CHI2003 Workshop

# Comparative Expert Reviews


## ABSTRACT

In this workshop we will try to obtain a better understanding of the strengths and weaknesses of the expert review and heuristic inspection methods. We will do this by comparing results of independent expert reviews, heuristic inspections and usability tests of the same state-of-the-art website carried out by participating expert usability professionals.



### Keywords

Usability evaluation methods, heuristic inspections, expert reviews, evaluator effect


## ORGANIZERS


Rolf Molich (Primary contact person)
DialogDesign
Skovkrogen 3
DK-3660 Stenlose, Denmark
Email: molich@dialogdesign.dk

Robin Jeffries
Sun Microsystems
17 Network Circle MS114
Menlo Park, CA, 94025
Email: robin.Jeffries@sun.com


## WORKSHOP GOALS

1. Obtain a better understanding of the strengths and weaknesses of expert review methods.
2. Set a benchmark against which other usability professionals can measure their expert review skills. We will do this by publishing all test reports, essentially allowing outsiders to evaluate the website on their own and compare their results with ours.
3. Provide a survey of the state-of-the-art within professional expert reviews and heuristic inspections. The applied methods and the submitted reports will provide a survey of the techniques that are actually being used by experts today.
4. Show participating usability professionals their strengths and weaknesses in usability review, which is one of the core processes of the usability profession.
5. Experiment with consensus building methods in order to determine whether consensus building among experts increases the quality of the review results.
6. Analyze the differences in the expert findings in detail in order to propose changes or important caveats to the method.

## COMPARATIVE USABILITY EVALUATIONS (CUE-X)

If accepted, this workshop will constitute the fourth comparative usability evaluation, CUE-4.

The basic idea behind all CUE studies has been to assemble a number of leading professionals physically or virtually, and ask them to work on the same usability problem. By comparing their solutions and discussing interesting differences in approach or result, we have gained important insights into the state-of-the-art within our profession.

CUE-1 was a comparative usability test of a Windows calendar application carried out in March 1998 by four professional teams [6]. The study generated a number of interesting results and set the stage for CUE-2.

CUE-2 was a comparative usability test of www.hotmail.com carried out in late 1998 by nine professional teams from all over the world [2,7]. The study was very successful because it showed that state-of-the-art websites contain a huge number of usability problems, and because it showed that many usability professionals make serious errors when conducting a usability test.

CUE-3 was a comparative test of expert evaluations conducted in September 2001 with 12 Danish usability professionals [4]. The study was intended as a pilot study. It was not completely successful, because the website that we evaluated (www.avis.com) contained too many trivial and obvious usability problems, and because we had not realized the importance of a strict consensus-building process.

## USABILITY EVALUTION METHODS

### Definitions

An expert review is an ad hoc method used by an expert usability professional to evaluate a user interface. The only thing you can say about it is that it doesn't require users other than the reviewer(s).

A heuristic inspection is the application of a set of commonly recognized usability heuristics by someone without particular knowledge of usability engineering to evaluate a user interface

[1,3,8]. The method can of course also be applied by usability professionals, which is what we will do in this workshop.

### Historic Overview

Usability studies and expert reviews have existed for as long as the field of human-computer interaction has. Heuristic inspection, as a method that could be applied by non-specialists, was invented by Nielsen and Molich [8] in 1990. Other inspection methods (methods that do not involve observing the target population using the application) followed, resulting in research comparing their efficacy. [3] discusses many of those papers and covers the challenges of doing this research well. The conclusions of the early work were primarily:

- Inspection methods tended to result in different problems being found by different evaluators. Agreement as low as 10% overlap between evaluators has been replicated several times [4,5].
- Many inspection methods turn up low severity problems, relative to usability testing [5].
- False-alarms (reporting usability problems that are either incorrect or are of such extremely low severity as to be worthless) has been a concern for several inspection methods [9].

More recently Cockton and Woolrych [1] have done an extensive study of heuristic inspection, in particular, comparing it to usability testing. They found that inspections led both to many missed problems (based on a plausible derivation of the full set of usability problems in the application) and many false alarms, and that they only found surface problems -- those that could be discovered with a few mouse clicks. Cockton and Woolrych cite expert review as the gold standard against which other inspection methods must be compared.

## WORKSHOP PLAN

### Workshop participants

1. At most twenty people, including the workshop organizers, will be admitted to the workshop based on duly submitted applications. The workshop organizers may invite selected, prominent members of the international usability community to the workshop.

