BEHAVIOUR & INFORMATION TECHNOLOGY, JANUARY–FEBRUARY 2004, VOL. 23, NO. 1, 65–74

Taylor & Francis
Taylor & Francis Group

Comparative usability evaluation

ROLF MOLICH†, MEGHAN R. EDE‡, KLAUS KAASGAARD§ and BARBARA KARYUKIN¶

†DialogDesign, Skovkrogen 3, DK-3660 Stenlose, Denmark; e-mail: molich@dialogdesign.dk

‡Wells Fargo, San Francisco, CA, USA; e-mail: meghan.ede@acm.org

§Yahoo!, Sunnyvale, CA, USA; e-mail: klausk@yahoo-inc.com

¶Xerox Corp., Wilsonville, OR, USA; e-mail: barbara.karyukin@office.xerox.com

Abstract. This paper reports on a study assessing the consistency of usability testing across organisations. Nine independent organisations evaluated the usability of the same website, Microsoft Hotmail. The results document a wide difference in selection and application of methodology, resources applied, and problems reported. The organizations reported 310 different usability problems. Only two problems were reported by six or more organizations, while 232 problems (75%) were uniquely reported, that is, no two teams reported the same problem. Some of the unique findings were classified as serious. Even the tasks used by most or all teams produced very different results – around 70% of the findings for each of these tasks were unique. Our main conclusion is that our simple assumption that we are all doing the same and getting the same results in a usability test is plainly wrong.

1. Introduction

If a usability test is to be a tool that produces reliable results suitable for making informed design decisions, it is required that two usability tests of the same piece of software, for instance a website, produce reasonably similar results. In particular, it is required that reasonable agreement on the problems that are defined as critical, and on the overall conclusion is reached. Otherwise a project manager who uses a professional usability test report to allocate development resources for correcting critical usability problems may be misled.

To the best of our knowledge there are few other reported comparative studies that have usability tested real software under industrial conditions using professional testers. This is backed by Hertzum and Jacobsen (2001). The only study known is the Comparative Usability Evaluation 1 study (CUE-1). CUE-1 is discussed below. Indeed, even published examples of realistic usability test reports are rare. An informal survey carried out by Rolf Molich revealed that no websites for usability testing companies contained sample reports. Test reports are confidential.

Hertzum and Jacobsen (2001) provide an overview of other comparative studies where the thinking aloud, cognitive walkthrough and heuristic inspection methods were compared. Gray and Salzman (1998) contains a critical study of comparative studies published before 1997, concluding that most of these studies have methodological flaws that make it difficult to conclude much about the reliability of the methods used.

In a study of the evaluator effect reported in Jacobsen *et al.* (1998) four evaluators – all HCI researchers – independently analysed the same set of videotapes of four usability test sessions. Each session involved a user thinking out loud while solving given tasks in a multimedia authoring system. As much as 46% of the problems were uniquely reported by single evaluators and 20% by only two evaluators. Jacobsen *et al.* (1998) conclude that there is a substantial evaluator effect in thinking aloud studies.

Boren and Ramey (2000) visited two organizations that do usability testing and watched how the tests were run. They found a great deal of variability in the way instructions to think aloud are given and in how test administrators interact with test participants during a test session. The study shows that even within the same organization there are inconsistencies in the way testing is implemented.

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0144929032000173951

1.1. *Comparative usability evaluation 1 (CUE-1)*

The Comparative Usability Evaluation 2 (CUE-2) study, which this paper is about, builds on the experience gathered in a previous project, CUE-1. In the CUE-1 project, four professional usability labs carried out independent usability tests of a Windows calendar management application, Task Timer for Windows. The CUE-1 project was conducted in early 1998. The results of CUE-1 were published in Molich *et al*. (1998).

The CUE-1 study showed that there were remarkable differences in approach, reporting, and findings between the labs. The most interesting result was that while a total of 141 usability problems were reported by the four labs, only one problem was reported by all four labs. Another problem was reported by three labs. Eleven problems were reported by two labs. Each of the remaining 128 problems (91%) were uniquely reported by single labs.

