Measuring
the User Experience
Collecting, Analyzing, and Presenting Usability Metrics

Second Edition

Tom Tullis
Bill Albert

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO
SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

careful participant recruitment, investing in wide task coverage is more fruitful than increasing the number of users."

One technique that works well to increase task coverage in a usability test is to define a set of tasks that all participants must complete and another set that is derived for each participant. These additional tasks might be selected based on characteristics of the participant (e.g., an existing customer or a prospect) or might be selected at random. Care must be exercised when making comparisons across participants, as not all participants had the same tasks. In this situation, you may want to limit certain analyses to the core tasks.

THE SPECIAL CASE OF MODERATOR BIAS IN AN EYE-TRACKING STUDY

One of the more difficult aspects of moderating a usability study is controlling where you look during the session. Moderators usually are looking either at the participant or their interaction on a screen, or some other interface. This works well, except in the case of an eye-tracking study. Most eye-tracking studies measure where participants look, and whether participants are noting key elements on the interface. As a moderator, it can be difficult *not* to look at the target when the participant is scanning the interface. Participants can pick up on this easily, begin to notice where *you* are looking, and use that information as a guide to target. It happens very quickly and subtly. While this behavior has not been reported in the user experience literature, we have observed it during our own eye-tracking studies. The best thing to do is to be aware of it, and if you find your eyes starting to wander to the target, simply refocus on the participant, what they are doing, or, if you have to, some other element on the page. Or don't sit in the room with the participant, if that's an option. When you're sitting with the participant in an eye-tracking study, there's also a greater chance that the participant will naturally look at you and away from the screen.

5.7 NUMBER OF PARTICIPANTS

There has been much debate about how many participants are needed in a usability test to reliably identify usability issues. [For a summary of the debate, see Barnum et al. (2003).] Nearly every UX professional seems to have an opinion. Not only are many different opinions floating around out there, but quite a few compelling studies have been conducted on this very topic. From this research, two different camps have emerged: those who believe that five participants are enough to identify most of the usability issues and those who believe that five is nowhere near enough.

5.7.1 Five Participants is Enough

One camp believes that a majority, or about 80%, of usability issues will be observed with the first five participants (Lewis, 1994; Nielsen & Landauer, 1993;

Virzi, 1992). This is known as the "magic number 5." One of the most important ways to figure out how many participants are needed in a usability test is to measure *p*, or the probability of a usability issue being detected by a single test participant. It's important to note that this *p* is different from the *p* value used in tests of significance. The probabilities vary from study to study, but they tend to average around 0.3, or 30%. [For a review of different studies, see Turner, Nielsen, and Lewis (2002).] In a seminal paper, Nielsen and Landauer (1993) found an average probability of 31% based on 11 different studies. This basically means that with each participant, about 31% of the usability problems are being observed.

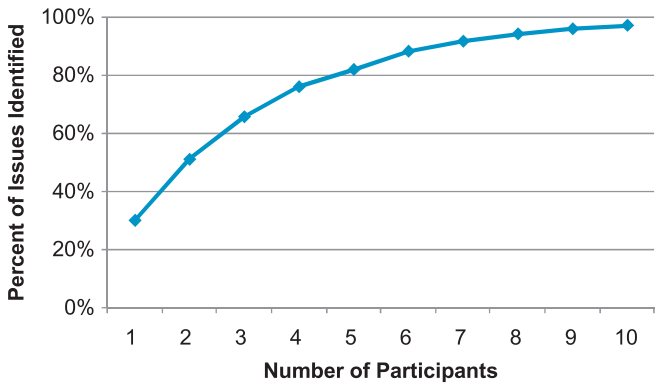

Figure 5.7 Example showing how many users are required to observe the total number of issues in a usability study, given a probability of detection.

Figure 5.7 shows how many issues are observed as a function of the number of participants when the probability of detection is 30%. (Note that this assumes all issues have an equal probability of detection, which may be a big assumption.) As you can see, after the first participant, 30% of the problems are detected; after the third participant, about 66% of the problems are observed; and after the fifth participant, about 83% of the problems have been identified. This claim is backed up not only by this mathematical formula, but by a\necdotal evidence as well. Many UX professionals only test with five or six participants during an iterative design process. In this situation, it is relatively uncommon to test with more than a dozen, with a few exceptions. If the scope of the product is particularly large or if there are distinctly different audiences, then a strong case can be made for testing with more than five participants.

CALCULATING *P*, OR PROBABILITY OF DETECTION

Calculating the probability of detection is fairly straightforward. Simply line up all the usability issues discovered during the test. Then, for each participant, mark how many of the issues were observed with that participant. Add the total number of issues identified with each participant and then divide by the total number of issues. Each test participant will have encountered anywhere from 0 to 100% of the issues. Then, take the average for all the test participants. This is the overall probability rate for the test. Consider the example shown in this table.

Participant	Issue 1	Issue 2	Issue 3	Issue 4	Issue 5	Issue 6	Issue 7	Issue 8	Issue 9	Issue 10	Proportion
P1	X		X		X		X	X	X		0.6
P2	X	X		X		X					0.4
P3			X			X		X	X	X	0.5
P4	X	X			X			X	X	X	0.6
P5				X	X		X				0.3
P6	X					X		X	X		0.4
P7		X	X		X			X	X	X	0.6
P8	X		X	X	X		X	X		X	0.7
P9		X	X	X		X	X		X		0.6
P10		X			X						0.2
Proportion	**0.5**	**0.5**	**0.5**	**0.4**	**0.6**	**0.4**	**0.4**	**0.6**	**0.6**	**0.4**	**0.49**

Once the average proportion has been determined (0.49 in this case), the next step is to calculate how many users are needed to identify a certain percentage of issues. Use the following formula:

$$1 - (1 - p)^n,$$

where n is the number of users.

