JUS
JOURNAL OF USABILITY STUDIES
Vol. 6, Issue 1, November 2010, pp. 8-24

Rent a Car in Just 0, 60, 240 or 1,217 Seconds? – Comparative Usability Measurement, CUE-8

Rolf Molich
DialogDesign
Skovkrogen 3
DK-3660 Stenlose
Denmark
molich@dialogdesign.dk

Jarinee Chattratichart, Ph.D.
Senior Lecturer
Faculty of Computing, Information Systems and Mathematics
Kingston University
Surrey KT1 2EE
United Kingdom
j.chattratichart@kingston.ac.uk

Veronica Hinkle
HCI Researcher
Software Usability Research Laboratory (SURL)
Dept. of Psychology
Wichita State University
Wichita, KS

Janne Jul Jensen, Ph.D. Postdoc
Aalborg University
Dept. of Computer Science
Selma Lagerlöfs Vej 300
DK-9220 Aalborg East
Denmark
jjj@cs.aau.dk

Jurek Kirakowski, Ph.D.
Director, Human Factors Research Group,
Enterprise Centre, North Mall
University College Cork,
Ireland.
jzk@ucc.ie

Jeff Sauro
Principal Usability Engineer
Oracle Corporation
1 Technology Way
Denver CO, 80237
jeff@MeasuringUsability.com

Tomer Sharon
User Experience Researcher
Google
76 Ninth Avenue
New York, NY 10011, USA
tsharon@google.com

Brian Traynor
Associate Professor, Faculty of Communications Studies,
Mount Royal University,
Calgary, Alberta, Canada
btraynor@mtroyal.ca

Abstract

This paper reports on the approach and results of CUE-8, the eighth in a series of Comparative Usability Evaluation studies. Fifteen experienced professional usability teams simultaneously and independently measured a baseline for the usability of the car rental website Budget.com. The CUE-8 study documented a wide difference in measurement approaches. Teams that used similar approaches often reached similar results. This paper discusses a number of common pitfalls in usability measurements. This paper also points out a number of fundamental problems in unmoderated measurement studies, which were used by 6 of the 15 participating teams.

Keywords

usability measurement, comparative usability evaluation, SUS, time-on-task, user satisfaction, unmoderated studies, unattended evaluations

Introduction

Traditional usability tests are usually a series of moderated one-on-one sessions that generate both qualitative and quantitative data. In practice, a formative usability test typically focuses on qualitative data whilst a summative test focuses on performance metrics and subjective satisfaction ratings.

Qualitative testing is by far the most widely used approach in usability studies. However, usability practitioners are discovering that they need to accommodate engineers, product managers, and executives who are no longer satisfied with just qualitative data but insist on performance measurements of some type. Quantitative usability data are becoming an industry expectation.

The current literature on quantitative methods aimed at practitioners is limited to a book by Tullis and Albert (2008), a website by Sauro (2009), and UsabilityNet—a project funded by the European Union (Bevan, 2006) and a few commercial offerings, for example, Customer Carewords (2009). All base their measures on the ISO 9241-11 (1998) definition of usability. Tullis and Albert's book describes the what, why, and how of measuring user experience from usability practitioner viewpoints. Customer Carewords focuses on websites and introduces several additional metrics such as disaster rate and optimal time. UsabilityNet identifies a subset of resources for Performance Testing and Attitude Questionnaires.

There are several psychometrically designed questionnaires for measuring satisfaction. Two of these are the System Usability Scale (SUS; Brooke, 1996) and the Website Analysis and MeasureMent Inventor (WAMMI) questionnaire (Claridge & Kirakowski, 2009). Many companies use their own questionnaires, but these may not have sufficient reliability and validity. Some instruments have also been developed to assess user mental effort as an alternative to satisfaction, for example Subjective Mental Effort Questionnaire (SMEQ; Zijlstra, 1993) and NASA-Task Load Index (TLX; Hart, 2006).

Sauro's work and applications are mostly a result of statistical analyses of real world usability data and go as far as proposing a method to compute a single composite usability metric called Single Usability Metric (SUM; Sauro & Kindlund, 2005). Tullis and Albert, and Sauro, stress the importance of strict participant screening criteria and reporting confidence intervals, especially with small sample sizes. However, it is not known how much of this and other recommended practices have actually been taken up by the industry.

There is, therefore, a need for information about best practice in usability measurements for practitioners. This formed the basis for the CUE-8 study, the outcomes of which are reported in this paper.

About CUE

This study is the eighth in a series of Comparative Usability Evaluation (CUE) studies conducted in the period from 1998 to 2009. The essential characteristic of a CUE study is that a number of organizations (commercial and academic) involved in usability work agree to evaluate the same product or service and share their evaluation results at a workshop. Previous CUE studies have focused mainly on qualitative usability evaluation methods, such as think-aloud testing, expert reviews, and heuristic inspections. An overview of the eight CUE studies and their results are available at DialogDesign's website (Molich, 2009).

Goals of CUE-8

The main goals of CUE-8 were

- to allow participants to compare their measurement and evaluation skills to those of their peers and learn from the differences,
- to get an impression of the methods and techniques used by practitioners to measure usability, and
- to discuss and identify best practices in measuring usability.

Method

In May 2009, 15 U.S. and European teams independently and simultaneously carried out usability measurements of the Budget.com website (see Figure 1). The measurements were based on a common scenario and instructions (Molich, Kirakowski, Sauro, & Tullis, 2009).

The scenario deliberately did not specify in detail which measures the teams were supposed to collect and report, although participants were asked to collect time-on-task, task success, and satisfaction data as well as any qualitative data they normally would collect. The anonymous reports from the 15 participating teams are publicly available online (Molich, 2009).

Teams were recruited through a call for participation in a UPA 2009 conference workshop.