Another interesting result was the considerable difference in approach. One team used a quantitative approach to usability testing, focusing mainly on product acceptance in the marketplace. Two teams used a qualitative approach, focusing mainly on usability problems. One team used both approaches.

It was also found that usability reports generated by the labs differed considerably from each other. They also differed somewhat from the recommendations presented in some of the recognized textbooks in the field (Dumas and Redish 1993; Rubin 1994).

2. Comparative usability evaluation 2 (CUE-2)

The CUE-1 study generated considerable interest. At conferences and in discussion groups, colleagues urged us to run a follow-up study to investigate whether the general trends that appeared in CUE-1 could be replicated. They also asked us to correct a number of problems that had shown up in CUE-1, for example to provide a common starting point for the study, to simulate discussions with the development team, and to increase the number of participating teams.

Therefore, nine other professional usability labs decided to undertake another, similar study in late 1998. This study is called CUE-2.

The purpose of CUE-2 was to:

- Continue the CUE research.
- Provide a survey of the state-of-the-art within professional usability testing of websites.
- Provide a basis for discussion about methodological and theoretical foundations of usability testing.
- 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 skills.
- Provide a basis for a panel discussion at CHI99 (Molich *et al*. 1999).

2.1. *Participating teams*

Nine teams participated in the study. The participants were:

Six industry usability labs:

- Framfab, Denmark,
  Lars Schmidt.
- Kommunedata, Denmark,
  Ann Damgaard Thomsen and Klaus Kaasgaard.
- NovaNET Learning, USA,
  Joseph Seeley.
- P5, The Netherlands,
  Wilma van Oel and Roel Kahmann.
- SGI (Silicon Graphics), USA,
  Barbara Karyukin.
- Sun Microsystems, USA,
  Meghan Ede.

One university lab, which sometimes carries out paid usability work:

- University of Maryland, USA,
  Kent Norman.

Two student teams:

- Technical University of Denmark,
  Torben Nørgaard Rasmussen, Asbjørn Johansen, and Tue Nørgaard.
  Faculty advisor: Professor Christian Gram, Dept. of Information Technology.
- Southern Polytechnic State University,
  Marji Schumann, Benjamin Speaks, Nadyne Mielke, Melany Porter, Anusuya Mukherjee, and Michael Quinby.
  Faculty advisor: Dr Carol Barnum, Humanities and Technical Communication.

All teams volunteered for the study. About 20 potential participants contacted Rolf Molich after the publication of CUE-1 and expressed interest in participating in a follow-up study. After hearing the timing requirements for the study, a number of them decided that they did not have the necessary resources.

The two student teams were included in order to get an idea of the difference between professional usability testing and usability testing carried out by inexperienced university students taking a usability class. Both teams attended regular classes in basic usability testing. The teacher of the class attended by the Danish student team was Rolf Molich (one of the CUE-2 organizers).

The teams agreed that each participating organization would cover all of its own expenses in connection with the evaluation.

2.2. *Comparative evaluation plan*

Eight of the usability tests took place in November and December 1998. One test was carried out in the spring of 1999.

Erika Kindlund (a participant in CUE-1) and Rolf Molich selected www.hotmail.com for this study because Hotmail is a widely known state-of-the-art website that requires no particular domain expertise, and because it was one of only two websites that responded favourably to our request for usability testing. Erika Kindlund and Rolf Molich were not part of any of the usability test teams.

At a date that was agreed upon well in advance with each team, the team received the following information by e-mail:

- The name and URL of the website to be tested (Hotmail).
- A client scenario (Molich 2003) describing the main goals of the usability test. The client scenario was provided to simulate a relationship with the development team and to establish a common starting point for all teams. Test goals were identified and prioritized in the scenario. The client scenario was written by Erika Kindlund in co-operation with Hotmail representatives.

The teams were free to use whatever methodology they considered appropriate for the evaluation. However, we asked the teams to carry out the test in a way that was as close as possible to their standards. The usability report was due three weeks after the disclosure of the website.

During the test period the usability teams had a 'marketing liaison' contact to Microsoft Hotmail in the event that they needed further clarification or feedback on the scope of their proposed studies. The access was by e-mail through an intermediary (Erika Kindlund). The intermediary recorded the questions and answers in order to provide a log of the interactions.