2. Each participant must agree to carry out an evaluation of a website selected by the organizers. The evaluation can either be an expert review, a heuristic inspection or a usability test with at least five users. Each participant can choose freely among the methods. The organizers will ensure that all methods are equally represented.

3. At most two people from any given company can participate. If two people from the same company participate, they must still carry out independent evaluations.

4. Minimum qualifications to participate: Must have at least five years' experience with professional usability work. Must have conducted at least five professional usability tests.

5. Participants must agree to have their usability reports published. Published reports will be anonymous unless participants explicitly agree to the opposite.

6. Each participating organization will cover all of its own expenses in connection with the workshop and they must pay the CHI2003 workshop fee.

7. In their application, participants must indicate whether they want to conduct an expert review, a heuristic inspection or a regular usability test with at least five participants.

8. The list of participants will remain secret until a few days before the workshop. In other words: While they conduct their review or inspection, participants will not know who else participates in the study.

### Before the workshop

1. At least five weeks before the workshop:
   The organizers will select a website that is suitable for evaluation. The criteria for selecting the website are:
   a. It must be state-of-the-art with respect to usability. It must not contain a significant number of trivial usability problems.
   b. The target group for the website must be the general public. In other words, the prospective attendees of the workshop must have no problems identifying themselves with the target group of the website.
   c. We will try to find a website where one or more members of the project team, in particular user experience specialists, are willing to attend the workshop.
   d. The language of the website must be English.

2. Four weeks before the workshop:
   Accepted participants will receive a scenario by email. The scenario includes the URL of the website and the context for the evaluation.
   Participants are expected to spend between 2 and 20 hours evaluating the web site. They will have one week before they have to hand in their results. They must use the chosen evaluation method. Teams that use the heuristic inspection or expert review methods must not involve end users of the website.

Participants must not attempt contact the website project team during the study. Any questions for the project team that might arise during the evaluation should be included in the report.

3. Three weeks before the workshop:
   Each participant must submit an individual, anonymous report to molich@nngroup.com. The report must be in a format that is defined by the organizers.
   Participants who are doing a regular usability test will have two weeks to complete their report, which is due two weeks before the workshop.
   A participant who does not submit a proper report in time will not be admitted to the workshop.

4. One week before the workshop:
   All test reports are made available to all participants on the world wide web.

**Report format**

The evaluation report format is strictly prescribed since the topic of differences in reporting format was covered adequately by the CUE-1 and CUE-2 studies. Evaluation reports must be written in English. A detailed report template and examples will be provided with the scenario.

An evaluation report must contain:

1. One page title, author and affiliation.

2. One page executive summary.
   - 2-3 most important problems
   - 2-3 most important positive findings
   - High-level usability recommendations aimed at the project manager, for example "Allocate more resources to quality assurance," or "Arrange site visits for project team members."

3. Comments on the website.
   - Each comment must be classified as a problem description or a positive comment.
   - Each problem must be rated on three scales:
     - Frequency (how often does the problem occur),
       1. Rarely
       2. Often
       3. Sometimes
     - Impact (how serious is the problem when it occurs),
       1. Minor (delays users briefly),
       2. Serious (delays users significantly but eventually allows them to complete the task),
       3. Catastrophic (prevents users from completing their task)
     - Persistence (will users learn how to get around the problem).
       1. Yes, quickly
       2. Only after encountering the problem several times
       3. No

   A gross way to determine whether a problem should be fixed could be to add the three numeric values and see whether the problem gets a high score.
   Participants should include as many comments as they would consider appropriate in an industrial setting.

4. Answers to specific questions about
   - Resources used (person hours, staff),
   - Specific evaluation methodology,
   - Quality assurance techniques employed, for example "peer review of the report."
   - Estimated confidence in the validity and completeness of the results.

   Participants will receive the specific questions together with the scenario

5. Additional comments.
   - Questions you would have liked to ask the website project team.
   - Comments on the study.

**At the workshop**

09.00 - 09.20 Brief introduction and presentation.

09.20 - 11.00 Consensus building sessions.