After conducting the measurements, teams reported their results in anonymous reports where they are identified only as Team A ... Team P. The teams met for a full-day workshop at the UPA conference.

Figure 1. The Budget.com home page as it appeared in May 2009 when CUE-8 took place.

The following is the common measurement scenario:

> *The car rental company Budget is planning a major revision of their website, www.Budget.com.*
>
> *They have signed a contract with an external provider to create the new website. Budget wants to make sure that the usability of the new website is at least as good as the usability of the old one. They want you to provide an independent set of usability measurements for the current website. These measurements will provide a baseline against which the new website could be measured by another provider.*
>
> *Your measurements must be made in such a way that it will later be possible to verify with reasonable certainty that the new website is at least as good as the old one. The verification, which is not part of CUE-8, will be carried out later by you or by some other contractor.*
>
> *Budget wants you to measure time on task and satisfaction for ... five key tasks.... Budget has clearly indicated that they are open to additional measurements of parameters that you consider important.*

> *Budget recently has received a number of calls from journalists questioning the statement "Rent a car in just 60 seconds," which is prominently displayed on their home page. Consequently, they also want you to provide adequate data to confirm or disconfirm this statement. If you disconfirm the statement, please suggest the optimal alternative that your data supports and justify it.*

The scenario is realistic but fictitious. The workshop organizers had limited contact with Budget.com, and they had no information on whether Budget was planning a revision of their website.

The measurement tasks were prescribed to ensure that measurements were comparable. The following were the five tasks:

1. Rent a car: Rent an intermediate size car at Logan Airport in Boston, Massachusetts, from Thursday 11 June 2009 at 09:00 a.m. to Monday 15 June at 3:00 p.m. If asked for a name, use John Smith and the email address john112233@hotmail.com. Do not submit the reservation.
2. Rental price: Find out how much it costs to rent an economy size car in Myrtle Beach, South Carolina, from Friday 19 June 2009 at 3:00 p.m. to Sunday 21 June at 7:00 p.m.
3. Opening hours: What are the opening hours of the Budget office in Great Falls, Montana on a Tuesday?
4. Damage insurance coverage: An unknown person has scratched your rental car seriously. A mechanic has estimated that the repair will cost 2,000 USD. Your rental includes Loss Damage Waiver (LDW). Are you liable for the repair costs? If so, approximately how much are you liable for?
5. Rental location: Find the address of the Budget rental office that is closest to the Hilton Hotel, 921 SW Sixth Avenue, Portland, Oregon, United States 97204.

Measurement Approaches

Table 1. Key Measurement Approaches

Approach	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
Participants total	22	9	20	14	11	15	12	60	20	20	313	43	10	15	20
Participants moderated	22	9	20	0	11	0	12	3	7	20	0	0	10	15	20
Participants unmoderated	0	0	0	14	0	15	0	57	13	0	313	43	0	0	0
# team members	8	2	1	1	1	1	1	1	1	7	1	2	4	2	1
Person hours used	30	81	24	28	26	30	40	38	59	88	21	44	80	39	128
Questionnaire	W,M	S	O	A	O	O	S	S	O	S	S	S	S,N	O	S
Time measured for	T	T	T	U	T	T	T	U	U	T	CS	CS	T	T	T
Result verified by	M	M	M	U	M	P	M	U	U	M	C	C	M	M	M
Include failed tasks	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No

Questionnaire: A=ASQ, M=SMEQ, N=NASA TLX, O=Own, S=SUS, W=WAMMI.
Time measured for: CS=Comprehend and complete task, T=Task completion, U=User defined
Result verified by: C=Multiple choice, M=Moderator, P=Professional, U=User

As shown in Table 1, nine teams (A, B, C, E, G, K, N, O, and P) used "classic" moderated testing. They used one-on-one sessions to observe 9 to 22 participants completing the tasks.

Six teams partly or wholly used unmoderated sessions. Teams sent out tasks to participants and used a tool to measure task time. Some teams used multiple-choice questions following each task to get an impression of whether the task had been completed correctly or not.

Four teams (D, F, L, and M) solely used unmoderated testing. Teams D, L, and M used a tool to track participant actions, collect quantitative data, and report results without a moderator in attendance. These teams recruited 14 to 313 participants and asked participants to complete the tasks and self-report. These teams used tools to measure task completion time. Team F used a professional online service (usertesting.com) to recruit and to video record users working from their homes; the team then watched all videos and measured times.

Two teams (H and J) used a hybrid approach. They observed 3 to 7 participants in one-on-one sessions and asked 13 to 57 other participants to carry out the tasks without being observed.

Team G included a comparative analysis of corresponding task times for Avis, Enterprise, and National. They also did keystroke level modeling to predict experienced error free task times.

Test Tasks

All teams gave all five tasks to users. Most teams presented the tasks in the order suggested by the instructions, even though this was not an explicit requirement. Team K and O repeated the car rental tasks (task 1 and 2) for similar airports after participants had completed the five given tasks. These teams reported significant decrease in time with repeated usage; task times for the repeated tasks were often less than half of the original times.

Measurements

All teams except one reported time-on-task in seconds. Team A reported time-on-task to the nearest minute. Some teams included time from task start until participants gave up in their time-on-task averages. Some of the teams that used unmoderated testing included time to understand the task in their time-on-task.

Other metrics reported included

- # of clicks minus # of clicks of the optimal path (to measure efficiency), and
- lostness ratio (optimal pageviews : average pageviews; optimal # of clicks : average # of clicks).