Each lab used its standard usability report format with one exception: The identity of the lab was neither directly nor indirectly apparent from the report. In addition, each usability lab reported the following in a separate addendum (Molich 2003):

- Deviations from its standard usability test procedure.
- Resources used for the test (person hours).
- Comments on how realistic the evaluation had been.

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

After all tests had been completed, the anonymous reports were made publicly available on the World Wide Web (Molich 2003).

The user interface of Hotmail did not change significantly during the test period. It has changed significantly since.

3. Results

This section presents observations from the nine usability reports.

All nine teams who agreed to participate in the study completed the study and submitted a report. All reports were delivered either on time or within a few days of the three-week time limit.

3.1. *Selection and application of methodology*

All teams, except team D, chose an approach that centred around the think-aloud method (Dumas and Redish 1993; Rubin 1994), as shown in table 1 and 2. Some of the teams supplemented the think-aloud method with inspection and various inquiry methods.

Team D had their test participants complete a questionnaire after a semi-structured exploration of the product. Part of the questionnaire was a standardized QUIS (Questionnaire for User Interaction Satisfaction) (QUIS 2003) rating of the product. No observational data were collected.

As shown in table 3, six of the nine teams chose to recruit both experienced and novice Hotmail users. Although this was not explicitly mentioned in the scenario, we think this is a reasonable decision because the usability of Hotmail for experienced users is as important as the usability for novice users.

Table 1. Methods employed.

Team	A	B	C	D	E	F	G	H	J
Number of test participants	7*	6	5	50	9*	5	11	4	6
Recording technique	Video	Video	Paper	Paper	Video	Paper	Video	Paper	Video
Testing (think aloud)	+	+	+		+	+	+	+	+
Inspection		+	+		+				
Inquiry: Interview	pr/af		af			af	pr/af		pr/af
Inquiry: Questionnaire		af		af	pr/af		af	pr	af
Inquiry: Standard tools				QUIS			SUS		

* = including one pilot test whose results are included in the data reported in this paper; pr = pre-test, af = post-test; QUIS = Questionnaire for User Interaction Satisfaction (QUIS 2003); SUS = System Usability Scale (Brooke 1996).

Table 2. Measurement approaches.

Team	A	B	C	D	E	F	G	H	J
Subjective data	+	+	+	+	+	+	+	+	+
Objective data: Time per task		+				+	+		+
Objective data: Task success	+	+	+		(+)	+	+	+	+
Objective data: Number of errors	+	+	+		+	+	+	+	+

" + " indicates that the team used this approach.

Table 3. Testing experienced and novice users.

Team	A	B	C	D	E	F	G	H	J
# experienced/novice Hotmail users	3/4	3/3	3/2	?	5/4	0/5	8/3	0/4	3/3
Tasks used to test experienced users	same	same	?	same	diff		reorder		diff

Same = Identical tasks for experienced and novice users completed in the same order. Reorder = Identical tasks completed in a different order. Diff = Different tasks for experienced and novice users.

3.2. *Interaction with the development team*

During the test period, the usability teams had access to a development team representative by e-mail. Two teams used this option.

Team F requested a 30 min phone conversation with a developer representative in order to clarify the scope and purpose of the test. When informed that questions could be submitted in writing, the team exchanged several e-mails with the intermediary. After the test had been completed, team F noted that '... all answers were the same. This is why I had no further questions. To me it seems that there was no real conversation'.

Team J asked 12 questions, mainly about user demographics and usability feedback from the Hotmail hotline. These questions appeared relevant.

Seven other teams did not avail themselves of the contact methods provided. It is unclear whether this was standard practice or due to the artificial quality of the study design where there would not be direct or on-going contact.

3.3. *Selection of test tasks*

The scenario given to the teams identified and prioritized 18 features that Hotmail Marketing and Engineering had identified as benefiting from user feedback. The five top priority features were registration, login, logout, viewing Hotmail with or without frames, and customization.

As table 4 shows, the overlap between the task sets chosen by the teams was limited. Teams arrived at different task sets regardless of the baseline. Almost half of the tasks used for testing Hotmail were unique, that is, they were used by only single teams. Only six tasks out of 51 were used by five or more teams.