So if you want to know the proportion of issues that would be identified by a sample of three users:

- $1-(1-0.49)^3$
- $1-(0.51)^3$
- $1-0.133$
- 0.867, or about 87%, of the issues would be identified with a sample of three users from this study

5.7.2 Five Participants is *Not* Enough

Other researchers have challenged this idea of the magic number 5 (Molich et al., 1998; Spool & Schroeder, 2001; Woolrych & Cockton, 2001). Spool and Schroeder (2001) asked participants to purchase various types of products, such as CDs and DVDs, at three different electronics websites. They discovered only 35% of the usability issues after the first five participants—far lower than the

80% predicted by Nielsen (2000). However, in this study the scope of the websites being evaluated was very large, even though the task of buying something was very well defined. Woolrych and Cockton (2001) discount the assertion that five participants are enough, primarily because it does not take into account individual differences.

The analyses by Lindgaard and Chattratichart (2007) of the nine usability tests from CUE-4 also raise doubts about the magic number 5. They compared the results of two teams, A and H, that both did very well, uncovering 42 and 43%, respectively, of the full set of usability problems. Team A used only 6 participants, whereas Team H used 12. At first glance, this might be seen as evidence for the magic number 5, as a team that tested only 6 participants uncovered as many problems as a team that tested 12. But a more detailed analysis reveals a different conclusion. In looking specifically at the overlap of usability issues between just these two reports, they found only 28% in common. More than 70% of the problems were uncovered by only one of the two teams, ruling out the possibility of the five-participant rule applying in this case.

THE EVALUATOR EFFECT

The Evaluator Effect (Hornbaek & Frokjaer, 2008; Jacobson, Hertzum, & John, 1998; Vermeern, van Kesteren, & Bekker, 2003) in usability testing suggests that UX professionals identify a different set of usability issues. In other words, there is little agreement or overlap in the usability issues identified by different UX professionals. The evaluator effect has been observed consistently in the CUE studies led by Rolf Molich (http://www.dialogdesign.dk/CUE.html). Most recently, CUE-9 (Molich, 2011) focused on the Evaluator Effect. Most of the 34 test team leaders in CUE-9 were confident that they found the most significant usability issues. However, there was little overlap in the issues. Furthermore, the test teams felt that running more participants would not help them identify more usability issues.

How do we reconcile this finding in the context of the recommended number of participants? It is easy for a UX professional to say that he found most of the usability issues after testing with 5–10 participants. In fact, they are usually very confident. But how do they really know unless they compare their findings to another UX professional? The fact is, they don't. It is quite possible that additional usability issues, often significant, may be uncovered with an independent assessment by another UX professional.

5.7.3 *Our* Recommendation

We recommend maintaining flexibility regarding sample sizes in usability tests. We feel it may be acceptable to test with 5–10 participants, using one UX team when the following conditions are met:

- It is OK to miss some of the major usability issues. You are more interested in capturing some of the big issues, iterating the design, and retesting. Any improvement is welcome.
- There is only one distinct user group that you believe will think about the design and tasks in a very similar way.
- The scope of the design is limited. There are a small number of screens and/or tasks.

We recommend increasing the number of participants and/or the number of UX teams when the following conditions apply:

- You must capture as many UX issues as possible. In other words, there will be significant negative repercussions if you miss any of the major usability issues.
- There is more than one distinct user group.
- The scope of the design is large. In this case we would recommend a broader set of tasks.

We fully realize that not everyone has access to multiple UX researchers. In this case, try to solicit feedback from any other observers. No one can see everything. Also, you might want to acknowledge that some of the major usability issues might not have been identified.

5.8 SUMMARY

Many usability professionals make their living by identifying usability issues and by providing actionable recommendations for improvement. Providing metrics around usability issues is not commonly done, but it can be incorporated easily into anyone's routine. Measuring usability issues helps you answer some fundamental questions about how good (or bad) the design is, how it is changing with each design iteration, and where to focus resources to remedy the outstanding problems. You should keep the following points in mind when identifying, measuring, and presenting usability issues.

1. The easiest way to identify usability issues is during an in-person lab study, but it can also be done using verbatim comments in an automated study. The more you understand the domain, the easier it will be to spot the issues. Having multiple observers is very helpful in identifying issues.
2. When trying to figure out whether an issue is real, ask yourself whether there is a consistent story behind the user's thought process and behavior. If the story is reasonable, then the issue is likely to be real.
3. The severity of an issue can be determined in several ways. Severity always should take into account the impact on the user experience. Additional factors, such as frequency of use, impact on the business, and persistence, may also be considered. Some severity ratings are based on a simple high/medium/low rating system. Other systems are number based.
4. Some common ways to measure usability issues are measuring the frequency of unique issues, the percentage of participants who experience

a specific issue, and the frequency of issues for different tasks or categories of issue. Additional analysis can be performed on high-severity issues or on how issues change from one design iteration to another.

5. When identifying usability issues, questions about consistency and bias may arise. Bias can come from many sources, and there can be a general lack of agreement on what constitutes an issue. Therefore, it's important to work collaboratively as a team, focusing on high-priority issues, and to understand how different sources of bias impact conclusions. Maximizing task coverage and including an additional UX team may be key.