- Split participants randomly into teams of three or four people each. People from the same organization must be in different teams.
- Commission each team to reach consensus on one list of comments on the website. The organizers will propose a number of consensus building methods, such as card sorting and the KJ-method. At the end of this session the team must provide a list of comments on which the team agrees.
- Discussions will be in English.
- Access to the website through the internet will not be provided by the workshop organizers. Workshop participants may bring computers with mobile internet access.
- Teams will be observed by the organizers in order to understand the consensus process; the organizers will not interact with the teams. We will consider videotaping some of the discussions provided that we can get the necessary permissions from the workshop participants.

| | |
|---|---|
| 11.00 - 11.20 | Coffee break |
| 11.20 - 12.00 | Presentation of usability test results.<br>• Are usability test results more reliable than results from expert evaluations? |
| 12.00 - 13.00 | Lunch |
| 13.00 - 13.30 | Comments from the website project team. |
| 13.30 - 15.10 | Plenum discussion:<br>• Problems and positive comments on the website.<br>• Evaluation methodologies.<br>• Were the consensus sessions worthwhile? How could they be improved?<br>• Social aspects of usability evaluation.<br>• CUE-4 process and validity.<br>• Was CUE-4 a worthwhile experience?<br>• How to improve CUE-4. |
| 15.10 - 15.30 | Coffee break |
| 15.30 - 17.00 | Plenum discussion (continued). |

### After the workshop

The organizers will write a paper about the results of the comparative evaluations. Interested workshop members will be invited to contribute to the paper.

## PANELIST DETAILS

*Robin Jeffries* is a Sun Microsystems Distinguished Engineer, where she runs the User Experience Office, part of the Office of the Chief Technologist. She works on user experience issues that impact multiple product groups, ranging from style guides to the design of next generation computer systems to research that improves our understanding of our customers.

She has done product development at Sun for 9 years. Prior to that she was a researcher at Hewlett-Packard Laboratories, at Carnegie-Mellon University, and at the University of Colorado.

Robin was recently the HCI lead for the *Java Look and Feel Design Guidelines: Advanced Topics.*

Robin was an author of one of the earliest papers comparing expert review with other usability methodologies [5]. She has a special interest in false alarms in usability evaluations.

*Rolf Molich* is a senior user experience specialist in the Nielsen Norman Group. Before joining NN/g, Rolf owned and managed DialogDesign, a small and successful Danish usability consultancy (www.dialogdesign.dk). Rolf conceived and coordinated the comparative usability evaluation studies CUE-1 and CUE-2 where four and nine usability labs, respectively, tested the same application. Rolf conceived and planned the CUE-3 study where 12 Danish usability professionals evaluated the same website.

Rolf has worked with usability since 1984; he is the co-inventor of the heuristic inspection method [8] (with Jakob Nielsen), and he is the author of the best-selling Danish book "User friendly computer systems", of which almost 25,000 copies have been sold.

Rolf was a principal investigator in the NN/group's recent large-scale usability test of 20 US e-commerce websites.

Rolf has previously taught several highly rated CHI tutorials and organized several successful CHI panels.

## REFERENCES


1. Cockton, G. and Woolrych, A., Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation, in *People and Computers XV*, eds. A. Blandford, J. Vanderdonckt and P.D. Gray. Springer-Verlag, 171-192, 2001.
2. CUE home page: http://www.dialogdesign.dk/cue.html.
3. Gray, W.D. & Salzman, M. (1998). Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods, *HCI*, 13(3), 203-261.
4. Hertzum, M., Jacobsen, N.E., and Molich, R. Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations., in *Extended Abstracts from CHI 2002* (Minneapolis MI, April 2002), ACM Press, 662-663.
5. Jeffries, R., Miller, J.R, Wharton, C., and Uyeda, K. User interface evaluation in the real world: A comparison of four techniques. *Proceedings of CHI'91*, ACM:New York, pp. 119-124.
6. 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.
7. Molich, R., Kaasgaard, K., Karyukina, B., Schmidt, L., Ede, M., van Oel, W. and Arcuri, M. Comparative Evaluation of Usability Tests, *CHI99 Extended Abstracts*, ACM Press, 83-84.
8. Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces, in *Proceedings of CHI´90* (Seattle WA, April 1990), ACM Press, 249-256.
9. Woolrych, A. and Cockton, G., Testing a Conjecture based on the DR-AR Model of UIM Effectiveness, in *Proceedings of HCI 2002, Volume 2*, eds. H. Sharp, P. Chalk, J. LePeuple and J. Rosbottom, British Computer Society, 30-33, 2002.