One team did not report tasks used for the test. Although we were able to deduce some of the tasks that this team had used from their findings, this team is not included in the summary in table 4.

The following list shows examples of how many teams tested various common tasks:

Table 4. Overlap in tasks used for testing Hotmail.

Total number of tasks	51	100%
Nine teams	?*	?*
Eight teams	2	4%
Seven teams	1	2%
Six teams	0	0%
Five teams	3	6%
Four teams	8	16%
Three teams	5	10%
Two teams	7	14%
Single teams, no overlap	25	49%

The table shows for example that two tasks were included in the task sets for eight teams, and that one task was used by seven teams. *One team did not report tasks used for the test.

Common tasks tested by most teams	# Teams
• Register and create a new account for yourself	All
• Send mail to one person	8
• Log out from Hotmail, or 'Leave pc for a while'	7

Common tasks tested by several teams	# Teams
• Login to Hotmail (after registering)	4
• Send mail with attachment	5
• Open attachment	4

Common tasks tested by a few teams	# Teams
• Forward simple mail	0
• Reply to mail	1
• Forward mail with attachment	1

Two teams used task sets tailored for experienced users (see table 3). Other teams gave experienced users the same tasks as inexperienced users.

3.4. *Test task flaws*

Some of the tasks used by the teams contained instructional bias like hidden clues. For example, one of the teams used the task 'Create a personal signature'. This task description contains a hidden clue: 'Signature' is a term used by Hotmail.

Tasks that contain hidden clues may test the test participant's ability to recognize a keyword rather than his/her ability to understand how the task is carried out.

Team H was the only team that used a task set without hidden clues (8 tasks). Team F did almost as well: Their task set contained 25 tasks with only one hidden clue. Almost all teams used the terms 'attachment' and 'POP mail' in their task descriptions. These Hotmail terms are hidden clues.

Once you are aware of the problem with hidden clues, it is simple to express tasks in a way that does not contain clues. Compare the following task with clues: 'Lois McClaran uses Hotmail. She lives in Indiana. Look her up in the Hotmail Membership Directory, and send her mail' to a similar significantly better scenario used by team C without clues: 'Lois McClaran uses Hotmail. She lives in Indiana. Send her mail'.

3.5. *Analysis*

Rolf Molich went through all of the nine test reports in order to determine the overlap in the usability problems reported by the teams.

Four of the nine teams compared their findings to the overall result list created by Rolf Molich. The remaining five teams did not have the time to do so. The comparison revealed 20 problems for which a team had not received proper credit. The four teams also detected two pairs of problems that expressed the same basic problem with different wordings; these problems were combined.

The comparison did not affect our main conclusions. Note that the comparison increased the total number of problems found in the study. The figures reported in this paper are based on the results after the comparison.

3.6. *Problems reported*

The main results are shown in table 5. The spreadsheet with the complete results of the analysis is available on the CUE home page (Molich 2003).

The overlap was remarkably limited. There was not a single problem reported by all nine teams or by eight teams. Only one problem was reported by seven teams.

Most surprising to us was that the difference in findings was significant even among the tasks that were used by all or almost all teams. Here are the results for two tasks tested by all teams:

- Registration – create a new account (see table 5)
  - One problem was reported by seven teams.
  - Out of 48 reported problems, 34 (71%) were uniquely reported by single teams.
- Compose an e-mail message and send it.
  - Two problems were reported by five teams.
  - Out of 23 reported problems, 16 (70%) were uniquely reported.

Interestingly, only 25% of all problems were reported by two or more teams; the remaining 75% were uniquely reported. As shown in table 6, 29 of the 232 uniquely reported problems were classified as 'serious'.

Less than 10 of the 310 reported problems turned out to be non-reproducible or incomprehensible.

