This forum addresses conceptual, methodological, and professional issues that arise in the UX field's continuing effort to contribute robust information about users to product planning and design. — **David Siegel and Susan Dray, Editors**

Are Usability Evaluations Reproducible?

Rolf Molich, DialogDesign

Are we all doing the same thing when we do a usability test or an evaluation? Here's how to find out.

Ingredients:

• A state-of-the-art website that is intended for the general population, for example, a car-rental website

• 15 or so professional UX teams.

Directions:

• Ask the teams to evaluate the usability of the website independently and simultaneously.

• Compare the anonymous usability test reports from the teams in a one-day workshop where all teams participate.

• Marvel at the substantial differences in approach, reporting, and results.

• Repeat over 10-plus studies.

Since 1998, this has been the recipe for 10 successful Comparative Usability Evaluation (CUE) studies with more than 140 participating teams. These studies have produced unique insights into how experienced UX professionals do usability testing.

In a CUE study, teams simultaneously and independently evaluate the same product. All of the teams are given the same test scenario and objectives for the same interface, most often a website. Each team then conducts a study using their organization's standard procedures and techniques, for example, usability testing or heuristic evaluation. After each team has completed its study, it submits its results in the form of an anonymous report. In a subsequent one-day workshop, all participants meet and discuss the reports, the differences between them, the reasons for the differences, and how to improve the test process. The differences are often stunning.

Participation in a CUE study is mostly driven by curiosity and an eagerness to learn. Participation is mostly free except that participants are asked to cover direct expenses, such as room rental. Participating teams and I are not compensated financially, except that CUE-5 and CUE-6 were organized as for-profit workshops for which I received a fee.

Most of the teams have been from the U.S., but a considerable number of German, English, and Danish teams have also participated. Most of the anonymous CUE-test reports are freely available [1].

PURPOSES AND NON-PURPOSES

The key purposes of all CUE studies have been:

• to survey the state of the art within professional usability testing of websites

• to investigate the reproducibility of usability test results

Insights

→ The total number of usability issues in modern, complex websites is much larger than you can hope to find in a usability test.

→ Five users are by far not enough to find 75 percent or even 25 percent of the usability problems in a complex website like a car-rental website.

• to allow participating experienced professionals to further increase their skills.

Early on, I also formulated some non-purposes:

• to pick a winner

• to make a profit.

OVERVIEW OF CUE-STUDIES

Table 1 provides an overview of the CUE studies.

