

The "Magic Number 5": Is It Enough for Web Testing?

Carol Barnum Southern Polytechnic State University, 1100 S. Marietta Pky, Marietta, GA 30060, USA

Nigel Bevan Serco Usability Services, 22 Hand Court, London, WC1V 6JF, UK. nevan@usability.serco.com

Gilbert Cockton University of Sunderland, PO Box 299, Sunderland, SR6 0YN, UK

Jakob Nielsen Nielsen Norman Group, 48921 Warm Springs Blvd., Fremont, CA 94539-7767, USA

Jared Spool User Interface Engineering, 242 Neck Road, Bradford, MA 01835, USA

Dennis Wixon Microsoft Corporation, One Microsoft Way, Redmond WA 98052, USA

ABSTRACT

Common practice holds that 80% of usability findings are discovered after five participants. Recent findings from web testing indicate that a much larger number of participants is required to get results and that independent teams testing the same web-based product do not replicate results. How many users are enough for web testing?

Keywords

Usability test, web,

OVERVIEW

When it comes to web usability testing, the magic number 5 is under attack. This number—5 participants will yield 80% of the usability findings—derives from research conducted by Nielsen [5], Virzi [8], and Lewis [2].

The challenge to this long-held and widely-used practice of conducting testing with an average of 5 users began with the results of a comparative usability evaluation [4], in which nine independent teams conducted usability evaluations of Hotmail, with no two teams finding the same set of problems. Of the 300 total problems found, 75% were reported by only one team. The challenge escalated with Jared Spool's report [7] on the results of web testing users purchasing CDs. Thinking it was safe to use the long-accepted number 5, Spool and his colleagues were surprised to see 247 problems identified by 18 different users, with major new findings being identified by each new user. Kessner et al. [1] found that six professional labs produced little overlapping findings in evaluating a UI.

Both Molich et al. [4] and Spool and Schroeder [7] conclude that web testing requires many more users to get reliable results. At the same time, Nielsen supports his original research for "comparable users who will be using the site in fairly similar ways" [5].

Woolrych and Cockton [9] assert that Nielsen's claim that "5 Users are Enough" [5] is based on a statistical formula [6] that makes unwarranted assumptions about individual differences in problem discovery, combined with optimistic setting of values for a key variable. They present evidence that the Landauer-Nielsen formula can fail to calculate an acceptable number of test users even for a small drawing editor, as the formula only holds for simple problem counts.

Medlock, Wixon et al at Microsoft recommend use of the RITE method [3] where changes to a product are made after having run only 1-3 participants. A survey found that 33 of 39 respondents had used a similar method of very rapid iterations and fixes at least once.

PANEL POSITIONS

Jakob Nielsen

First, let me clarify that I have always recommended 3-4 users as the optimal number for most usability studies. The reason for my more widely known recommendation to *plan* for 5 users is that this allows for one or even two no-shows while still getting to observe 3-4 users. Thus, 5 is the magic number during planning, whereas the magic number for the actual test is 3-4.

The standard recommendation to observe 3-4 users refers to studies conducted during a user-centered design process where an interaction design needs to be debugged for usability. The "discount usability" philosophy explicitly recognizes that this will not be a perfect study that will discover everything that's possible to know about the design, but we accept this trade-off in return for having more iterations in the design process and conserving resources for subsequent evaluations of these iterations.

The only situation where more than 5 users would be recommended for traditional user interface debugging would be a design with below-average usability personnel and bad project management. If the test facilitator is not very observant and if the project is slow to act on the findings, then it will be most cost-effective to test 10-15 users per iteration.

Gilbert Cockton

Five users can be enough some times. These are extremely rare. The view that 5 users could be enough was always naive. The statistical arguments behind it were ill informed [9]. Where people have tested more than 5 users, it is absolutely clear that, unless problems are very easy to find, not only are more problems found, but also that the profile of problems as regards frequency and severity changes radically with further users. The unpalatable fact is that one cannot predict in advance how many users are needed. Nor can one be specific about the risks of testing with 5, 10 or 15 users in advance of the evidence.

What we do understand are the variables that influence problem yield, for example, test user diversity, test protocol design, task performance diversity, application complexity, design quality (problems are easier to find for

Copyright is held by the author/owner(s).

CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.

