Comparative Evaluation of Usability Tests

Rolf Molich, DialogDesign, Skovkrogen 3, 3660 Stenlose, Denmark, molich@acm.org

Ann Damgaard Thomsen, Kommunedata, Lautrupparken 40, 2750 Ballerup, Denmark, adt@kmd.dk

Barbara Karyukina, SGI, 890 Industrial Blvd., Chippewa Falls, WI 54701, USA, barbarak@cray.com

Lars Schmidt, Networkers, Vermundsgade 40, 2100 København Ø, Denmark, ls@networkers.dk

Meghan Ede, Sun Microsystems, Inc., 1601 Willow Rd., 17 Network Circle, MPK17-101, Menlo Park, CA 94025, USA, meghan.ede@sun.com

Wilma van Oel, P5, Overtoom 283, 1054HW Amsterdam, The Netherlands, w.vanoel@p5-adviseurs.nl

Meeta Arcuri, Hotmail, Microsoft Corporation, 3590 N. First Street San Jose, CA, meeta@hotmail.com

ABSTRACT

Seven professional usability labs and one university student team have carried out independent, parallel usability tests of the same state-of-the-art, live, commercial web site. The web site used for the usability tests is www.hotmail.com, a major provider of free web-based e-mail. The panel will discuss similarities and differences in process, results and reporting.

Keywords

Usability testing, evaluation, usability test problems, thinking aloud, usability report, usability problem description, hotmail.

BACKGROUND

In early 1998 four professional usability labs performed independent usability tests of a Windows calendar management application, Task Timer for Windows. The results of this comparative study was published in [1]. The study is called "Comparative Usability Evaluation", CUE-1.

All usability tests were carried out by experienced usability professionals employed by the labs.

The study showed that there were remarkable differences in approach, reporting, and findings between the labs. The most interesting result was perhaps that while a total of 141 usability problems were uncovered by the four labs, only one problem was found by all four labs. Another problem was found by three labs. Eleven problems were reported by two labs. Each of the remaining 128 problems were reported by only one lab.

Another interesting result was the considerable difference in approach. Some teams used a quantitative approach to usability testing, focusing mainly on product acceptance in the marketplace. Other teams used a qualitative approach to usability testing, focusing mainly on usability problems.

We also found that the usability reports generated by the labs differed considerably from each other. They also differed from the recommendations presented in some of the recognized textbooks in the field, like [2] and [3].

The study has generated considerable interest. Therefore, a number of other professional usability labs decided to undertake another, similar study in the fall of 1998.

PURPOSE

The purpose of the Comparative Usability Evaluation 2 (CUE-2) is to

a. Provide a survey of the state-of-the art within professional usability testing of web-sites.

b. Set a benchmark against which other usability labs can measure their usability testing skills.

c. Investigate whether the general trends that appeared from the first CUE study can be replicated

d. Show participating usability labs their strengths and weaknesses in one of the core processes of the usability profession through non-offensive self-assessments of usability testing.

e. Provide the basis for an interesting and entertaining panel discussion at CHI99.

Basic Rules

a. Each participating usability lab will conduct a usability test of the selected web-site and write a report of the test.

b. All usability test reports will be anonymous and publicly accessible.

c. Each participating organization will cover all of its own expenses in connection with the Comparative Usability Evaluation.

d. The purpose of CUE-2 is not to select a "winner", or to criticize the approach or the findings by any of the participating labs.

How CUE-2 Extends CUE-1
The main improvements in CUE-2 over CUE-1 are:

*   CUE-2 tests a state-of-the-art web site instead of a four year old Windows 3 program. In CUE-1, user expectations for calendaring programs and GUI interaction in general far exceeded the capabilities of the test product.

*   Access to development team representatives during the usability test by e-mail. The questions and answers were recorded in order to provide a log of the interactions.

*   Improved mission statement (scenario). The teams in CUE-1 were not given a high level task set to focus on. Consequently, they all tested different parts of the user interface. This meant that the resulting usability data collected in the reports was hard to compare.

*   Student teams participate on an equal foot with the professional teams in order to examine the difference between professional usability testing and usability testing carried out by inexperienced university students taking a basic usability course.

COMPARATIVE EVALUATION PLAN
The usability tests took place in the fall of 1998.

Each usability lab carried out a "normal" usability test of the Hotmail web site and reported the results in a usability report.

Each lab used its standard usability report format with one exception: The identity of the lab is neither directly nor indirectly apparent from the report.

In addition, each usability lab reported
- Deviations from its standard usability test procedure.
- Resources used for the test (person hours).
- Comments on how realistic the exercise has been.

The participating usability labs did not communicate during the test period.

The web site used for the test, www.hotmail.com, was selected by the panel organizer in close cooperation with the Advisory Committee. The name of the web-site was disclosed to the participating labs exactly three weeks before the reports were to be delivered to the panel organizer.

After all tests had been completed and the test reports had been received by the panel organizer, copies of all reports were distributed to each of the participating labs. The reports will also be made publicly available on the World Wide Web. Details will be provided at the panel session.

Other teams
Two more usability labs have participated in the evaluation. For reasons of space, representatives from these labs will not appear on the panel.

*   Joseph Seeley, NovaNET Learning, Inc., Champaign, IL 61820, USA, seeley@nn.com
*   Kent Norman, Dept. of Psychology, University of Maryland, College Park, MD 20742-4411, USA, kent_norman@lap.umd.edu

The members of the university student team were:
*   Torben Nørgaard Rasmussen, Asbjørn Johansen, and Tue Nørgaard. Faculty advisor: Christian Gram, Dept. of Information Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark, cg@it.dtu.dk

We wish to acknowledge significant contributions to the usability tests by the following individuals: Klaus Kaasgaard (Kommunedata), and Roel Kahmann (P5).

Advisory Committee
The panel organizer (Rolf Molich) has been advised by an Advisory Committe consisting of:
*   Nigel Bevan, Serco Usability Services (UK)
*   Scott Butler, Rockwell Software (USA)
*   Erika Kindlund, Intraspect, Inc. (USA)
*   Jean Scholtz, NIST (USA)
*   Bonnie John, Carnegie-Mellon University (USA)

PANEL CONTENTS
The panel will present the results of CUE-2. The Hotmail user experience manager will participate in the panel discussion and represent the client perspective.

The panel will discuss similarities and differences in process, results and reporting. The panel will discuss the difference between usability testing and good usability testing.

The purpose of the panel is to show different approaches to the usability testing task, and to discuss whether professional usability testing is an art or a mature discipline that turns out reproducible results.

REFERENCES
1. Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J., and Miller, D.   Comparative Evaluation of Usability Tests, in *Proceedings of UPA98 (Usability Professionals Association 1998 Conference)* (Washington DC, June 1998), UPA, 189-200.
2. Dumas, J.S., and Redish, J.C.   *A Practical Guide to Usability Testing*. Ablex, Norwood NJ, 1993
3. Rubin, J. *Handbook of Usability Testing*, John Wiley, New York NY, 1994