



A Critique of "How To Specify the Participant Group Size for Usability Studies: A Practitioner's Guide" by Macefield

Rolf Molich

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

Abstract

This critique addresses important issues overlooked by Macefield in his recent paper in JUS. The participant group size depends on the purpose of the test. It also depends on test quality—if poor methodology is used, participant group size is unimportant. Evaluator quality and number of involved evaluators affect usability studies more than participant group size. The Comparative Usability Evaluation (CUE) studies have shown that it is infeasible or impossible to find all serious usability problems on a typical website. This critique argues that five users are enough to drive a useful iterative cycle.

Keywords

Usability study, participant group size, test sample size

Introduction

Today, no usability conference seems to be complete without one or more heated debates on participant group sizes for usability studies (Bevan, 2003; Molich, Bachmann, Biesterfeldt, & Quesenbery, 2010). In this respect, Macefield's recent article on group sizes for usability tests is timely (Macefield, 2009). However, when reading Macefield's article I found that a number of important issues were not addressed. With this critique I would like to draw the readers' attention to some of these issues.

Relevant Real-World Data

We must base discussions on topics such as the number of test participants on relevant real-world data. We should not base them only on analyses of simple systems such as Faulkner (2003) and Woolrych and Cockton (2001) or on studies conducted with undergraduates, such as Woolrych and Cockton's study (2001), unless, of course, it can be proven that undergraduates perform in a way that is comparable to usability professionals. My personal experience from teaching introductory usability engineering classes at the Technical University of Denmark is that only the best 10-20% of my students hand in reports that are at a professional level.

In 1998, I started a series of Comparative Usability Evaluations (CUE) to provide real-world data about how usability testing is carried out in practice. 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, report their evaluation results anonymously, and discuss their results at a workshop. Usually, 12-17 professional teams participate. Teams conduct their studies independently and in parallel using their favorite evaluation approach, which is most often "think aloud" usability testing or expert review.

CUE-1 to CUE-6 focused mainly on qualitative usability evaluation methods, such as think-aloud testing, expert reviews, and heuristic inspections. CUE-7 focused on usability recommendations. CUE-8 focused on usability measurement. An overview of the eight CUE studies and their results is available at <http://www.dialogdesign.dk/cue.html> (Molich, 2010).

The CUE-2, CUE-4, CUE-5, and CUE-6 studies reported remarkably similar overall results. As an example, consider the following key results from the CUE-4 study (Molich & Dumas, 2008), where 17 teams analyzed the usability of the website for the Hotel Pennsylvania in New York:

- Sixty percent of the issues (205 issues out of 340) were reported by single teams only.
- Forty serious and twenty-one critical issues were reported by single teams only.
- No issues were reported by all teams. The best overlap was that two issues were reported by 15 of the 17 teams.
- The whole study identified 28 critical problems, 89 serious problems, and 13 bugs. Even if a team had ingeniously identified all these 130 important problem issues, reporting them in a usable way would be difficult. For websites of the same size as the Hotel Pennsylvania website or larger, it appears almost inevitable that many important issues will not be reported.
- The maximum overlap was 30 issues. It occurred between team H, which tested 12 participants and reported 67 issues, and team M, which tested 15 participants and reported 56 issues.
- The minimum overlap was one issue. Team J, which tested 7 participants and reported 19 issues, had only one issue in common with team S, which tested 6 participants and reported 40 issues.
- The limited overlap could be interpreted as a sign that some of the teams, such as team J, had conducted a poor evaluation. Our interpretation, however, is that the usability problem space is so huge that it inevitably leads to some instances of limited overlap.

The Total Number of Usability Issues Is Close to Infinite

The CUE studies show that it is impossible—or at least infeasible—to find all usability issues in a realistic website or product because the number is huge, most likely in the thousands. This has important implications for discussions of participant group size. Because you can't find all problems anyway, go for a small number of participants and use them to drive a useful iterative cycle where you pick the low-hanging fruit in each cycle.

It could be argued that most industry usability tests are conducted by one team. While the CUE studies show that one team finds only a small fraction of the problems, they probably find most of the problems *that the one team will find*. A discussion of participant group size is not complete without at least mentioning that varying the number of evaluators will affect results considerably and probably more than varying the participant group size.

The "infinite" number of usability issues also has important implications for many of the studies of participant group size because they assume that the total number of issues is known.

The Group Size Depends on the Purpose

A discussion of the participant group size only makes sense if the purpose of a usability test is known as shown in Table 1.

