalities of usability testing

6. Remote usability test sessions by the non-expert evaluators
7. Aggregation and analysis of the overall results.

Each of these stages is described in more details in the following chapters.

2.1 Expert Evaluations

In the first stage the expert evaluations were performed in Finland during June 2006. We performed expert evaluations for the English version of the portal using heuristic usability evaluation method and Nielsen's 10 heuristics [6]. Due to delays in the development process, some parts of the portals were missing and it could not be evaluated in its entirety. The result of the expert evaluation was a list of 94 prioritized usability problems including six catastrophe and 18 major problems. Usability problems were rated on a scale 1-4 with 4 being the most severe [6]. Conducting an expert evaluation in the beginning of the evaluation procedure allowed us to familiarize ourselves with the portal and find out its overall usability before planning the user tests.

2.2 Planning of the Usability Tests

After the expert evaluations we were able to plan the user tests. The user test were planned to be as simple as possible to help the remote test moderators to run the test sessions as easily as possible. Traditional usability test [6,11], using thinking aloud technique was chosen and its physical requirements were kept to a minimum. The test environment only needed a computer with Internet connection and a video camera to record observations for later analysis.

Two to three evaluators were to be present at each test session; a moderator was responsible for running the test session and other evaluator(s) were responsible for taking notes and observing the test situation. The test setting was explicitly designed not to require a fully furnished usability laboratory with a one-way mirror and multiple video recorders.

Test sessions were planned to last 45-90 minutes and consist of the following parts:

1. *An introduction*, during which the moderator briefly explains the user the test procedure and other ethical consideration [6] and asks the user to complete the pre-test questionnaire. The pre-test questionnaire requests background information such as personal details (age, sex, education and job description or study subject), and asks user open ended questions about her use of information technology and experiences with Internet services.
2. *The actual test tasks*. During the test tasks the user is asked to think aloud while performing the tasks. The user is given, one at a time, 13 tasks covering the core features of the portal, including seeking job descriptions and specific job details, conducting a skills test and an aptitude test, and searching information about available training from the portal. Five tasks included in the test setup required a modified scenario, only a variation in a few words, to cater to the different user groups; job-seekers and counselors.

3. *The debriefing after the tasks*. At the end a drawing assignment is given to the user and she is asked to draw the structure of the system as she recalls it. In the debriefing the moderator also goes through a prepared list of questions about the portal and asks the user for further comments.

2.3 Usability Tests in Finland

The first usability tests were conducted during October 2006. At that time the portal was still under development and it had not been localized into Finnish. All the test materials were in our native language and also the tests were carried out in Finnish. Altogether seven users were selected matching the characteristics of the target groups. A couple of days before the actual tests a pilot test was performed with one user, leaving six users for the actual tests. Our test sessions did take place in our usability laboratory, but the recording was done using a free standing digital video camera instead of the built-in equipment. In each test session three evaluators were present and sessions were conducted according to the set procedure. Afterwards the observation notes were completed and verified by reviewing the video recordings.

After the user tests the findings of the heuristic evaluation and usability tests were combined to produce a master list of usability problems. From those problems a total of 30 subsets were identified and rated using the same severity scale as earlier (catastrophic, major, minor or cosmetic). As a final result we delivered to the project coordinator a standalone evaluation report, in which the severity rated usability problems and corresponding suggestions for improvements were grouped according to the main parts of the portal.

2.4 Usability Test Manual

We used our tests as a baseline to provide the other partners detailed instructions with an accurate observation framework to be used to report the critical incidents in their tests. As said, for the Finnish tests all the test materials were in Finnish and so had to be translated into English for dissemination to the other project partners.