Key Measurement Results

Table 2. Reported Key Measurement Results for Task 1, Rent a Car—All Times Are in Seconds

Key results	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
Reported time-on-task	180	133	134	323	210	209	157	108	207	195	148	251	451	306	328
Minimum time	60	66	105	156	93	128	74	0	60	126	18	110	243	180	134
Maximum time	900	242	172	647	373	1001	260	1244	349	353	570	1217	1012	582	677
Confidence low	141	103	123	269	145	162	113	63	171	170	139	216	328	199	260
Confidence high	327	163	143	402	258	293	219	154	243	220	157	288	574	413	395
Success rate	95	89	91	21	98	93	83	34	65	75	97	63	60	73	90
Rent car in xx secs	R	180	120	M	OK	240	OK	M	R	OK	OK	90		M	180
Qualitative results	12	No	5	No	No	16	3	6	No	17	68	No	79	No	19

Rent car in xx seconds: M=More research needed, OK=Current statement "Rent a car in just 60 seconds" is OK or defensible, R=Rephrase statement, number=replace "60" with number.
Team A, C, D, E, H, K, N, and O did not provide confidence intervals in their reports. Confidence intervals were provided after the workshop.

Eleven of the fifteen teams reported the mean (average) of their time-on-task measurements. Three teams (A, D, and F) reported the median and one team (G) reported the geometric mean.

Seven of the fifteen teams reported confidence intervals around their average task times. Confidence intervals are a way to describe both the location and precision of the average task time estimates. They are especially important for showing the variability of sample sizes. Eight teams left this information off (we are unsure if they calculated it). Some of these teams indicated that they were not computing confidence intervals on a regular basis. Another claim

that was heard among these teams was that "you cannot come to statistically significant conclusions with such small sample sizes." The information was provided after the workshop.

The methods used to compute the confidence intervals for time-on-task were

- Team B, J, L, M, P: MS Excel Confidence-function; and
- Team F, G: The "Graph and Calculator for Confidence Intervals for Task Times" (Sauro, 2009).

Team L, M, and P did not provide information in their report about the computation method. The information was provided after the workshop.

Nine of the fifteen teams provided qualitative findings even though qualitative findings were not a subject of this workshop. Teams argued that they obtained a considerable insight during the measurement sessions regarding the obstacles that users faced and that it would be counterproductive not to report this insight.

The format and number of reported qualitative findings varied considerably from 3 qualitative findings provided by team G to 79 qualitative comments, including severity classifications, provided by team N. Qualitative findings were reported in several formats: problem lists, user quotes, summarized narratives, word clouds, and content analysis tables.

"Rent a Car In Just 60 Seconds"

The scenario asked teams to provide data to confirm or disprove the prominent Budget home page statement "Rent a car in just 60 seconds" (see Figure 1) and to suggest an alternative statement, if required.

Table 2, row "Rent car in xx secs," shows that there was little agreement, with one team choosing not to respond. Several teams suggested that Budget's statement may indicate a lack of understanding of users' needs, because their study showed that time was not of the utmost importance to participants—it was more important to get the right car and the right price, and to ensure that the reservation was correct. In other words, Budget's implicit assumption that users value speed above all may be wrong. It was also suggested that the statement could rush users through the process.

Four teams suggested that the current statement was OK or defensible. Team L said, "Since it is theoretically possible to rent a car in less than 60 seconds, it is technically not an incorrect statement and thus it is OK to keep it." Team E and G argued similarly. Team K suggested that Budget should "focus on improving the efficiency and intuitiveness of the car rental process and pay less emphasis on getting the time you spend on renting a car through the site."

Three teams suggested that additional research was required, but none provided further details about what was required.

The five teams that provided specific times suggested times from 90 to 240 seconds; it was not always completely clear how they arrived at their suggestion.

We have asked Budget to comment on these findings, to no avail.