Table 1. Optimal participant group size for various purposes of a usability test based on the CUE studies and the author's experience.

| Main Purpose | Explanation | # participants |
|---------------------------|---|----------------|
| Political | Demonstrate to skeptical stakeholders that serious usability problems exist in their product. Demonstrate that usability testing can find some of them. | 2-3 |
| Find serious problems | Drive a useful iterative cycle: Find serious problems, correct them, find more serious problems. | 4-8 |
| Find all serious problems | Find all serious usability problems in a non-trivial product. | Unknown |
| Find all problems | Find all usability problems in a non-trivial product. | Unknown |
| Measure usability | Measure key usability parameters, for example time to complete key tasks and subjective user satisfaction. | >20 |

Table 1 says that the total number of participants required to find all usability issues or even all serious usability problems is unknown at this time ("infinite"). What we do know is that in four CUE studies that each involved 9-17 teams, at least 60% of the issues were reported by single teams only. We conjecture that if we had done additional testing with for example 20 more teams and hundreds of additional participants, several hundred additional issues would have been discovered and reported. Only a few of the reported issues were invalid, for example not reproducible or in conflict with commonly accepted usability definitions.

Macefield makes a similar point that group size should depend on the purpose of the test, for example in his Figure 1. But he fails to make the important points that tests can be run for political purposes and that finding all problems is infeasible for most real-world systems.

For usability measurements, the group size depends on the desired level of confidence. Computing the level of confidence is difficult because measurement results are not normally distributed (Sauro, 2010).

If Test Quality Is Poor, Group Size Doesn't Matter

If an evaluator uses poor test methodology, the results will be poor irrespective of the participant group size. Results of usability tests depend considerably on the evaluator (Jacobsen & Hertzum, 2001). This evaluator effect has been confirmed by the CUE studies. The CUE studies and the practical experience of this author show one of the reasons why the evaluator effect exists: Test quality is sometimes a problem. In other words, at least some usability tests are conducted with poor use of the "think aloud" methodology. Problems particularly arise in the following areas:

- Invalid test tasks: Most CUE teams used tasks with hidden clues. Other problems are: "humorous" or unrealistic tasks, task sets that do not cover key user tasks, and experienced users that are tested with tasks aimed at novices.
- Bad facilitation: Examples are that the evaluator provides hidden clues, the evaluator talks too much, and there is no debriefing.

The author's personal experience from certification of usability professionals in usability testing indicates that about 50% of professional evaluators make one or more severe errors in their test planning, facilitation, or reporting.

Conclusions and Practitioner's Takeaway

The following were the main findings in this article:

- Never pretend that you can find all usability problems—or even all serious usability problems—in a usability test.
- Use 2-3 participants if your main goal is political—to demonstrate to skeptical stakeholders that serious usability problems exist in their product and that usability testing can find some of them.
- Use 4-8 participants to drive a useful iterative cycle: Find serious problems, correct them, find more serious problems.
- The number of evaluators and their quality affects results considerably and probably more than participant group size.

References

- Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J.M., & Wixon, D.R. (2003). The "magic number 5": Is it enough for web testing? *CHI Extended Abstracts 2003*, 698–699.
- Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments and Computers*, 35(3), 379-383.
- Jacobsen, N. E., & Hertzum, M. (2001). The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13, 421–443.
- Macefield, R. (2009). How To Specify the Participant Group Size for Usability Studies: A Practitioner's Guide. *Journal of Usability Studies*, 5 (1), 34–45.
- Molich, R., & Dumas, J. (2008). Comparative Usability Evaluation (CUE-4). *Behaviour & Information Technology*, 27, 263–281.
- Molich, R. (2010). CUE – Comparative Usability Evaluation. Retrieved on May 3, 2010 from <http://www.dialogdesign.dk/cue.html>.
- Molich, R., Bachmann, K., Biesterfeldt, J., & Quesenbery, W. (2010). "Five users will find 85% of the usability problems" - and other myths about usability testing. *Proceedings of the Usability Professional's Association – International Conference 2010*, In press.
- Sauro, J. (2010). Measuring Usability - Quantitative Usability, Statistics & Six Sigma by Jeff Sauro. Retrieved on May 6, 2010 from <http://www.measuringusability.com/>.
- Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.) In *Proceedings of IHM-HCI 2001 Conference*: Vol. 2, (pp. 105- 108). Toulouse, France: Cépadèus.

About the Author



Rolf Molich

Rolf Molich owns and manages DialogDesign, a small Danish usability consultancy. Rolf conceived and coordinated the comparative usability evaluation studies CUE-1 to 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).