Participating in CUE-8, Comparative Usability Evaluation

Anders Bruun, Janne Jul Jensen, Mikael Skov & Jan Stage
Department of Computer science
Aalborg University
Selma Lagerlöfs Vej 300
DK-9220 Aalborg East
{ bruun, jjj, dubois, jans}@cs.aau.dk

ABSTRACT
This paper reports on the usability evaluation conducted for the participation in the eighth Comparative Usability Evaluation (CUE-8). It elaborates on the history of the CUE series, then reports in detail on the usability evaluation conducted and the results of it. Finally the overall results of the CUE-8 workshop are explored and the lessons learned from the workshop are presented.

Keywords
Usability measurement, comparative usability evaluation, time-on-task, satisfaction rating, success rate, quantitative data analysis, SUS, TLX

INTRODUCTION
Traditional usability evaluations are a series of moderated sessions involving a user and a test leader, and it generates both quantitative and qualitative data. This type of qualitative test is the most common usability test. However, usability practitioners find themselves having to accommodate managers who prefer measurements over qualitative data [2] in order to be able to benchmark and measure progress. The ISO 9241-11 [1] standard defines usability in terms of effectiveness, efficiency, and satisfaction and provides examples of metrics to measure them. Most commonly used by practitioners are success rate, time-on-task, satisfaction rating, and error rate. This became the focus of the eighth CUE workshop in which 15 teams participated. The website chosen for evaluation was budget.com, a car rental service website.

BACKGROUND
For eight years running, Rolf Molich has organised a reoccurring workshop on comparative usability evaluation, often referred to as the CUE-workshops [3]. Each year these workshops include a number of professional usability teams that volunteer their skills to evaluating a chosen product or service applying the methods, tools, techniques and procedures they would normally use for a similar evaluation. This generates a large amount of empirical data that is otherwise rarely available creating an ideal basis for comparison of results.

In 2009 the eighth CUE workshop took place at the Usability Professionals' Association (UPA) Conference in Portland, OR, USA on June 9th 2009. Molich had found that there is no general agreement on what best practice in usability task measurement is [4]. Thus, the purpose of this year's CUE workshop was to discuss the state-of-the-art in usability task measurement based on the results gathered from each teams' evaluation of a particular website and compare practical approaches to usability task measurement based on the assumption that "you can't manage what you can't measure". This differed from previous years' CUE workshops as they have been focusing on qualitative evaluations, rather than quantitative.

Ahead of the workshop each of the 15 participating teams were asked to conduct a usability evaluation of the car rental service website budget.com. The tasks were fixed and the same for all teams although comments and changes were possible on drafts beforehand. Each team was asked to carry out an independent evaluation parallel with the other teams, using the methods, tools, techniques and procedures they would normally use for a similar evaluation. Each team was, however, encouraged to measure efficiency (e.g. time-on-task) effectiveness (e.g. completion rate and errors) and satisfaction (e.g. post-task and post-test ratings). The System Usability Scale (SUS) was suggested as a post-test questionnaire if a team was unfamiliar with measuring post task and post test satisfaction. Each team was also expected to be willing to spend 10-30 hours on the evaluation and preparation of a report before the workshop.

Upon completing the evaluation, each team was asked to produce an anonymised usability report containing their results to the organizers ahead of the workshop. These reports would then form the basis of the workshop.

OUR CONTRIBUTION TO CUE-8

Our evaluation of budget.com took place on May 19th 2009 in our usability lab. It involved 10 users, two test leaders and two loggers.

Procedure

The usability evaluation was carried out by the authors of this paper.

Picture 2: The evaluation setup seen from within the observation room.

The participants were assigned a 45 minute slot each in a test plan and two of the authors were assigned as alternating test leaders, while the two others would operate the data collection equipment. The participants were asked to think aloud to supply an insight into their train of thought during their task solving. Upon completing their task solving, each of the participants were subjected to a NASA TLX test to measure their mental workload during the evaluation.

	Gender	Age	Internet exp.	Budget.com exp.	Renting cars online exp.
TP1	F	41	Every day	Never	1 time
TP2	F	45	Every day	Never	10+ times
TP3	F	35	Every day	Never	0 times
TP4	F	34	Every day	Never	2-10 times
TP5	F	28	Every day	Never	0 times
TP6	M	30	Every day	Never	0 times
TP7	M	28	Every day	Never	2-10 times
TP8	M	27	Every day	Never	1 time
TP9	M	26	Every day	Never	2-10 times
TP10	M	23	Every day	Never	0 times
Avg.	--	31.7	Every day	Never	--
High	--	45	Every day	Never	10+ times
Low	--	23	Every day	Never	0 times

Table 1: Demographic data of the participants.