Reported Time-on-Task

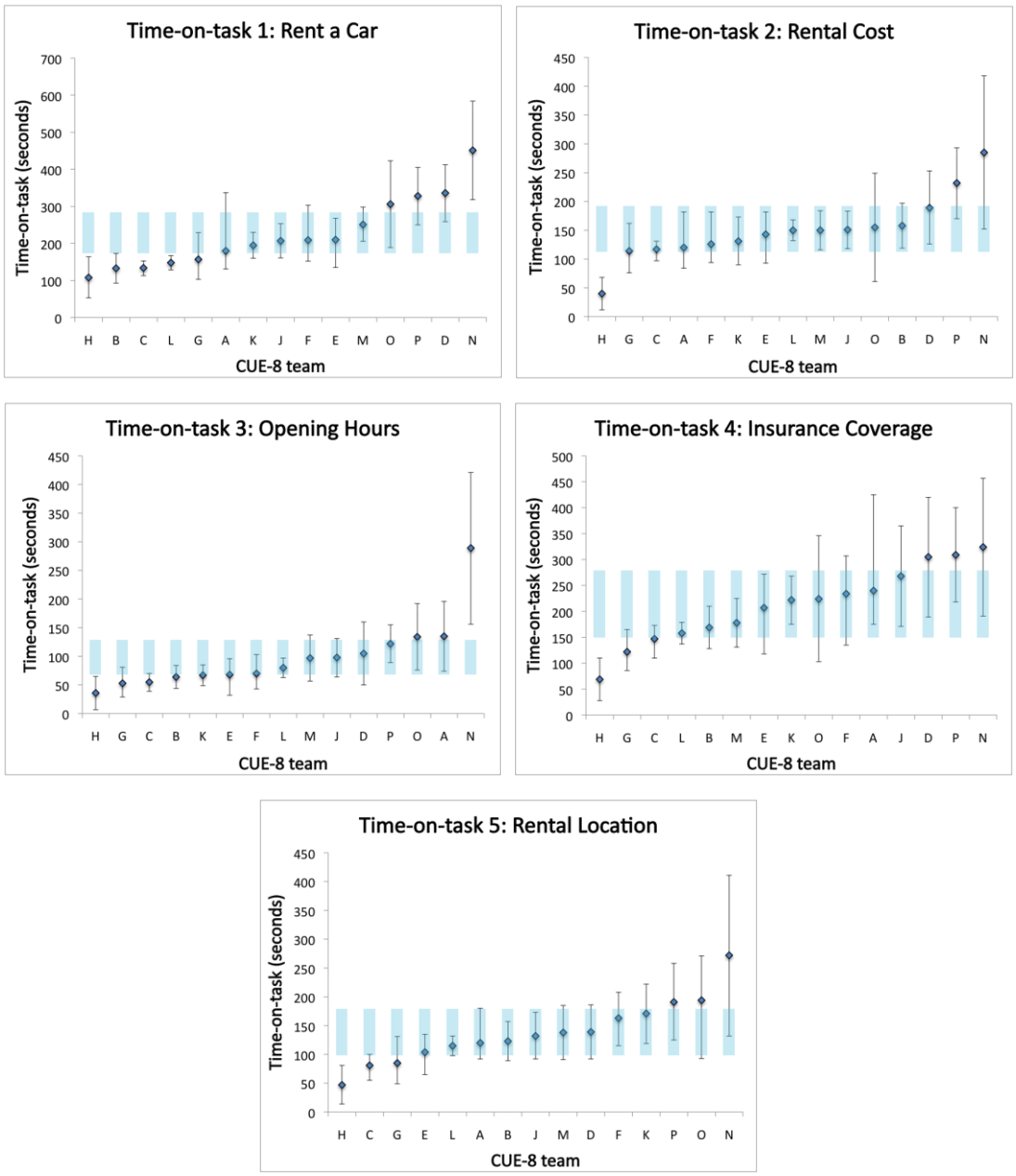

Figure 2. Reported time-on-task for tasks 1-5: The diamonds show the time-on-task reported by each team. The black vertical lines show the 95% confidence interval reported by the team before the workshop, or computed after the workshop for the teams that initially did not provide confidence intervals. The light blue bars show the average of the Confidence Intervals (CIs) for each task, that is, average upper CI limit to average lower CI limit. The graphs also show that some teams reported completion times that are not centered between the upper and lower limit of the confidence interval, for example team A, task 1. This is because these teams chose to report the median as the central measure.

Satisfaction Measurements

Of the 15 teams who participated in CUE-8, eight teams used the System Usability Scale (SUS) as the post-test standardized questionnaire. Four of the teams used their own in-house questionnaires and one used a commercially available questionnaire (WAMMI). Of the teams who used SUS, one team modified the response options to 7-scale steps instead of five. The data from this team was not used as part of the SUS analysis. The remaining seven teams left the scale in its original 5-scale form and provided scores by respondent.

Discussion

The following sections discuss that CUE-8 is not a scientific experiment, the measurement approaches, computing time-on-task, reporting uncertainty for time-on-task, qualitative results, reproducibility of results, participant profiles, satisfaction measurements, handling failed tasks, productivity, contaminated data, measuring time-on-task, and the usability of a remote tool.

CUE-8 Is Not a Scientific Experiment

From previous CUE studies (Molich, 2009) it is known that there is a wide range of approaches by professional teams when undertaking qualitative evaluations. The main motivation for CUE-8 was to see to what extent there was variation in approach to quantitative measurement.

The teams who participated were essentially a convenience sample. They were not recruited specifically to provide either best practices or state-of-the-art measurement techniques. For this reason we abandoned one of our original goals, to investigate whether usability measurements are reproducible. Nevertheless, there was a good mixture of qualifications between teams within CUE-8. Although the range of variation was wide, it could well be wider if a more systematic random sample of teams over the world was taken, but such a comprehensive systematic study would most likely be cost prohibitive.

The results from CUE-8 cannot therefore be generalised or summated into averages with sampling confidence intervals to produce overall trends. Methodological purity of this kind is not accessible in the real world. What we present is essentially 15 separate case studies showing 15 different approaches to the quantitative measurement and reporting of time, performance, and satisfaction. The benefit of CUE-8 is that in this area of quantitative evaluation, we can comment on what appear to us to be the strengths and weaknesses within each case study and present them as take aways.

Measurement Approaches

Team C discouraged participants to think aloud. Team K and N on the other hand explicitly asked participants to think aloud. At the workshop it was discussed whether think aloud increases or decreases total time-on-task. Some argued that the mental workload increases when thinking aloud, thus causing additional problems to arise. It may also impact task completion time, as participants tend to occasionally pause their task solving to elaborate on an issue. Others argued that think aloud forces participants to consider their moves more carefully thus decreasing time-on-task. This would be an interesting variable to investigate further.

A take away from the study was that instructions and tasks must be precise and exhaustive for unmoderated, quantitative studies because there was no moderator to correct misunderstandings. Even in moderated studies the moderator should not have to intervene because this influences task time. Unfortunately, this often means that tasks get quite long, so participants don't read all of the instructions or tasks. For example, in order to provide the necessary details, task 1 and 4 became so wordy that some participants overlooked information in the tasks. Some teams declared measurements from misinterpreted tasks invalid; others did not report their procedures. Instructions and tasks should be tested carefully in pilot tests.

Our study shows that task order is critical; there is a substantial learning effect as shown by the two teams who repeated task 1 and 2 after task 1-5.

A few teams presented the tasks in random order. At the workshop it was pointed out that this conflicts with a reasonable business workflow; task 1 or 2 should be first because the vast majority of users would start with one of these tasks.

Computing Time-on-Task

There is substantial agreement within the measurement community that measures such as time-on-task are not normally distributed because it is common to observe a positive skew in such data, that is, there is a sharp rise from the start to the center point of the distribution but a long tail back from the center to the end. Under such conditions, the mean is a poor indicator of the center of a distribution. The median or geometric mean is often used as a substitute for the mean for heavily skewed distributions (Sauro, 2009). Using the median censors data or discards extreme observations.

