Rent a car in just 60, 120 or 240 seconds – Comparative Usability Measurement

Rolf Molich, DialogDesign (Denmark), molich@DialogDesign.dk
Tomer Sharon, Google Inc. (USA), tsharon@google.com
Jurek Kirakowski, University College Cork (Ireland), jzk@ucc.ie

Thursday 27 May 2010
10.30 to 12.00

Rent a car in just 60, 120 or 240 seconds – Comparative Usability Measurement

Rolf Molich, DialogDesign (Denmark), molich@DialogDesign.dk
Tomer Sharon, Google Inc. (USA), tsharon@google.com
Jurek Kirakowski, University College Cork (Ireland), jzk@ucc.ie

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 the usability of the car rental website Budget.com. The study documents a wide difference in approaches. Teams that used similar approaches often reached surprisingly similar results. The paper discusses a number of common pitfalls in usability measurement. The paper also points out a number of fundamental problems in unmoderated measurements, which were used by 6 of the 15 participating teams.

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 in (Molich, 2009).

Method

In May 2009, 15 US and European teams independently and simultaneously carried out usability measurements of the Budget.com website. The measurements were based on a common scenario and instructions (Molich, Kirakowski, Sauro, & Tullis, 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, and met for a full-day workshop at the UPA conference.

The main goal of CUE-8 was to gather information about the state-of-the-art in usability measurement. 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).

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

1. Rent a car
   Rent an intermediate size car at Logan Airport in Boston, Massachusetts, from Thursday 11 June 2009 at 09.00 am to Monday 15 June at 3.00 pm. If asked for a name, use John Smith, 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 pm to Sunday 21 June at 7.00 pm.

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. The repair will cost 2,000 USD. Your rental includes LDW, Loss Damage Waiver. 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. Legend: Questionnaire: A=ASQ, M=SMEQ, N=NASA TLX, O=Own, S=SUS, W=WAMMI.

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

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 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. Team 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 video 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.

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.

Key Measurement Results

Table 2. Reported key measurement results for task 1, Rent a Car. All times are in seconds.
Legend: 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.

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

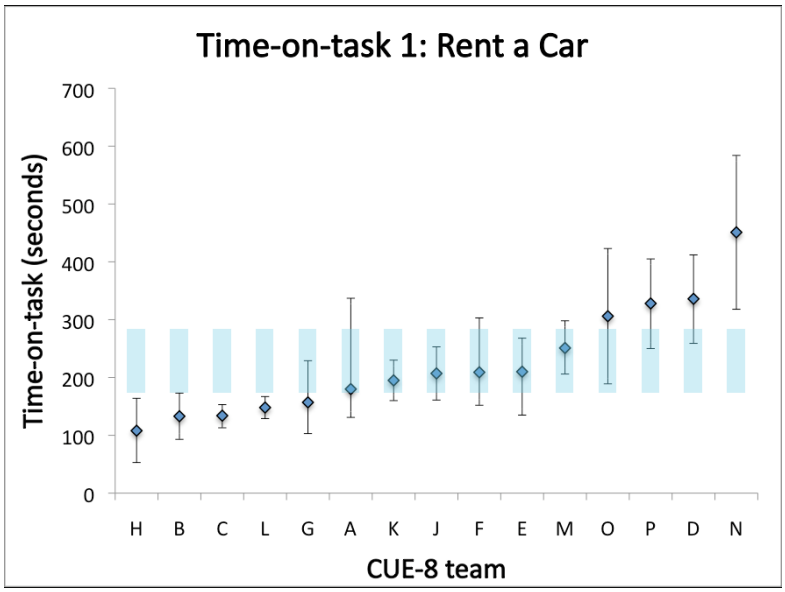

Figure 1. Reported time-on-task for task 1. The diamonds show the time-on-task reported by each team. The black vertical lines show the 95% confidence interval reported by the team before the workshop, or computed after the workshop for the teams that initially did not provide confidence intervals. The 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 since these teams chose to report the median as the central measure.

Eleven of the fifteen teams reported the mean (average) of their time-on-task measurements. Three teams (A, D, 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. Some of these teams indicated they are 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 methods used to compute the confidence intervals for time-on-task were:

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

Nine of the fifteen teams provided qualitative findings even though qualitative findings were not a subject of this study. 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.

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 five-scale form and provided scores by respondent.

Discussion

Computing Time-on-task

There is substantial agreement within the measurement community that measures such as time-on-task are not normally distributed since 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 are 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.

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. Since 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 which would "normalize" the distributions mathematically (Sauro, 2009).

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 1.

Eyeballing shows that the results from 6 teams (A, E, F, J, K and M) are in reasonable agreement for all five tasks. Two more teams (B and L) agree with the 6 teams for all tasks except task 1. Two teams (D and O) agree with the majority for three tasks. On the other hand, five teams mostly report diverging results. Team H and N consistently diverge from the other teams.

An analysis of the teams' approaches reveal 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.
- 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. 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 US.
- 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.

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 23 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".

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 are 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 2. 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.

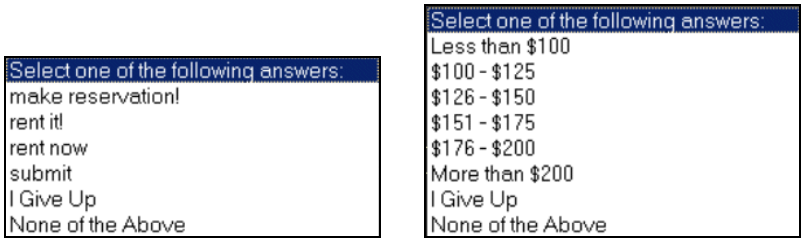

Figure 2. Multiple choice answers for task 1 and 2 used by team L for determining if their participants had obtained the right answer in unmoderated sessions. For task 1 participants were asked to find the label of the button that performed the rental; the correct answer is "rent it!". For task 2 the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

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 weakness 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.

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.

Practitioner's Takeaway

- Adhere strictly to precisely defined measurement procedures to get reproducible results.
- Report time-on-task, success/failure rate and satisfaction.
- Understand the inherent variability from samples and 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 insight regarding the 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 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

This paper is a short version of a paper that has been submitted to the Journal of Usability Studies. The authors gratefully acknowledge the contributions of the five additional authors of the JUS paper:
Janne Jul Jensen - Aalborg University (DK), Jarinee Chattratichart - Kingston University (UK),
Brian Traynor - Mount Royal University (CAN), Jeff Sauro - Oracle Corporation (US), and
Veronica D. Hinkle - Wichita State University (US).

CUE-8 depended entirely on the enthusiastic support from the 34 CUE-8 participants of which 19 attended the UPA 2010 workshop.

References

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

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

Sauro, J. (2009). Measuring Usability - Quantitative Usability, Statistics & Six Sigma by Jeff Sauro. Retrieved on September 11, 2009 from http://www.measuringusability.com/