In addition to the actual tests materials, we also provided the project partners with a detailed test specification. This specification, or a usability test manual, did not only cover the actual test procedure, but also the underlining rationale for the specific test tasks. This additional contextual information was needed for the partners to accurately localize the test tasks and scenarios to their respective languages. So, the manual included detailed instructions how to set up and carry out a usability test for an Internet portal for vocational counseling:

- Resource estimate. The manual described briefly the human resources needed for testing, based on having two test users.
- Guidance for localizing the test instructions and tasks. For the remote evaluations all of the materials were to be localized to German, Danish, Slovenian and Romanian languages to match the local versions of the portal. In the manual we argued the reasons why to localize the materials and instructions and how to do it.
- Setting up a usability test. We gave simplified instruction for the non-expert evaluators about how to carry out usability tests including physical test setup and recording, selecting test users, planning test's timetable, and test procedure

(including five phases: preparation, introduction, during the test tasks, debriefing and after the tests). All the necessary test materials were appended to the test manual: a background questionnaire to be filled in by the user at the beginning of the test, individual test tasks to be handed to the user during the test, scenarios and test tasks to be used by the moderator during the test, the drawing assignment, debriefing interview questions, and a checklist for the moderator of a usability test.
- Performing a pilot test before the actual tests.
- Framework for observing and analyzing the success during the tests, and reporting the test data. The local testers were given an observing guide (an observation form with example data and points of interest about the test tasks), which briefly presented a simple analysis strategy for studying the success in tasks. The observation guide and observation form were to help the remote test personnel in making observation notes during the test and reporting findings to project coordinator.

The hypothesis was that by providing comprehensive enough instructions the reported data from the local partners' tests would be comparable and valid.

2.5 Training Session

In mid-October 2006 we provided the project partners a very concise, about five hour training session or introduction to usability testing. The training session was carried out during a project workshop in Graz, Austria and it emphasized the practical side of usability testing and mainly tried to increase the partners' awareness to usability issues. All project partners responsible for testing were present. The training agenda was based on our usability test manual. We gave the partners very brief examples of expected results and experiences from our already conducted usability tests. The actual report of our test results was not delivered to the project partners prior to their respective test sessions. Thus, their observed phenomena and found usability problems were not influenced by our results.

2.6 Remote Tests in Other Partner Countries

The remote international usability tests with 11 users were to be executed without our participation in five European countries solely based on our usability manual and training. Local test moderators interacted with native participants in their respective native language in the local contexts. They were responsible for implementing and running and recording the tests, and reporting the findings. The few qualifications for the local testers were being native speakers of their local languages, attendance to our training session and fluency in written English.

Usability tests in partner countries were carried out in November 2006. The project coordinator reported us the following about the tests:

- The project coordinator had performed tests with one low skilled job-seeker and two counselors using the German language version of the portal. According to the partner all materials, including the checklists, were translated to German before the tests. Two evaluators were present at each test.

- The second Austrian partner took advantage of the already translated materials provided by the project coordinator and also tested the German language version of the portal with two low skilled users.
- Due to scheduling problems the partners in Romania and Denmark could not conduct usability tests at all.
- The Slovenian version of the portal was not completed in time. Instead the tests in Slovenia were conducted using the English version of the portal with one low skilled job-seeker and one counselor. In consequence of not having a localized version the Slovenian partner reported having had problems related to terminology during the tests.

2.7 The Overall Results

The project coordinator was responsible for collecting all the test data from all the evaluations and then analyzing and aggregating the final results. The overall test data consisted of results and suggestions for improvements provided by us (usability inspection and tests with six users) and test data from the local tests provided by other project partners (usability tests with seven users). These aggregate results are to be made a few months after the writing of this paper.

3 Reliability and Validity of the Test Results

As mentioned in the previous chapters we prepared an observation form for the non-expert evaluators to help them report their findings. The provided forms were filled for every one of the realized 13 test sessions. We have used these observation forms to compare the results from the test sessions executed by both usability experts (us) and persons not fluent in usability testing methodology. While the forms did not give us the full richness as if we had attended the usability tests ourselves, they did mark us the critical incidents and gave a rough picture of the tests in general.