There are, as alternatives, a variety of statistical techniques that will "correct" a skewed distribution in order to make it symmetrical and therefore amenable to summary using means and standard deviations. Team F and G used such an approach. The rest reported time-on-task the way it is usually reported in the HCI literature: untransformed data are the norm.

Reporting Uncertainty in Time-on-Task

At the workshop it was argued that usability practitioners mislead their stakeholders if they were not reporting confidence intervals. Understanding the variability in point estimates from small samples is important in understanding the limits of small sample studies. Confidence intervals are the best way to describe both the location and precision of the estimate, although the mathematical techniques of computing confidence intervals on sample distributions from non-normal populations are still a matter of controversy in the statistical literature.

In order to compare teams' confidence intervals, all teams must meet the same screening criteria for participants. As discussed in the Participant Profiles section, this was not the case in this study where convenience samples were often employed.

If the sample on which the measures were taken is from a normally distributed population, the mean is a useful measure of the average tendency of the data, and the variance is a useful measure of variability of the data. The confidence interval is a statistic that is derived from the computation of the variance and also assumes normality of population distribution.

Because time-on-task is not normally distributed, means, variances, and confidence intervals derived from variances are possibly misleading ways of estimating average tendency and variability. There are a number of ways of getting over this as was displayed in our teams: some teams used medians (which are not sensitive to ends of distributions), others used a transformation that would "normalize" the distributions mathematically (Sauro, 2009).

Qualitative Results

As shown in Table 2, four of the reports included 10-20 qualitative usability findings. This seems to strike a useful balance between reporting informally a few quantitative findings, which 6 teams did, and reporting a high number of qualitative findings, which team L and N did (68 and 79 findings, respectively).

Reproducibility of Results

Did the teams get the same results? The answer is no, but the reported measurements from several teams—sometimes a majority—agree quite well as you can see from Figure 2.

Eyeballing shows that the results from six teams (A, E, F, J, K, and M) were in reasonable agreement for all five tasks. Two more teams (B and L) agreed with the six teams for all tasks except task 1. Two teams (D and O) agreed with the majority for three tasks. On the other hand, five teams mostly reported diverging results. Team H and N consistently diverged from the other teams.

We examined the overlap between confidence intervals for any pair of teams for each task. Overlap scores were computed based on the overlap between confidence intervals in percent. For example, for task 1 team B reported the confidence interval [103, 163], while team E reported [145, 258]. Team B's overlap score is (163-145)/(163-103) = 30%, while team E's comparable score is (163-145)/(258-145) = 16%. The total overlap score for a team is the sum of the team's 14*5=70 overlap scores for the 14 other teams and 5 tasks. This scoring method favors teams that reported narrow confidence intervals that overlap with the wider intervals of many other teams, which seems fair. Team L did best by this scoring method, followed by team J and M. See the complete results in Figure 3.

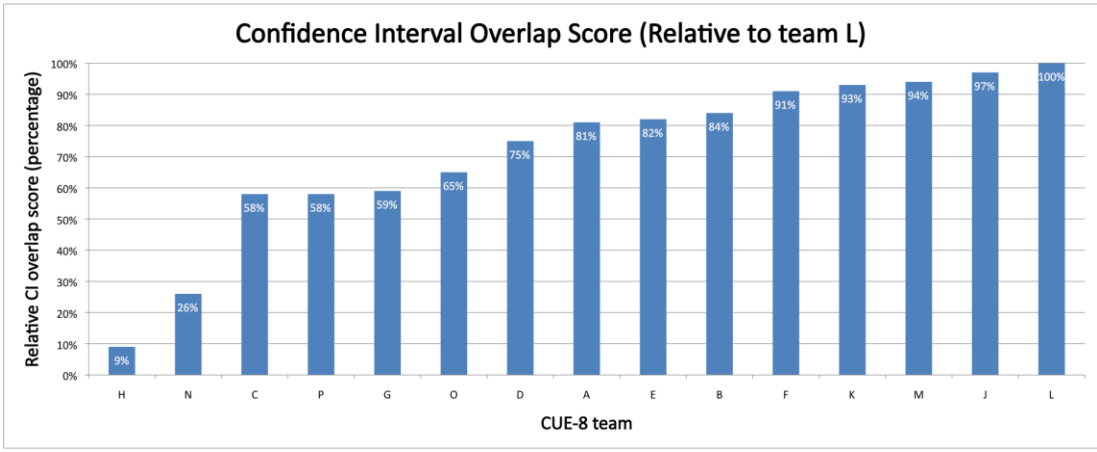

Figure 3. Overlap between findings: This scoring attempts to quantify the eyeballing "Yeah, I think the overlap between team J, K, L, and M is pretty good."

An analysis of the teams' approaches reveals the following sources for diverging results:

- Equipment error, such as reporting a task time of zero seconds, which team H did. It is difficult to assess with complete certainty that any given reading at the extremes of a distribution is due to equipment error, although a task time of zero surely must be.
- External factors such as a poor Internet connection.
- Participants who did not follow instructions.
  See the section Cleaning Contaminated Data.
- Participants who repeatedly had to consult task descriptions while they were working on a task, especially if it was awkward to move between the online instructions and the test site.
- Not recruiting sufficiently representative users of the site.
  See the subsection Participant Profiles.
- Definition of "time-on-task"
  See the subsection Measuring Time-on-Task.
- Lack of experience: All participating teams were professional in the sense that team members get paid to do usability work. A few teams acknowledged to having never conducted a quantitative usability evaluation before. Their motivation for participating may have been the opportunity of getting to know this specific area better. Whether or not they would have agreed to participate, had the evaluation been actual consultancy work for budget.com, rather than a workshop is not certain, but it is clear that the results reported by these teams differed from the rest.

Participant Profiles

Recruiting was an important reason that some teams reported diverging results. Not all teams seemed to use strict participant screening criteria; some used convenience samples.

