

How many testers are enough?

Dr. Sarah Burton-Taylor. Director – WUP Revised September 2006

Introduction

We are often asked 'how many testers do we need for usability testing?'. The answer is 'it depends' - are you trying to:

- Unearth all the usability issues on a site?
- Identify the 'big' usability issues that stop users having a good user experience?
- Do an academic or benchmarking study?
- Inform a complete site redesign?
- Develop a user centred organisational culture?
- Persuade a senior manager that there is a major problem with the site – which he denies!

It also depends on:

- The budget and time available - inevitably there are always constraints on both
- The diversity of the site's target audiences - the goals they wish to achieve, their knowledge and their experience
- The strategic importance of the web site - is it mission critical?

So the optimum solution is going to be a trade off between research objectives, time and money available, user variability and strategic importance.

Organisations often seem concerned about finding all the problems on a site. We would argue this is the wrong concern

How many usability problems?

Organisations often seem concerned about finding *all* the problems on a site. We would argue this is inappropriate:

- Firstly, it's very difficult to know when you've found 100% of the usability problems – an additional tester will always find another problem
- Secondly, do you have the resources to fix everything? Usually, organisational constraints limit the problems that can be fixed
- Thirdly, it is unrealistic to expect to fix everything: usability problems are often related - fixing one problem is likely to surface another one.

For the organisations we work with (large public and private sector) the more sensible question is, "How many testers do we need to find the 'big' usability issues that stop users having a good experience?"

Nielsen (1998) reports on research by Molich & Gram that rated usability issues by severity. They found an average of:

- 11 usability 'catastrophes' that prevented users from achieving their goals
 - 20 'serious' usability problems which interfered but did not prevent goal achievement
 - 29 cosmetic usability problems which irritated users but did not stop goal achievement.
-

20% of the problems account for 80% of the bad experience

This study suggests that 18% of the known usability issues identified resulted in a bad user experience. This fits with our experience of undertaking user experience research on hundreds of websites: typically a relatively small proportion of usability problems account for the bulk of the bad experience. Whilst it is difficult to quantify '100%' of the problems, and how 'bad' a bad experience is, 'Pareto's theory' seems to operate - 20% of the problems cause 80% of the bad experience. So finding and fixing the big problems will have a massive effect on the user experience.

So, what do others say?

Virzi (1992) found that c.80% of known usability problems could be surfaced with 5 testers, and that 3 testers would reveal the most severe problems. Nielsen and Landauer (1993) also suggested that 5 users will surface c.80% of known usability problems, and that 3 testers will reveal nearly 70% of these problems. Nielsen (2000) indicates that there is a law of diminishing returns – *"the third tester will do many things that you have already observed with the first or second user...[and] generate a small amount of new data...after the fifth user you are wasting your time by observing the same findings repeatedly but not learning much new"*.

This is certainly what we find - virtually all the 'big' problems surface in the first 3 or 4 testers. Nielsen's work suggests that 15 users will reveal all the known usability problems in a design, but recommends a more effective spend of three iterative tests with 5 users.

Perfetti & Landesman (2001) suggest that more than 8 testers are needed to detect all usability issues. But they do not suggest doing these all at once; rather, they advocate a programme of ongoing iterative testing *"bringing in a user or two every week"*.

Woolrych & Cockton (2001) have undertaken a heuristic evaluation of Landauer & Nielsen's formula, which suggests that 3 or 5 testers are adequate where the variance for individual testers is low, but that more are needed to find all usability problems where tester variability is high – however, they don't give a number! A recent feature on 'How many testers are really enough?' in the User Experience magazine reviewed the question from the perspective of various authors and studies, and again concluded that using small groups of representative users iteratively offered much better value, in all senses, to clients than large groups of testers. Rolf Molich, in an article for UI 11 2006 Conference also states that "*the number of users needed for web-testing depends on the goal of the test. If you have no goal, then anything (including nothing) will do*". He suggests 3-4 users to "*sell usability*" into the organisation, 5-6 users to find "*catastrophic problems*" in an iterative process, and over 50 users to find all usability problems in an interface. Faulkner (2003) has established that the variability of the testers can have enormous impact on the number of usability issues discovered: she found that different groups of 5 testers surfaced anything from 55% to 99% of 'known' usability problems. She concludes "*test users must be representative of the target population. The important and often complex issue, then, becomes defining the target population*". This is something we believe is absolutely critical i.e. defining the target audience and then recruiting and testing accordingly.

So, how many testers then?

As we said at the beginning it's a trade off between the research objectives, time and money available, user variability and strategic importance. Quality, not quantity, is the issue – the quality of the evaluation, its results, and the use to which those results are put. Better to test smaller numbers iteratively and fix the problems in between, rather than test large numbers at any one time. Often one large problem will dominate and prevent testers progressing through the site – with large sample sizes clients will simply see testers fail repeatedly on the same task!

Research without action is a waste

Research without action is a waste: the outcomes need to have full commitment from the key organisational stakeholders in order to result in action, and our whole approach underpins this philosophy. We're not the only people who think this: Mark Hurst (2003) writes "*Changing the organization is the most difficult and most important part of user experience work...user experience only matters if it has real world results*".

We, therefore, encourage clients to actively observe three testers in order to get experiential knowledge of the user experience: we may have done additional unobserved testing in advance, but we find that what they see has the biggest impact on the client stakeholders – not what they read in a report. After the observed testing, on the same day, we facilitate a meeting with the client to drive out actions that they agree are relevant and actionable, and are framed within their constraints. Clients never say that they haven't found enough usability problems (on the contrary!), and they gain significant insight into the user experience, and are committed to take action to make relevant improvements.

Even one tester can make a difference

Even one tester can make a difference. Many of our clients have brought in senior managers, content producers or marketing communications staff to see just one tester, and have found that making them 'walk in the user's shoes' can have a profound effect on their understanding of what makes an effective website.

Further details of all our services can be found at: www.wupltd.co.uk.

© The Web Usability Partnership Ltd
Unit 15, Lansdowne Court, Bumpers Farm,
Chippenham, Wilts SN14 6RZ
Tel: 01249 444 757
Email: info@wupltd.co.uk
Web: www.wupltd.co.uk

References

- Virzi, R.A., (1992), Refining the test phase of usability evaluation: how many subjects is enough?
- Landauer T.K. & Nielsen J. (1993), A mathematical model of the findings of usability problems
- Nielsen, J. (1998), cost of user testing a website
- Nielsen, J. (2000), Why you only need to test with 5 users
- Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough
- Perfetti, C. & Landesman, L. (2002), Eight is not enough
- Faulkner, L. (2003), Beyond the five-user assumption: benefits of increased sample sizes in usability testing
- Hurst, M., (2003), Organizational dynamics: The most important user experience method
- Usability Professional Association - User Experience Magazine Volume 4, Issue 4, 2005 "How many users are really enough?"
- 'Usability Testing Best Practices: An Interview with Rolf Molich' – UI 11 2006 Conference Articles (www.uie.com)