3.7. *Content and format of usability test reports*

Since one of the goals of our study was to gather insight into the everyday practices of usability professionals, we also examined the content and format of the usability test reports. The following violations of good reporting practices were observed as outlined for example in the standard textbook *A Practical Guide to Usability Testing* (Dumas and Redish 1993):

- Report too long.
  Too many problems reported. A usability report that describes 75 or even 150 usability problems is difficult to read and sell to developers and designers. If no other agreement has been made with the customer, only a manageable number of problems should be reported, perhaps 15–60. It is an important task for a usability professional to prioritize the full list of usability problems so that only the most important ones are reported.
- No executive summary.
- No severity classification of problems.
  Some reports did not distinguish between serious problems and minor details.
- No indication of how many users encountered a problem (frequency).
- No positive findings.
  One report started by saying 'Generally, the users were very happy about Hotmail'. The rest of the report contained more than 30 problem descriptions without any positive findings to substantiate the initial claim.
- Unattractive, unprofessional layout.
  The layout is important for selling the results to busy developers.
- Unclear or vague problem descriptions that required time consuming deciphering or even clarification from the test team.

Table 5. Overlap in problems reported for all tasks and for Registration.

	All tasks		Registration	
Total number of problems	310	100%	48	100%
Nine teams	0	0%	0	0%
Eight teams	0	0%	0	0%
Seven teams	1	0.3%	1	2%
Six teams	1	0.3%	0	0%
Five teams	4	1%	0	0%
Four teams	5	2%	2	4%
Three teams	17	5%	4	8%
Two teams	50	16%	7	15%
Single team, no overlap	232	75%	34	71%

Columns 2 and 3 show the overlap in findings for all tasks. Columns 4 and 5 show the overlap in findings for the registration task, which all teams used (see the section 'Discussion – Differences in Tasks').

Table 6. Important characteristics of usability test reports.

Team	A	B	C	D	E	F	G	H	J
# Pages	16	36	10	5	36	19	18	11	22
Executive Summary?	Yes	Yes	No	No	No	Yes	No	Yes	Yes
# Screen shots in report	10	0	8	0	1	2	1	2	0
# Levels in severity scale	2	2	3	1	2	1	1	3	4
# Problems reported	32	149	18	10	67	76	41	18	25
# Problems reported only by this team	13	105	4	0	33	36	19	10	12
# Serious problems reported only by this team	5	16	2	–	3	–	–	1	2
# Positive findings reported	0	8	4	7	24	25	22	4	6

Table 7. Summary of resources used to test www.hotmail.com.

Team	A	B	C	D	E	F	G	H	J	Mean
Person hours used for test	136	123	84	16	130	50	107	45	218	112
# Usability professionals involved	2	7	1	1	3	1	1	3	6	3.0
Number of test participants	7	6	5	50	9	5	11	4	6	6.6

Team D is not included in the Mean.

3.8. *Resources used*

The number of person hours used by each team are shown in table 7. The figures do not include the time spent by test participants. The figures range from 45 hours to 218 hours. There seems to be no simple relation between the number of hours used, the length of the report and the number of problems reported. The figures for team D are not directly comparable to the other figures; the 16 hours were used to write a briefing to the test participants, accumulate the results and write the report.

4. Comments from Hotmail

After the tests had been completed, the Hotmail usability team reported:

- New findings ~4%.
- Validation of known issues ~67%.
  - Previous finding from Hotmail lab tests.
  - Finding from on-going inspections.
- Remainder – beyond Hotmail Usability ~29%.
  - Business reasons for not changing.
  - Out of Hotmail's control (partner sites).
  - Problems generic to the web.

Hotmail did not go into details about which 4% of our problems were new findings because usability findings are considered confidential. However, Hotmail has told us that the one problem detected by seven teams was a new finding. Hotmail usability engineers also indicated that the limited usability test resources are normally focused on new features to be released (such as localization to German, French and Japanese) rather than old features.

5. Discussion

5.1. *Methodological approaches*

All teams, except team D, used the same methodology. However, the implementations of the methodology may have varied significantly.

All teams used their standard usability test procedure as requested. The basic methodology was the same. Yet there was very little overlap in the problems identified. Our study doesn't show that the same procedure and the same tasks result in the detection of different problem sets; it shows that lots of different procedures lead to the discovery and reporting of lots of different problems.