The following are examples of questionable recruiting:

- Some of the participating European teams recruited participants who did not have English as a primary language. This caused both language and cultural biases. Task 4 (Loss Damage Waiver conditions) was particularly affected by this. One team selected participants mainly based on sufficient knowledge of English.
- Even if the European participants had good English, Budget.com is not for Europeans. Budget has separate websites such as Budget.be, Budget.dk, and Budget.co.uk for Europeans—even for renting cars in the U.S.
- Only team F, O, and P had resources to pay for their participants. Because of funding problems some teams recruited friends and colleagues (in particular, students) instead of a representative sample of Budget.com users. Some teams recruited only coworkers.

- Team F recruited users through usertesting.com. Similarly, team L's participants were all coworkers that were used to using an in-house online test tool and participating in the company's tests. Logistically, this worked well but at the workshop it was pointed out that the participants might be "professional" usability testers who conducted many test sessions per month. We don't know if and how this affected results.

Because we had no contact with Budget, it was not technically possible to recruit people who were actually visiting the site.

Satisfaction Measurements

As with many of the metrics collected, there was variability in the SUS scores the teams reported. The SUS scores are shown in Figure 4. An analysis of overall variance shows that there is a statistical difference between SUS scores, $F(6,451)=6.73$, $p < .01$, which can be attributed to variation between teams. There are two groups of scores that differ significantly from each other (between each group), but which do not differ statistically within each group.

One cluster of four teams (B, K, L, and G) generated SUS scores within 7-10% of each other (73, 77, 78, and 80). The other cluster of three teams scores (M, H, and P) were within 4-6% of each other (62, 65, and 66). Table 3 shows the mean and standard deviation for the teams' SUS scores in the two clusters.

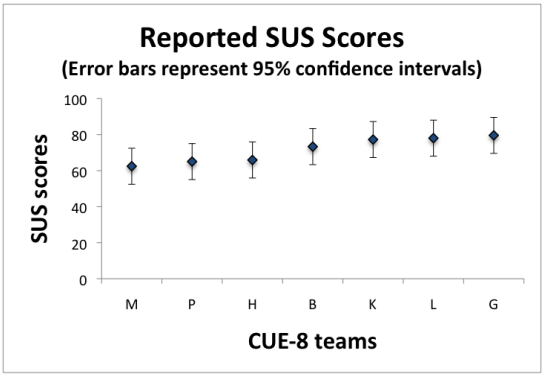

Figure 4. Reported SUS scores for the seven teams that used the original SUS measurement

Table 3. Mean SUS Scores Organized by Clusters of Scores, Standard Deviation (SD)

Cluster	Team	Mean	SD	Average
1	B	73.3	18.0	77.1
	K	77.3	12.0	
	L	78.0	19.9	
	G	79.6	10.8	
2	M	62.4	19.6	64.5
	H	65.9	22.2	
	P	65.0	19.3	

Other researchers (Bangor, Kortum, & Miller, 2008) have pointed out that SUS has shown to be positively skewed with an "average" score for websites and web-applications of approximately 68 out of 100 (with a standard deviation of 21.5). The average SUS score for the first cluster of 77.1 suggests Budget.com is a better than the average website falling in the 66th percentile of the Bangor et al. dataset. The average SUS scores for the second cluster of 64.5 suggests Budget.com is a worse than average website falling in the 43rd percentile of the Bangor et al. dataset.

It is unclear whether or not the differences observed in the SUS scores are a reflection of SUS being inadequate for measuring websites. It is likely that many of the observed differences occurred due to the different participants and evaluation procedures used by the teams.

The SUS scores can be contrasted with the score from the one team who used WAMMI (team A). The Budget.com WAMMI Global User Satisfaction score was in the 38th percentile suggesting it to be a below average website (the industry average is at the 50th percentile). Because there was only a single WAMMI data point, it is not possible to know how much more or less WAMMI scores would fluctuate compared to the SUS scores.

Handling Failed Tasks

Ten out of 15 teams chose to include time-on-task for tasks where participants gave up or obtained an incorrect answer in their calculations of mean or median time-on-task. For the failed tasks they used the time until the participant gave up. At the workshop, it was successfully argued that these figures are incompatible. The time until a participant gives up or finds an incorrect result is irrelevant for time-on-task; reported time-on-task should include only data from successfully completed tasks. Failed times are still useful as their own metric called *average time to task failure*. If you only report one measure, then report task completion times and exclude failure.

Some teams argued that three separate results are of equal importance: time for successful completion, success rate (or failure rate), and disaster rate—that is, the percentage of participants who arrive at a result they believe in but which is incorrect.

It is useful to differentiate between task success and failure in satisfaction results. For instance one may report that "X% of participants who successfully completed task 2 gave a satisfaction score higher than Y%."

Some teams opted for using a binary code for task success, whereas other teams used error-based percentages to classify success (0/50/100% or even 0/25/50/75/100%).

Productivity—Team Hours Spent per Participant

In examining the average times spent on moderated versus unmoderated or hybrid studies (60 hours vs. 37 hours), there was surprisingly a lot of overhead for unmoderated tests.

Productivity varied remarkably from 4 minutes per participant (team L) to 9 hours per participant (team B). Median productivity was 2:22 hours per participant.

Unmoderated studies pay off when the number of participants gets large. For example, team L ran 313 unmoderated participants in 21 hours, which is about 4 minutes per participant, whereas team D ran 14 unmoderated participants in 28 hours, which is about 2 hours per participant, similar to the moderated test ratio. Team L had both the largest number of participants (313) and the lowest person hours used (21). This impressive performance, however, came at a cost as described in the next section.

Cleaning Contaminated Data, or Killing the Ugly Ducklings

Teams who used unmoderated sessions all reported some unrealistic measurements.