The following Table 2 summarizes the main differences between expert and non-expert evaluators when reporting the test data and interacting with the users during the tests.

Table 2. Differences between expert and non-expert evaluators when reporting the test data and interacting with the users during the tests

Type \ Evaluator	Non-expert Evaluators	Expert Evaluators
Reporting the observations	Reported the exact user behavior as a sequence	Reported the user actions in relation to the overall goal
Reported critical incidents	Reported equally all incidents, emphasis on positive comments	Reported incidents relating to usability problems, emphasis on negative comments
Quality of the reported observations	Heterogeneous between the different partners	Uniform among the usability experts
Interaction with the users during the test	Frequent interaction with the users, several assists during a test	Very minimal interaction with the users, assistance involving foreign terms

As the first line in the Table 2 shows we managed to make more observations about the reason why the users did what they did during the test. For instance when we reported how the users interpreted some element in the user interface the non-expert evaluators only reported that the users had difficulties with the element.

The cause for the difference can be explained by the major difference in the observers' experiences with usability testing i.e. their moderator skills. Other option is the individual differences in the users' ability to think aloud or the non-expert evaluators' inability or reluctance to promote the users' thinking aloud.

All in all the results from both experts and non-experts are very consistent. Our findings (based on 6 user tests) cover almost 90% of all the test observations. Similarly the remote tests reported over 70% of our results. All the critical and major usability problems where reported by both groups, except for those arisen from lack of interaction (see the navigation bar example in the following paragraph). Thus, the results from remote tests validated our findings with very good accuracy. In addition there seems to be only a few culture or language specific usability problems.

The single most interesting difference in the observations was the usability problems relating to the portal's navigation bar (including a navigable bread crumb trail) depicted in Fig. 2. In our evaluations none of the test users grasped the functionality of the navigation bar's bread crumb trail and actually only a few noticed or commented the whole bar at all during the tests. In the other hand majority of the remote test users were reported to use the navigation bar, but it is unclear from the reported incidents whether they navigated thru the actual bread crumb trail.

Fig. 2. The portal's navigation bar, with a bread crumb trail

Based on the usability evaluations the portal's user interface was redesigned. The navigation bar was relocated to the top of the screen and its look and feel was changed dramatically. Another major redesign was the removal of the wizard-like bread crumb trail (both from the navigation bar and the job pages) and the introduction of tabbed browsing to bind each job description into a more concrete unit.

A small online survey was done before (N=13) and after (N=30) the user interface changes. In survey, among other questions, three questions where asked relating to the user interface:

1. How would you rate the design of the site?
   (5=Excellent, 1=poor)
2. How would you rate the clarity of the page structure?
   (5=Excellent, 1=Poor)
3. Was it easy to find the information you were looking for?
   (3=Yes, 2=Half-and-half, 1=No)

The below Fig. 3 shows the averages of the survey answers. Even though the survey reached only a relatively small number of people the changed towards better

(or yes) is clear. Uncannily, the improvement in both the design and the clarity is almost equal.

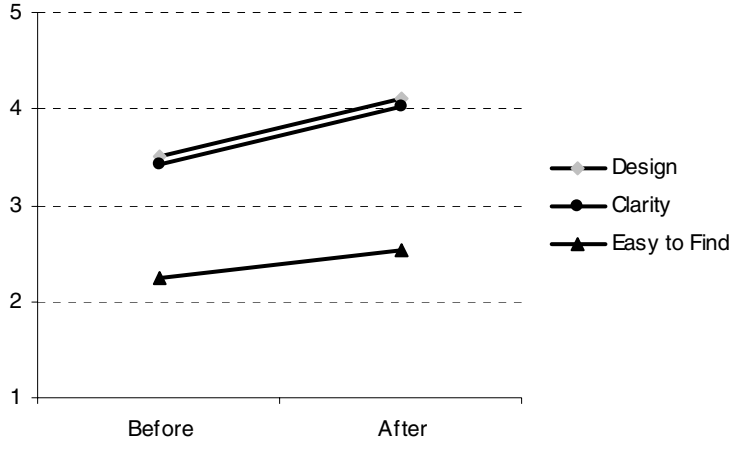

Fig. 3. Before and after user ratings for the portal design, clarity of structure and easiness to find information