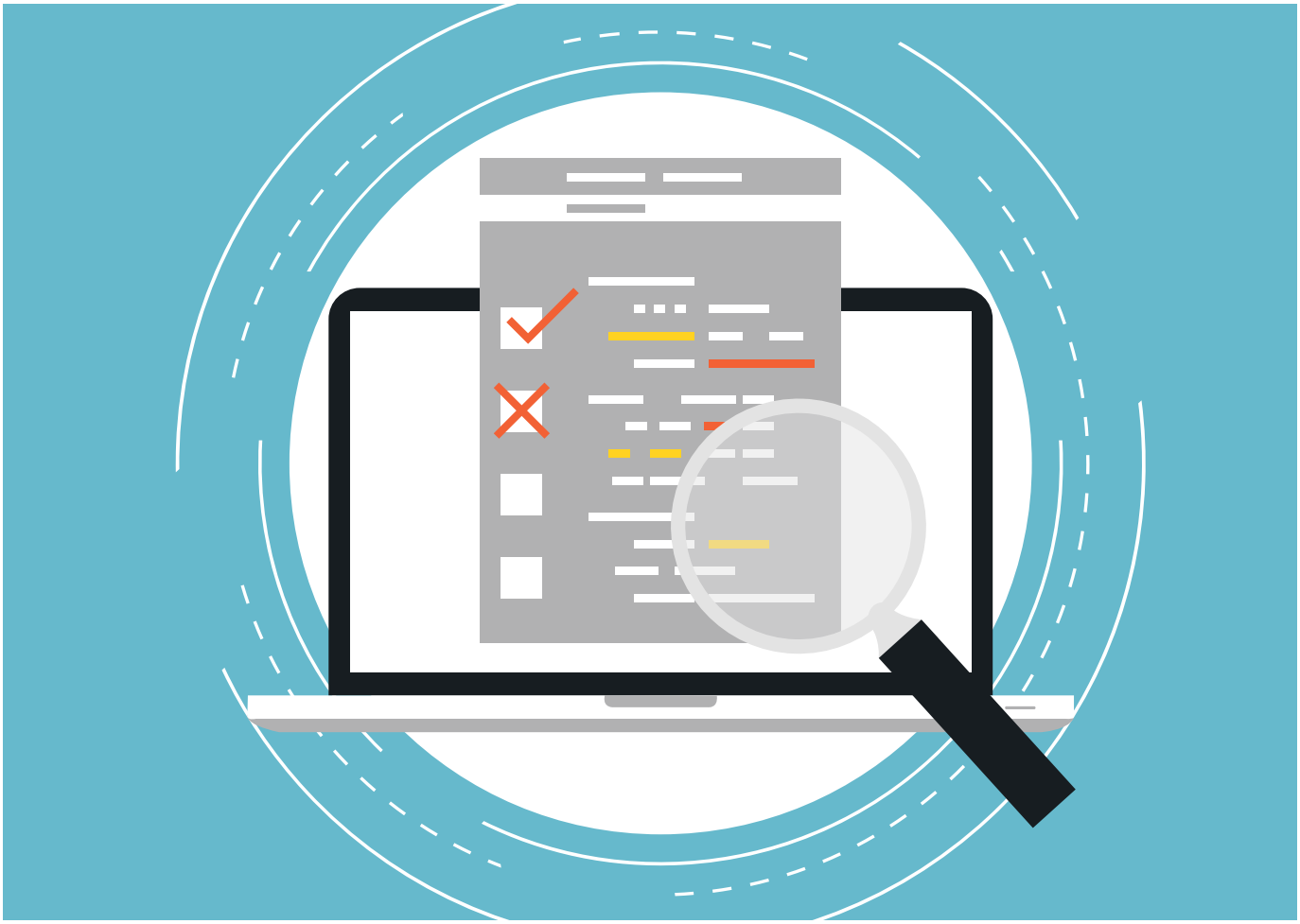
CUE-1 to CUE-6 were classic CUE studies; they all studied usability evaluation methods, in particular usability test, expert review, and heuristic inspection.

In CUE-7, six experts made recommendations for fixing specified real usability problems on IKEA's website. The purpose was to derive a set of usable recommendations for writing recommendations [3].

CUE-8 focused on the measurement of usability, in particular task time. In this study, the usability test tasks were prescribed [4].

CUE-9 focused on the evaluator effect. Thirty-five participants in the U.S. and in Germany independently analyzed the same five videos of unmoderated usability test sessions of a truck-rental website. In this study, all test tasks and test participants were the same. The study confirmed the evaluator effect: Different moderators report somewhat different issues even though they watched the same videos [5].

CUE-10 studied test moderation. Sixteen moderators, mostly professionals with considerable experience, video-recorded themselves and their test participants during



usability tests of an airline website. The purpose was to compare approaches to test moderation.

KEY FINDINGS

An overview of the key findings from the CUE studies is shown in Table 2. The following sections discuss these findings in more detail.

Huge number of issues. The total number of usability issues for the state-of-the-art websites that we have tested is huge. It is much larger than you can hope to find in one usability test. For example:

- CUE-2 found 310 usability problems in Hotmail in 2001.
- CUE-4 found 340 usability problems on a hotel website in 2003.
- CUE-9 found 223 usability problems on a van-rental website in 2011.

The figures in Table 3 indicate that all studies were far from finding all usability issues. In particular, the substantial number of issues that were reported by single teams only indicate that the 310, 340, and 223 usability issues may just be the tips of the icebergs.

If we had asked further teams to participate in any of these studies, or if the participating teams had run additional test sessions with different test tasks, additional problems would have been found.

Takeaway: Never claim or assume that you can carry out an exhaustive usability test of a website. Exhaustive testing may be possible within limited function areas.

Five users are not enough. It is a widespread myth that five users are enough to find 85 percent of the usability problems in a product. As shown in Table 3, the CUE studies have consistently shown that even 15 or more professional teams report only a fraction of the problems.