ACM 1-58113-637-4/03/0004

poor designs!), problem reporting procedures and the usability goals set for a product (no goals, no problems!) However, we do not yet have a formula into which these variables can be plugged. When we do, the answer will rarely be 5.

We must fully share this reality with clients. Usability is about risk management. Risks diminish as we test more users. The break even on cost-benefit is product specific. For some, one user is enough, for others even 100 will be too few.

Carol Barnum

The problem that has arisen from the Molich and Spool studies is that the apparent conclusion—5 is nowhere near enough for web usability testing—has been widely disseminated, causing great alarm in the usability community. Questions have arisen as to whether it's worth testing if the results are as invalid as have been claimed by these two studies.

Our own web studies, conducted according to the approaches prescribed by the early researchers, and in close cooperation with the client/sponsor, suggest that very similar results can be duplicated by different test teams.

Five *is* enough for web testing:

- when the original discount model for testing is followed
- when the results of testing are understood and clearly communicated
- when there is close cooperation between the client/sponsor and the test team
- when the results are used for diagnostic purposes and team learning
- when the expected result is insight, not validation.

Dennis Wixon

The problem of determining how many users one must test in order to have a reasonable expectation of uncovering all the problems that users will have is one that has been the subject of much excellent research and theorizing.

Unfortunately when considered from the viewpoint of designing real products in the real world it is the WRONG PROBLEM TO STUDY.

The goal of most iterative tests in a commercial world is to produce the best possible product/design in the shortest time at the lowest possible cost with the least risk.

Consequently, design and usability teams are well advised to fix interface problems as rapidly as possible and continue testing and fixing until time runs out or a usability metric is achieved. Such an approach (we call it RITE, Rapid Iterative Testing and Evaluation) [3] concentrates statistical power where it belongs – verifying that fixes actually work for users.

While I agree with Gilbert that we must be honest with our clients, it is also critical that we understand their concerns and serve their needs.

Jared Spool

We have a crisis. And, this argument is in the dead center of it.

On one level, the notion that 'N' users is required for testing (whether 'N' be 5, 8, or some other reasonably small number) is an academic question. Design teams are limited by both time and resources, so they'll test as many users as they can within their constraints. They'll be done when they are done. For whatever value of 'N' they choose, 'N' users will always be better than zero.

But, in the grand scheme, there's a much larger and more insidious problem. Practitioners accept usability testing as the cornerstone of their user-centered design practices. Yet, here we have a panel of very smart, well-respected, world-renowned experts that can't come to any agreement on the basic elements of a quality testing protocol.

For years, the common belief was that Landauer and Nielsen had it right. They had produced a very simple formula which could be graphed in the much cited 'parabola of optimism', as we've come to call it. Because we found that simple software usability tests fit the curve, we believed all was well with the world. We knew what to do.

Understanding how many users for testing is more than just an academic calculation. It's the basis of faith in everything we're trying to do to make the world better. We need to come to agreement and we need to do it quickly.

REFERENCES

1. Kessner, M., Wood, J, Dillon, R.F., and West, R.L. On the reliability of usability testing. *CHI 2001*. Poster.
2. Lewis, J.R. Sample sizes for usability studies: Additional considerations. *Human Factors* 36, 368-378 (1994).
3. Medlock, M.C., Wixon, D., Terrano, M., Romero, R., Fulton, B. (2002). Using the RITE method to improve products: a definition and a case study. *Proc. Usability Professionals Association* (Orlando FL, July 2002).
4. Molich, R., et al. Comparative evaluation of usability tests. *CHI 99 Extended Abstracts*, ACM Press, 83-84.
5. Nielsen, J. Why you only need to test with 5 users. Alertbox (2000) www.useit.com/alertbox/2000319.html
6. Nielsen, J., and Landauer, T.K. A Mathematical Model of the Finding of Usability Problems, in *Proc INTERCHI '93*, 206-213
7. Spool, J., and Schroeder, W. Testing web sites: Five users is nowhere near enough. *CHI 2001 Extended Abstracts*, ACM Press, 285-286.
8. Virzi, R. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors* 34, 457-468 (1992).
9. Woolrych, A., and Cockton, G. Why and When Five Test Users aren't Enough, in *Proc. IHM-HCI Conference: Volume 2*, 105-108, 2001.