Table 2, row "Minimum time," shows that few observed participants were able to complete the rental task in anywhere near 60 seconds. Teams agreed that it was impossible even for an expert who had practiced extensively to carry out the reservation task (task 1) in less than 50 seconds. Yet, team H reported a minimum time of 0 seconds for successful completion of this task; 22 of their 57 measurements were below 50 seconds. Team L reported a minimum time of 18 seconds for successful completion of the same task; 6 of their 305 measurements for this task were below 50 seconds.

Some of the teams decided to discard measurements that appeared to be too fast or too slow, in other words, they decided to "to kill the ugly ducklings." For example, the cleaning procedure used by team L was to delete

- any participants whose total time for the study was less than 4 minutes (10 participants),
- any participants who got none or only one task correct (8 participants),
- any individual task time <5 seconds because the participant was probably not doing the task in earnest, and
- any individual task time >600 seconds because the participant was probably interrupted.

It is not clear from the reports how teams came up with criteria for discarding measurements. Apparently, the criteria were based on common sense rather than experience or systematic studies. Team M used click path records to check measurements that looked suspicious; they also discarded measurements where "the test tool must have encountered a technical problem capturing page views and clicks across all tasks."

The teams hypothesized that participants had either guessed or pursued other tasks during the measurement period. However, by discarding data based solely on face value, teams admitted that their data were contaminated in unknown ways. It could then be argued that other data that appeared valid at first glance were equally contaminated. Example: Team F analyzed the data from their unmoderated videos and found measurements that appeared realistic but were invalid. They also found a highly suspicious measurement where the participant used almost 17 minutes to complete the rental task, which turned out to be perfectly valid; the participant looked for discounts on the website and eventually found a substantial discount that no one else discovered.

Measuring Time-on-Task

In both moderated and unmoderated testing it is difficult to compensate for the time used by the participant to read the task multiple times while solving the task. In unmoderated testing it is difficult to judge if the participant has found the correct answer unless they include video recordings or click maps, which may take considerable time to analyze. Multiple-choice questions are an option, which was used by some teams as shown in Figure 5. However, some of the answers changed during the period where the measurements occurred making all choices incorrect, and some participants might have been able to guess the right answer from the multiple choice list.

At the workshop, team M argued that automated unmoderated tools such as Userzoom, which they used, allowed them to validate answers also by URLs (reaching a certain page as a way to validate the successful or unsuccessful completion of a task). Team M argued that they had more issues with moderated sessions where moderators and participants were discussing issues while the clock was measuring time-on-task.

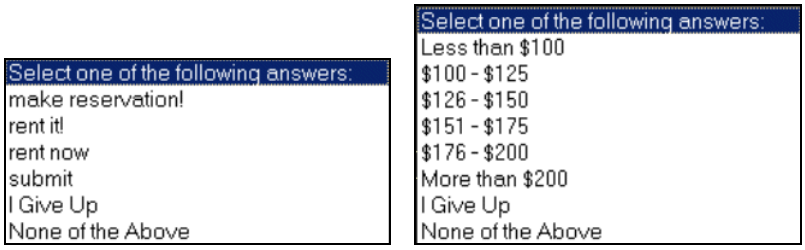

Figure 5. Multiple choice answers for task 1 and 2 used by team L for determining if their participants had obtained the right answer in unmoderated sessions. For task 1 participants were asked to find the label of the button that performed the rental; the correct answer is "rent it!" For task 2 the answer varied. Most often the rental price was in the $176-$200 range, but on some days it was more than $200.

The discussion at the workshop concluded that it is not completely clear what you are measuring as time-on-task in an unmoderated session: Time to complete task? Time to comprehend and complete task? Time to comprehend and complete task plus time to remember to click the "Task finished" button? Time to complete test task and other parallel tasks? Also, irrelevant overhead varied considerably from participant to participant. Some grasped a task almost instantly, some printed the task, and some referred back to the task description again and again. In moderated sessions, in comparison, the moderator ensures some consistency in measuring time-on-task.

Usability of Remote Tool

The ease of use of the remote tool, the clarity of the instructions, etc., has a considerable impact on unmoderated participants' performance. For example, one of the teams used a tool that hid the website when participants indicated that they had completed a task; this made it

unrealistically difficult for participants to answer the follow-up questions that checked whether or not the task was completed correctly.

Conclusion

Usability metrics expose weaknesses in testing 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 know how reliable results are or how consistent methods are when all you are producing are problem lists. Of course, you can also show anything you want with statistics—but while you can, it is harder with statistics than without.

Unmoderated measurements are attractive from a resource point of view; however, data contamination is a serious problem and it is not always clear what you are actually measuring. While both moderated and unmoderated testing have opportunities for things to go wrong, it is more difficult to detect and correct these with unmoderated testing. We recommend further studies of how data contamination can be prevented and how contaminated data can be cleaned efficiently.

For unmoderated measurements the ease of use and intrusiveness of the remote tool influences measurements. Some teams complained about clunky interfaces. We recommend that practitioners demand usable products for usability measurements.

Practitioner's Take Away

CUE-8 confirmed a number of rules for good measurement practice. Perhaps the most interesting result from CUE-8 is that these rules were not always observed by the participating professional teams.

- Adhere strictly to precisely defined measurement procedures for quantitative tests.
- Report time-on-task, success/failure rate and satisfaction for quantitative tests.
- Exclude failed times from average task completion times.
- Understand the inherent variability from samples. Use strict participant screening criteria. Provide confidence intervals around your results if this is possible. Keep in mind that time-on-task is not normally distributed and therefore confidence intervals as commonly computed on raw scores may be misleading.
- Combine qualitative and quantitative findings in your report. Present what happened (quantitative data) and support it with why it happened (qualitative data). Qualitative data provide considerable insight regarding the serious obstacles that users faced and it is counterproductive not to report this insight.
- Justify the composition and size of your participant samples. This is the only way you have to allow your client to judge how much confidence they should place in your results.
- When using unmoderated methodologies for quantitative tests ensure that you can distinguish between extreme and incorrect results. Although unmoderated testing can exhibit a remarkable productivity in terms of user tasks measured with a limited effort, quantity of data is no substitute for clean data.