Participants

The evaluation included ten participants. All participants were employees in our organization or spouses of the evaluators. As the website should appeal to a wide demographic profile we chose participants of differing age, differing job profile and an even number of males and females. Each participant was given a bottle of wine for their participation. Their demographic data can be seen in table 1.

The Evaluation

All ten evaluation sessions were carried out in the usability laboratory of our organization (See figure 1).

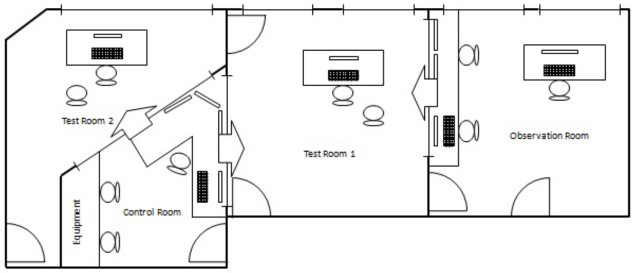

Figure 1: The layout of the laboratory used.

After greeting and briefing each participant, they were placed by the PC and given the tasks one by one always in the same order. They were asked to clearly state when they felt they had completed the task. The test leader would only help in case the participant got stuck.

After each task, the participant was asked to answer the corresponding question in a SUS-questionnaire. The evaluation was stopped if this exceeded the assigned 45 minutes by more than five minutes (happened once). Upon completion of the evaluation, the participants answered the rest of the SUS questionnaire.

Each session was completed by having the participant fill out a questionnaire regarding their demographic data.

RESULTS OF OUR EVALUATION

Our results address the three categories of the ISO: Efficiency, effectiveness and satisfaction. We furthermore also compiled a problem list and the results of the NASA TLX test, the results of which will not be presented in this paper.

Efficiency (Time)

The participants were rather diverse in terms of efficiency. On average, they spent almost 1600 seconds (~26 minutes) on task completion. However, they were rather different on task completion with one participant using only 997 seconds (16 minutes) and another using 2719 seconds (~45 minutes). See table 2 for further details.

Our results seem to challenge the statement on the front page of budget.com where it is claimed that you can rent a car in 60 seconds. All our participants spent more than 4 minutes on this task.

Task 2 had a relatively high task completion time for a task that to some degree was a repetition of task 1. It could be expected that the completion time would reflect some learning from task 1 but this seems minimal.

Participant	1	2	3	4	5	Total time
TP1	277	137	160	352	*540*	1466
TP2	328	158	141	160	*210*	997
TP3	395	236	822	268	198	1919
TP4	279	176	161	307	165	1088
TP5	243	317	391	208	220	1379
TP6	585	486	419	318	283	2091
TP7	615	437	154	404	275	1885
TP8	*1012*	496	302	909	--	2719
TP9	356	176	105	145	304	1086
TP10	*417*	231	230	171	253	1302
Avg.	450.7	285	288.5	324.2	272	1593.2
High	1012	496	822	909	540	2719
Low	243	137	105	145	165	997

Table 2: Task completion time for the participants. Gray italic numbers indicate that the task was not solved or that the test leader provided extensive help, while two dashes indicates that the participant was asked to proceed by the test leader.

Effectiveness (Task Completion)

We measured effectiveness from task completion. Four participants never fully completed two tasks (task 1 and 5). Either they realized they could not complete the task, e.g. find specific information, or they simply failed to provide a correct answer to the question specified in the task.

Our strong focus on usability problem identification (as the primary result of our evaluation) results in very few non-completed tasks: Encouraging the participants to continue trying to solve the tasks usually provides more insight into the problems of the application. However, this also often means that the participants manage to finish tasks they would otherwise not have finished, working on their own, as they would simply have given up earlier.

Satisfaction (System Usability Scale, SUS)

All participants filled in a System Usability Scale (SUS) questionnaire as a measure for satisfaction.

SUS scored (1-100)			
TP1	63	**TP6**	48
TP2	90	**TP7**	63
TP3	73	**TP8**	32
TP4	58	**TP9**	70
TP5	33	**TP10**	53
Avg.	59		
High	90		
Low	32		

Table 3: The SUS scored on a scale from 1 to 100.

Looking at the SUS questions after each task (table 4), we can see that the participants perceived the first task as relatively easy (2.8). This is somewhat surprising as they spent considerable more time on this task than anticipated. On the other hand, they were more negative towards task 4 (3.8) where they have to find information about insurance. This task caused several problems for more of the participants.