The applied methodologies seemed to correspond well to established practice in the area (Dumas and Redish 1993; Rubin 1994). We saw a few possible mistakes in the use of the methodology. We do not believe, however, that these mistakes influenced our main results considerably. The possible mistakes were:

- It is not apparent from the test reports that the teams interviewed experienced users about their actual use of Hotmail, asking them to show for instance what parts of Hotmail they used daily, sometimes and never.
- None of the teams reported that they had considered a competitive analysis, even though the scenario said 'Hotmail's biggest competitors are: Yahoo Mail and Netscape WebMail'. No one mentioned the possibility of comparing Hotmail to Outlook Express.

5.2. *Interaction with the development team*

After the CUE-1 study several of the participating teams said that their study would have been more focused if they had had access to the development teams. We therefore offered such access through an intermediary in CUE-2.

Two of the teams (F and J) took advantage of this offer. The remaining seven teams did not request information from customer marketing or development beyond what was in the scenario. Five teams (A, B, C, E and G) noted after the test that the CUE-2 test deviated significantly from their standard procedure in that they usually work quite closely with the client in determining features that might benefit from user feedback, etc. There is no explanation why these teams did not attempt to establish contact with the client.

It is possible that the test would have been more tightly focused had it gone through the normal process of negotiation for services. Results might have been more consistent had fewer areas been subjected to testing, but no team attempted to raise this question towards the client although several noted it after the test.

Interaction with the client apparently is not a must. In our experience most professionals are willing to carry out evaluations of websites belonging to the competitor of a client, where there is no access to the web team.

5.3. *Competence*

A critic of our study has argued '... if you take 11 mediocre (at least, unproven) usability teams [CUE-1 and CUE-2 excluding student teams] and you give them

a hard problem, do you find out something interesting?' (Anonymous CHI 2001 reviewer).

The CUE-2 teams represented the state-of-the-art in the usability testing field. They covered a wide range of attributes, including in-house teams, contractor/vendor teams, and different locations in the USA and in Europe. Most of these teams were well-established and had been operating for many years. If these teams are 'mediocre' or 'unproven' then we have a problem as a profession.

The two student test reports are not easily distinguishable from the professional reports. Several people have unsuccessfully tried to identify which two of the nine anonymous test reports were written by the student teams. Most have not had even one report right. At a CHI2000 tutorial, 70 attending usability professionals were unable to distinguish between the professional and the student reports.

Some people have argued that usability professionals and their methods should ideally be measured on how effectively they introduce usability improvements into the product. We agree. We also agree that our study has not measured (and has not attempted to measure) this critical success factor, except for some observations on the professionalism of the content and format of the usability test reports.

5.4. *Differences in tasks*

The scenario asked the teams to test 18 major features in Hotmail. This is impossible within the time limits of an ordinary usability test, thus the teams had to pick and choose what features to test. As we have discussed above, results might have been more consistent had fewer areas been subjected to testing. However, it is not at all unrealistic for a usability team to receive a request for testing 18 features all at once. The large number of tasks may have affected the number of findings, however not as significantly as one may think. Note that the difference in findings stands out even among the tasks used by all or most teams. See table 5.

Instead of testing high priority, general tasks (as per the scenario), some teams included tasks in their task set that were lower in Hotmail's priority. They did so without any explanation as to why they determined that these tasks were more important. One example is: 'It's your mother's birthday December 14. You want Hotmail to remind you a day in advance so you can remember to buy a present'. Although the reminder function was included in the feature list provided by Hotmail, teams overlooked more common, higher prioritized features that pertain to the core functionality of the website, such as sending attachments.

There were few reported attempts to select tasks based on what real users considered important in Hotmail. It seems that most of the task sets were based on the usability teams' perception of what was important. Three of the nine teams (A, B, and J) gave some kind of a rationale for their selection of tasks. Team B carefully mapped their tasks onto the requirements from Hotmail.

5.5. *Huge range of content*

The limited overlap between findings may be a result of the huge range of content on commercial sites. In other words, current websites may be so huge and contain so many usability problems that within the limited scope of a single usability test one can hope to identify only a fraction of even the serious usability problems.

This viewpoint is supported by Spool and Schroeder (2001), who conducted 49 usability tests of four e-commerce websites and identified 378 problems that prevented people from completing their purchase.