Never claim or assume that you can carry out an exhaustive usability test of a website.

If you vary the task set, the moderator, or the usability test procedure, new problems will be found. Many serious or critical problems cannot be discovered by a particular moderator or by a particular task set [6].

Takeaway: Five users—or 20 users, or even 100 users—will find only a small fraction of the usability problems. The last row in Table 3 shows that they will not even find all serious or critical problems. Nothing in the CUE studies contradicts, however, that five users may be enough to drive a useful iterative process.

No gold standard. CUE results consistently show that usability testing is not the expensive, high-quality gold standard against which all other methods can be measured.

Takeaway: Use usability testing as

IMAGE BY BLOSSOM STAR / SHUTTERTOCK.COM

Study	# Teams	Time	What each participating team did
CUE-1	4	March 1998	Usability test of Task Timer for Windows
CUE-2	9	December 1998	Usability test of Hotmail.com (7 professional teams, 2 student teams)
CUE-3	11	September 2001	Usability inspection of Avis.com
CUE-4	17	March 2003	Evaluated the usability of HotelPenn.com [2]
CUE-5	13	August 2005	Evaluated the usability of IKEA's online wardrobe planner
CUE-6	13	October 2006	Evaluated the usability of Enterprise.com
CUE-7	8	March 2007	Made recommendations for fixing six usability problems on IKEA.com [3]
CUE-8	15	June 2009	Measured the usability of key tasks on Enterprise.com [4]
CUE-9a	17	June 2011	Analyzed five videos of usability tests of UHaul.com – Atlanta US [5]
CUE-9b	18	August 2011	Analyzed five videos of usability tests of UHaul.com – Chemnitz DE [5]
CUE-10	16	May 2018	Moderated three test sessions of Ryanair.com; sessions were video recorded

→ Table 1. An overview of the 10 CUE studies that have been conducted to this point.

Huge number of issues	The total number of usability issues for a modern website is huge, 300 or more.
Five users are not enough	Five users are not enough to find 75% or even 25% of the usability problems on a website.
No gold standard	Usability testing is not the perfect method against which all other methods can be measured.
Expert reviews are useful	Expert reviews produce results of a quality comparable to usability tests.
Unusable test reports	The quality of the usability test reports varied dramatically.
Task design	There was virtually no overlap between tasks used by different teams.
Few false alarms	Almost all reported issues were valid.

→ Table 2. Key findings from CUE studies.

Study	CUE-2	CUE-4	CUE-9
Number of participating teams	9	17	35
Total number of issues = problems + positive findings	310	340	223
Issues reported by			
– All participating teams	0	0	0
– More than 75% of the teams	1	3	4
– 40% to 75% of the teams	10	17	18
– At least 3 teams but less than 40% of the teams	17	64	76
– 2 teams	50	51	35
– Single teams only	232	205	90
	75%	60%	40%
Serious or critical problems reported by single teams only	29	61	17

→ Table 3. Number of issues reported by one or more teams.

part of a market basket of evaluation methodologies.

Expert reviews are useful. CUE-4 indicated that expert reviews produce results of a quality comparable to usability tests—at least when carried out by experts.

Most issues reported by usability testing were also reported by expert reviews and vice versa. Few false problems were identified. As expected, expert reviews seem to require slightly fewer resources than usability tests.

CUE-3 indicated that professionals with limited experience may have problems using expert reviews. This is one of the few places in the CUE studies where we saw false alarms.

Takeaway: Use expert reviews when usability and subject matter experts are available and resources are scarce. I would like to add a personal warning that was not tested in the CUE-studies: Use expert reviews with great care in organizations that have a low usability maturity.

Unusable test reports. The quality

of the usability test reports varied dramatically. In CUE-2, the size of the nine reports varied from five pages to 52 pages—a 10-times difference! Some reports lacked positive findings, executive summaries, and screenshots. Others were complete with detailed descriptions of the team's methods and definitions of terminology. By looking through the different reports, we can quickly pick out the attributes that would make our reports more helpful to our clients.

Takeaway: Make your usability test reports usable for the target audience: management and developers.