Acknowledgements

We thank the anonymous reviewers for their insightful comments on this paper.

The following were the CUE-8 participants:

- Aalborg University (DK): **Janne Jul Jensen**, **Anders Bruun**, Jan Stage, Mikael B. Skov
- Bentley College (US): **Tamara Rose**
- Bureau of Labor Statistics (US): **Jean Fox**
- DialogDesign (DK): **Rolf Molich**
- FamilySearch.org (US): **Tyson Stokes**

- Fidelity Investments, team 1 (US): **Donna P. Tedesco**, Marisa Siegel
- Fidelity Investments, team 2 (US): **Tom Tullis**
- Google, Inc. (US): **Tomer Sharon**, **Daniel "Danny" Wildman**
- Kingston University (UK): **Jarinee Chattratichart**, Martin Colbert
- Mount Royal University (CAN): **Brian Traynor**
- Nielsen Norman Group (US): **Janelle Estes**
- Oracle Corporation/Measuring Usability LLC (US): **Jeff Sauro**
- University of Cork (IRL): **Jurek Kirakowski**, **Raegan Murphy**, Mike Brown, Noirin Curran, Nigel Claridge, Eve Griffin, Mary Joyce, Anthony Yiu
- UserPlus (BE): **Lonneke Spinhof**
- Wichita State University (US): **Veronica D. Hinkle**, **Amanda Smith**, Barbara Chaparro, Doug Fox, David Libby, Shivashankar Naidu, Justin Owens

The 19 people whose names appear in bold participated in the workshop. The teams are listed in an order that provides no clues regarding their team code.

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24, 574—594.

Bevan, N. (2006). About UsabilityNet. Retrieved on November 23, 2009 from http://www.usabilitynet.org/about/aboutusa.htm

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

Claridge, N., & Kirakowski, J. (2009). WAMMI - What is WAMMI? Retrieved on October 28, 2009 from http://www.wammi.com/whatis.html

Customer Carewords. (2009). Task Performance Indicator. Retrieved on September 11, 2009 from http://www.customercarewords.com/task-performance-indicator.html

Hart, S. G. (2006). Nasa-Task Load Index (Nasa-TLX); 20 Years Later. Human Factors and Ergonomics Society Annual Meeting Proceedings, General Sessions, pp. 904-908(5). Human Factors and Ergonomics Society.

ISO 9241-11 (1998). International Organization for Standardization – Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability.

Molich, R. (2009). CUE–Comparative Usability Evaluation. Retrieved on October 1, 2009 from http://www.dialogdesign.dk/cue.html

Molich, R., Kirakowski, J., Sauro, J., & Tullis, T. (2009). Comparative Usability Task Measurement (CUE-8) instructions. Retrieved on October 1, 2009 from http://www.dialogdesign.dk/CUE-8.htm

Sauro, J. (2009). Measuring usability - Quantitative usability, statistics & six sigma. Retrieved on September 11, 2009 from http://www.measuringusability.com/

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA.

Tullis, T., & Albert, B. (2008) *Measuring the user experience*. Burlington MA: Morgan Kaufmann.

Zijlstra, R. (1993). *Efficiency in work behaviour: A design approach for modern tools*. Delft, Netherlands: Delft University Press.

About the Authors

Rolf Molich
Molich owns and manages DialogDesign, a small Danish usability consultancy. Rolf conceived and coordinated the comparative usability evaluation studies CUE-1 - CUE-8 where a total of almost 100 professional usability teams tested or reviewed the same applications. Rolf is the co-inventor of the heuristic inspection method (with Jakob Nielsen).

Jarinee Chattratichart, Ph.D.
Chattratichart is a senior lecturer at Kingston University, UK. As a lecturer, she enjoys raising students' awareness of HCI and teaching user interface design, requirements engineering, UML, and database design. As a researcher, her major interest and publications are usability evaluation methods and performance metrics.

Veronica Hinkle
Hinkle has an M.A. in Anthropology and currently works on her Ph.D. Dissertation in Human Factors Psychology at Wichita State University. She is a researcher at the Software Usability Research Laboratory (SURL) and her research interests include qualitative and quantitative methodology, user-centered design, user experience, and interface design.

Janne Jul Jensen, Ph.D.
Jensen is a Postdoc at Aalborg University currently working on the project "Web Portal Usability." She earned her PhD degree in HCI in 2009. She has collaborated internationally with academia and privately held companies on projects spanning web accessibility for motor handicapped users to usability for children, resulting in numerous publications.

Jurek Kirakowski, Ph.D.
Kirakowski is the Director of the Human Factors Research Group. He participated in the first ever CUE workshop as well as numerous EC-funded projects on usability testing and measurement. For the past 30 years, he has spent a lot of his time developing and standardising user satisfaction questionnaires.

Jeff Sauro
Sauro is a Six-Sigma trained statistical analyst focusing on quantitative usability methods and has published and presented on the topic at CHI, UPA, HCII, and HFES conferences. He holds a Masters from Stanford University, maintains the website MeasuringUsability.com, and currently works for Oracle in Denver, CO.

Tomer Sharon
Sharon is a user experience researcher at Google New York since 2008, a Bentley University graduate, a founder and past president of UPA Israel, and a former UX Magazine editorial board member. In the field since 1999, he has worked as a researcher at Check Point Software Technologies in Israel, the company that invented the computer firewall.

Brian Traynor
Traynor is an Associate Professor in Information Design. His research interests include Job Performance and Information Comprehension, Design Teaching Methods, and User Attribution of Blame. He spent 20 years in Telecom and IT Technical Communications before returning to an academic environment.