	Task 1	Task 2	Task 3	Task 4	Task 5
TP1	1	2	3	5	7
TP2	5	1	2	3	2
TP3	2	2	7	1	2
TP4	3	2	2	4	1
TP5	1	1	6	3	2
TP6	4	4	4	4	4
TP7	2	5	1	6	4
TP8	6	4	2	4	--
TP9	2	3	2	3	5
TP10	2	2	3	5	2
Avg.	2.8	2.6	3.2	3.8	3.2
High	6	5	7	6	7
Low	1	1	1	1	1

Table 4: The SUS rating of each task from each participant on a scale from 1 to 7, 1 being easiest, 7 being hardest.

THE CUE-8 WORKSHOP

The results of our evaluation were compiled into a usability report which was submitted to the organizers of the CUE-8 workshop. The organizers had before the workshop produced some comparison results derived across all of the reports submitted and these results were presented at the workshop. Each team would present their results and based on the reports and the presentations, an extensive discussion of the results took place.

Based on the workshop the following overall lessons learned were derived:

Lesson 1: Unmoderated usability evaluations are only more cost effective than moderated usability evaluations when the sample size is large: Surprisingly unmoderated evaluations proved to have a lot of overhead compared to moderated evaluations with small sample sizes. This was attributed to the extra work of cleaning up the data of the evaluation.

Lesson 2: It is advisable to use recognized questionnaires rather than to make your own: Own brand questionnaires tend to be less regular and may not discriminate between the tremendous variety there is between users, thus causing warped when doing the statistical analysis afterwards.

Lesson 3: Cleaning contaminated data from unmoderated usability evaluations poses serious challenges: The data of unmoderated evaluations often contain flawed data in the form of unrealistically high or low time-on-task or low error rate. This is usually dealt with through a cleaning procedure setting some thresholds. However, multiple teams found that with these procedures there were outliers being discarded that were valid and inliers that were erroneous and should have been discarded but were not. Thus, unmoderated evaluations come at a cost.

Lesson 4: Using mean and median for time-on-task should be done carefully: Often mean and median are used for reporting the average time on task in a usability evaluation. However, as time-on-task is not normally distributed, the mean is a poor indicator of the centre of a distribution. The median may be used instead but it censors data or discards extreme observations instead. An uneven distribution can be handled with the right statistical tools, but unfortunately it rarely is.

Lesson 5: Confidence intervals are valuable for describing the location and precision of the results: Often however, these are not computed and reported. This could be a valuable addition to a field that mostly takes a qualitative approach to usability evaluation. Eight of the 15 teams did not report confidence intervals for their data.

Lesson 6: Reproducing results between teams is possible to some extent: Six of the 15 teams agreed on all five tasks within a 95% confidence interval. Two more teams agree with the six teams for all tasks except task 1. Two teams agree with the majority for three tasks. On the other hand, five teams mostly report diverging results. Two teams consistently diverge from the other teams.

CONCLUSION

Usability metrics expose the weaknesses in usability evaluation methods (recruiting, task definitions, user-interactions, task success criteria, etc) that likely exist with qualitative testing but are less noticeable in the final results. With qualitative data it is difficult to compute the reproducibility of the results due to their qualitative nature. This in return prevents us from assigning confidence intervals, which can be a valuable metric.

Unmoderated measurements are attractive from a resource point of view with large sample sizes; however, data contamination is a serious problem and it's not always clear what you are actually measuring. Furthermore cleaning the data poses a number of challenges not trivially overcome. We recommend further studies of how data contamination can be prevented and how contaminated data can be cleaned efficiently.

ACKNOWLEDGMENTS

We wish to thank Rolf Molich for the possibility of participating in the CUE-8 workshop. We also wish to thank the participants of our usability evaluation, as it would not have been possible without them.

REFERENCES

1. ISO (1998) ISO Standard 9241-11. Guidance on Usability, International Organization for Standardization.
2. Molich, R., Kirakowski, J., Sauro, J. & Tullis, T. (2009) Comparative Usability Task Measurement (CUE-8) Instructions. Retrieved on December 1, 2009 from http://www.dialogdesign.dk/cue-8.htm.
3. Molich, Rolf (2009) CUE - Comparative Usability Evaluation. Retrieved on December 1, 2009 from http://www.dialogdesign.dk/CUE.html
4. Usability Professionals' Association (2009) Comparative Usability Task Measurement (CUE-8). Retrieved on December 1, 2009 from https://www.usabilityprofessionals.org/upa_conference/app/schedule/show_detail/10176/for:2009