- Limit yourself to at most 25 pages, possibly by leaving out some of the less important issues. Remember: You haven't found all the issues anyway.
- Include a one-page executive summary and place it at the beginning of the report.
- Include positive findings.
- Include screenshots, possibly with callouts, to make the report more informative and attractive for its users.

Task design. In CUE-2, nine teams created 51 different tasks for the same UI. We found each task to be well designed and valid, but there was scant agreement on which tasks were critical. If each team used the same best practices, then they should have derived similar tasks from the test scenario. But that isn't what happened. Instead, there was virtually no overlap. It was as if each team thought the interface was for a completely different purpose.

Takeaway: During task design, focus on key tasks rather than secondary tasks, however interesting they may be.

Few false alarms. We rigorously evaluated each reported issue. We paid particular attention to problems reported by single teams only. We found almost all described problems to be reasonable and in accordance with generally accepted advice on usable design. The only exception was CUE-3, where a few of the less experienced teams reported questionable issues.

Takeaway: If you have some usability experience and report issues based on experience and observation, your reported problems should be reliable.

DISCUSSION

Some have raised criticism of the CUE studies. For example:

- "A matching process was used to identify the usability findings that reported the same usability issues. This process involved judgment and may be unreliable."

While several experts reached consensus about the matching of all reported findings, we acknowledge that others may group the findings differently and that this may affect the results—but not to the extent that it will affect the general conclusions.

- "Participating teams had too much freedom. You should have prescribed more details of study, for example, the test script and the test tasks. This would have made results much more uniform."

We deliberately decided not to tell some of the world's most recognized usability-test experts how to do a usability test. This helped us gain additional insights. Also, in CUE-9 all participating teams watched the same videos but still reported different results.

- "Participants did not get paid. The results of commercial studies would have been much better."

This does not match our experience; also, none of our participants agreed that this would have made a significant difference.

CONCLUSION

The CUE studies raise some central questions for the future research of usability-testing techniques. How can we construct tests that find the important usability problems as quickly as possible? And how can we improve our practices so different teams will consistently find the same problems?

The practices of all of the teams in the CUE studies needed review, formalization, and a general tightening up. In all probability, since the teams were professional, everyone can benefit from reviewing the practices. We can use this analysis to hold a mirror up to our own work. These long-overdue experiments provide valuable material for sharpening individual usability practices. The CUE participants have done a great job of opening our eyes to the possibilities for improvement.

Usability evaluation should no longer be an artistic activity where freedom prevails. To reap the maximum benefit from usability evaluations, standard procedures should be defined and enforced, just like doctors are obligated to use standard procedures for most treatments.

The CUE studies show that our simple assumption that we are all doing the same thing when we do a test or an evaluation is incorrect.

ACKNOWLEDGMENTS

Special thanks are due to the more than 140 usability specialists who have participated in the CUE studies over the past 20 years. Many participants enjoyed the experience so much that they participated in several studies.

ENDNOTES

1. http://www.dialogdesign.dk/ CUE.html
2. Molich, R. and Dumas, J.S. Comparative usability evaluation – CUE-4. *Behaviour & Information Technology 27*, 3 (2008), 263–281.
3. Molich, R., Hornbæk, K., Krug, S., Scott, J., and Johnson, J. Recommendations on recommendations. *User Experiencee 7*, 4 (2008), 26–30.
4. Molich, R., Chattratichart, J., Hinkle, V., Jensen, J.J., Kirakowski, J., Sauro, J., Sharon, T., and Traynor, B. Rent a car in just 0, 60, 240 or 1,217 seconds? – Comparative usability measurement, CUE-8. *Journal of Usability Studies 6*, 1 (2010), 8–24.
5. Hertzum, M., Molich, R., and Jacobsen, N.E. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology 33*, 2 (2013), 143–161.
6. Lindgaard, G. and Chattratichart, J. Usability testing: What have we overlooked? *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 2007, 1415–1424.

Rolf Molich manages DialogDesign, a tiny Danish usability consultancy. In 2014, he received the UXPA Lifetime Achievement Award for his work on the Comparative Usability Evaluation project. He is vice president of the UXQB, which develops and maintains the CPUX certification. He is also the co-inventor of the heuristic evaluation method.

→ molich@dialogdesign.dk

DOI: 10.1145/3278154