4 Conclusions

Remote usability evaluations in an international context can be either very costly or low on results. General guidelines emphasize the risks and obstacles involved in international testing at a distance and guide practitioners towards very traditional, safe and therefore costly usability evaluation methodology [3]. In our study the local subject matter specialists, who were not familiar with usability engineering, were harnessed to execute usability tests in additional two European countries. These non-experts were successful in conducting remote usability tests i.e. recruiting users, organizing and moderating the tests and reporting their observations. The non-expert test moderators perceived the offered materials and training adequate for them to perform the tests. However, our analysis indicates that the observation form provided for the test personnel might have been too suggestive and thus slightly biased the made observations.

Analysis of the reported observations revealed that the results from expert and non-expert tests supported each other almost perfectly. This is in line with earlier studies where minimal training has been reported to give non-experts adequate knowledge to identify, report and rate the severity levels of usability problems they encountered [9,10].

The main difference between the observations of non-experts and experts were the ability to see the big picture (e.g. overall goals) and to produce reports of uniform quality. As suggested by Bojko et al. [12] more test situation training might have helped to make more accurate observation and include also the observers own

interpretation of the user actions. However, it would have also made the process more cumbersome and more expensive.

All in all, the process undeniably produced a better version of the career portal, and the remote test results validate that the portal caters also to the needs of the users in all the partner countries. This shows promise that non-expert personnel can be effectively utilized to carry out usability tests with only minimal training, presupposing there is an experienced usability team coordinating the evaluation.

Acknowledgement. The authors of this paper wish to acknowledge the participation, funding and support of the Leonardo ICT CTO project and the persons therein that conducted the remote usability tests and allowed us to compare their observations to ours.

References

1. Thompson, K.E., Rozanski, E.P., Haake, A.R.: Here, There, Anywhere: Remote Usability Testing that Works. In: Proceedings of the 5th conference on Information technology education (SIGITE'04), pp. 132–137. ACM Press, New York (2004)
2. Krauss, F.S.H.: Methodology for remote usability activities: A case study. IBM Systems Journal 42(4), 582–593 (2003)
3. Dray, S., Siegel, D.: Remote Possibilities? International Usability Testing at a Distance. Interactions Journal 11(2), 10–17 (2004)
4. Gaver, W., Dunne, T., Pacenti, E.: Cultural probes. Interactions 6(1), 21–29, ACM Press, New York (1999)
5. Mattelmäki, T.: Design Probes. University of Art and Design Helsinki, Helsinki (2006)
6. Nielsen, J.: Usability Engineering. Academic Press Inc., New York (1993)
7. Dix, A., Finlay, J., Abowd, G., Beale, R.: Human-Computer Interaction, 3rd edn. Pearson, London (2004)
8. Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., Carey, T.: Human-Computer Interaction. Addison-Wesley, New York (1994)
9. Castillo, J.C., Hartson, H.R., Hix, D.: Remote Usability Evaluation: Can Users Report Their Own Critical Incidents? In: Proceedings of the Conference on Human Factors on Computing Sytems (CHI'98): Summary, pp. 253–254. ACM Press, New York (1998)
10. Hartson, H.R., Castillo, J.C.: Remote Evaluation for Post-Deployment Usability Improvement. In: Proceedings of the Conference on Advanced Visual Interfaces (AVI'98), pp. 22–29. ACM Press, New York (1998)
11. Dumas, J.S., Redish, J.C.: A Practical guide to usability testing. Greenwood Publishing Group Inc., USA (1999)
12. Bojko, A., Lew, G.S., Schumacher, R.M.: Overcoming the Challenges of Multinational Testing. vol. 12(6), pp. 28–30 (2005)