Our teams reported 436 usability problems in Hotmail. After combining duplicates (same usability problem reported by several teams), 310 different usability problems remained. As shown in table 5, 75% of these problems were reported uniquely, by single teams. At least 29 serious problems were reported by single teams. The exact number of serious problems reported by single teams is difficult to determine because three teams did not assign a severity to their reported problems.

The limited overlap suggests that if we had continued to test with more teams, we would have discovered many more problems. There doesn't appear to be a threshold; as we add more teams, we not only get more results, but also more significant, serious and unique results.

None of our teams commented that the usability of Hotmail was any worse (or any better) than the websites they normally tested.

5.6. *Are five users enough?*

Nielsen (2000) concludes that a usability study with five users will find 85% of the usability problems. Nielsen (2000) bases his conclusion on the mathematical model in Nielsen and Landauer (1993). Virzi (1992) also looks at the issues of 'how many subjects are enough' and arrives at a similar conclusion.

The CUE-2 study puts a different spin on these conclusions. In our study, nine teams reported 310 usability problems in Hotmail, of which 85% would

represent 264 problems. None of the teams reported this many problems even though all but one team used five or more test participants (Team H used four). The range in reported problems was a low of 10 (3% of all reported problems) to a high of 149 (48%), with the remaining seven teams reporting between 18 (6%) to 76 problems (25%).

There didn't seem to be any relation between the number of test participants and the number or type of reported problems. The team using the most participants (Team D, 50 participants) reported the fewest problems (10, none uniquely reported) and the team using the least participants (Team H, four participants) reported the next fewest problems (18, of which 10 were uniquely reported, including one serious problem not reported elsewhere). That is, even using only four test participants resulted in uniquely reported, serious problems. The most problems were reported by Team B (149, of which 105 were uniquely reported) who used six test participants, which was about average for the teams (modified mean, excluding Team D = 6.6 test participants). Excluding the results for Team D, which did not use the think aloud protocol and found no problems that other teams hadn't also found, all the remaining teams reported interesting and useful results but none came close to discovering 'all' problems or reaching the 85% threshold.

It appears that five test participants are enough to find worthy problems, but that one test is, by itself, unlikely to find anywhere near 85% of the problems regardless of the number of test participants. It may very well be true, however, that for a given test team and a given set of tasks, testing more than five users will not reveal a significant number of new problems.

It also appears that the complexity of current state-of-the-art websites like Hotmail is much larger than the complexity of the systems used to derive the Nielsen and Landauer model. In fact, the complexity is so large that it cannot be covered by one test no matter how many test participants it uses. Only iterative testing of all areas of the site would come close to identifying most of the problems.

The limited overlap between the findings suggests that even the combined effort of our nine teams has not found 85% of the problems.

5.7. *Use of a trained evaluator versus unattended testing*

Team D asked their 50 test participants to do an unattended exploration of Hotmail. When comparing the main results from team D to that of the other teams (table 6) we note that this team has reported far fewer problems than any other team. Also, team D was one of the two teams that did not report the one serious problem reported by seven other teams – the area of largest overlap among the teams.

Unattended testing didn't lead to any more (in fact, quite a bit less) reported problems and didn't provide insights that other methods didn't also provide. Since only one team used this method, however, further studies are needed before the reliability of the unattended test method can be seriously understood.

6. Conclusions

6.1. *Consistency of usability testing across organizations*

Our results document a wide difference in reported problems, methods, tasks and usability reports from nine teams who usability tested the website www.hotmail.com. Each usability team except one (Team D) produced unique results which the team classified as important. Seventy-five per cent of the 310 usability problems were only reported by single teams. At least 29 serious problems were reported by single teams.

We have demonstrated that the effectiveness of a usability test is dependent on the chosen tasks, the methodology, and the persons in charge of the test. All usability tests and testers are not equal – even amongst professional organisations. Our results are contrary to the common belief.

After our study was completed, Kessner *et al*. (2001) conducted a similar study with comparable results.

6.2. *Large number of usability problems*

Assuming that Hotmail indeed represents the state-of-the-art within website usability, we can conclude that the number of usability problems in a typical website may be so large that one cannot hope to find more than a fraction of the problems in an ordinary usability test.

6.3. *Everyday practices of usability professionals*

Variations in the everyday practices of usability testing and their deviation from the textbook recommendations may have contributed to the limited overlap in the findings. Our results document a wide difference in:

- selection and application of methodology
- selection of test tasks
- formulation of test tasks

- problems reported
- content and format of usability test reports.

6.4. *Recommendations*

We offer the following recommendations for development teams, usability professionals and their managers:

- Realize that there is no foolproof way to identify usability flaws. Usability testing by itself can't develop a comprehensive list of defects. Use an appropriate mix of methods.
- Place less focus on finding 'all' problems. Realize that the number of usability problems is much larger than you can hope to find in one or a few tests. Choose smaller set of features to test iteratively and concentrate on the most important features.
- Realize that single tests are not comprehensive. They are still useful, however, and problems detected in a professionally conducted single test should be corrected.
- Increase focus on quality and quality assurance. Prevent methodological mistakes in usability testing such as skipping high-priority features, giving hidden clues or writing usability test reports that are not fully usable.

Acknowledgements

Thanks to Meeta Arcuri and Rob Aseron of MSN Hotmail for allowing us to use the Hotmail website for the test. Thanks also to Nigel Bevan of Serco Usability Services, Erika Kindlund of Intraspect Software (now working for Intuit), Anker Helms Jørgensen of the IT University of Copenhagen, Joseph S. Dumas of Oracle Corp., and an anonymous CHI2001 reviewer for insightful comments on early drafts of this paper.

References

Boren, M. T. and Ramey, J., 2000, Thinking aloud: Reconciling theory and practice. *IEEE Trans. Professional Communication*, **43,** 261–278.

Brooke, J., 1996, SUS – A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland (eds) *Usability evaluation in industry* , (London: Taylor & Francis), pp. 189–194.

Dumas, J. S. and Redish, J. C., 1993, *A Practical Guide to Usability Testing* (Norwood: Ablex).

Gray, W. D. and Salzman, M. C., 1998, Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, **13,** 203–261.

Hertzum, M. and Jacobsen, N. E., 2001, The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, **13,** 421–443.

Jacobsen, N. E., Hertzum, M. and John, B. E., 1998, The evaluator effect in usability studies: Problem detection and severity judgements. Human Factors and Ergonomics Society 42nd Annual Meeting, 5–9 October 1998 (Santa Monica: Human Factors and Ergonomics Society), pp. 1336–1340.

Kessner, M., Wood, J., Dillion, R. F. and West, R. L., 2001, On the reliability of usability testing. Conference on Human Factors in Computing Systems: CHI 2001, 31 March–5 April 2001 (extended abstracts) (Seattle: ACM Press), pp. 97–98.

Molich, R., 2003, Comparative usability evaluation – CUE. http://www.dialogdesign.dk/cue.html.

Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J. and Miller, D., 1998, Comparative evaluation of usability tests. Usability Professionals Association 1998 Conference, 22–26 June 1998 (Washington DC: Usability Professionals Association), pp. 189–200.

Molich, R., Kaasgaard, K., Karyukina, B., Schmidt, L., Ede, M., van Oel, W. and Arcuri, M., 1999, Comparative evaluation of usability tests. Conference on Human Factors in Computing Systems: CHI99, 15–20 May 1999 (extended abstracts) (Pittsburgh: ACM Press), pp. 83–84.

Nielsen, J. and Landauer, T. K., 1993, A mathematical model of the finding of usability problems. Conference on Human Factors in Computing Systems: INTERCHI '93, 24–29 April 1993 (New York: ACM Press), pp. 206–213.

Nielsen, J., 2000, Why you only need to test with 5 users. http://www.useit.com/alertbox/20000319.html.

QUIS, 2003, About the QUIS, version 7.0. http://www.lap.umd.edu/quis/

Rubin, J., 1994, *Handbook of Usability Testing* (New York: John Wiley).

Spool, J. and Schroeder, W., 2001, Testing web sites: Five users is nowhere near enough. Conference on Human Factors in Computing Systems: CHI 2001, 31 March–5 April 2001 (extended abstracts) (Seattle: ACM Press), pp. 285–286.

Virzi, R. A., 1992, Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, **34,** 457–468.