

The Science of Usability Testing

Jean E. Fox
Bureau of Labor Statistics

Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

Usability testing has evolved from the stringent methods of experimental psychology, to less controlled, more qualitative tests, to the wide variety of methods used today. As the methods have evolved, researchers have studied many aspects of usability testing with the goal of better understanding how to best implement, plan, conduct, and interpret tests. In this paper, I will discuss findings from some of the research on three aspects of usability testing: the number of participants, the number of trained observers, and the use of the think-aloud method.

Key Words: Usability Testing, Think Aloud

Introduction

User experience professionals use many methods to assist with the research, design, and evaluation of all kinds of products and systems. This paper focuses on usability testing, just one of the many user experience methods (examples of other methods include focus groups, card sorting, and expert reviews). The goal of this paper is to provide research-based guidance for planning and conducting usability tests.

In a usability test, typical users perform tasks typical for that user group in a typical environment with a website, system, or any other kind of product with a user interface. Government statistical agencies might test their data collection instruments (including paper, telephone, and web collection), their internal data analysis systems, as well as their public data dissemination websites. During a usability test, test administrators collect qualitative data such as observations and participant comments, and quantitative data such as task times and success rates. By analyzing these data, the test administrators can identify usability problems with the product or system, then recommend improvements to address these problems.

Usability testing has evolved from the rigorous methods of experimental psychology, to less controlled, more qualitative tests, to the wide variety of methods used today. As the methods have evolved, researchers have studied many aspects of usability testing with the goal of better understanding how to best plan, conduct, and interpret tests. This research helps practitioners design and conduct more efficient and effective usability tests. In addition, the research provides insight that is helpful for interpreting and presenting the usability test results.

There has been quite a bit of research on many aspects of usability testing, but in this paper I will focus on three issues:

- Sample size: How many participants do you need?
- Evaluator effect: How many trained observers do you need?
- Think aloud: How does thinking aloud affect participants' performance?

Researchers have studied other aspects of usability testing as well, including topics such as (1) the severity of problems uncovered, (2) comparisons across different approaches to usability testing, such as remote and unmoderated testing, and (3) comparisons to other usability methods. These are all important topics, and I encourage the reader to explore them, but for the purposes of this paper, I have chosen to focus on sample size, the evaluator effect, and the think-aloud method. There has been a fair amount of research on these topics, and they may be of particular interest to the Federal statistical community.

Sample Size

The earliest usability tests were extensions of experimental psychology studies, with all the same scientific rigor, including a sufficient sample size. As the method evolved, it became more informal and flexible, and less bound to

the statistical requirements of scientific research. As a result, user experience practitioners began to explore how a smaller sample size would impact the results, since tests with fewer participants would be faster and less expensive.

Studies on this topic tend to follow the same approach. The researcher conducts usability tests, then identifies the number of usability problems each participant uncovered (it is up to each researcher to define and identify what a usability problem is). By taking random samples of smaller groups of participants, the researcher could identify what percent of all the problems (uncovered by all the participants in the study), on average, you could expect to find with each sample size. For example, to find out how well a sample of two participants would do, the researcher would take random combinations of two participants and calculate how many unique problems the two participants together uncovered, then determine what percent these two participants together found of the total number of problems uncovered by all of the participants. The researcher would create many random combinations of different size samples to determine how sample size impacts the percentage of problems uncovered.

In some of the earliest work, Virzi (1990, 1992) used this method with two studies. In one, he studied an electronic calendar. Twenty undergraduate students served as participants, and he found 40 problems. In his second study, he evaluated a voicemail system with 12 general population participants, and found 13 usability problems. By creating random samples of different sizes, he found that on average, four or five participants would reveal about 80% of the usability problems uncovered by all the participants, with diminishing returns with additional participants.

Based on this research, user experience professionals began using smaller sample sizes in usability tests. Additional research also supported their findings, such as Nielsen and Landauer (1993), who reported on five usability tests on a variety of applications from an office setting (e.g., word processing and calendar systems).

However, not everyone agreed with this conclusion. Spool and Schroeder (2001) conducted a study with sites that sell music and videos with participants from the general population. They identified 378 usability problems. They reported that for one site, their first five participants uncovered just 35% of the problems, and they continued to find new problems with all 18 of their participants. In this case, their research design was different, as they did not generate random samples, but instead focused on just their experience.

Faulkner (2003) also concluded that testers may want to include more than five participants. She ran a usability test of a weekly time sheet system with 60 participants. With 100 random samples of five participants, she found that on average, the five-person samples uncovered 85% of the problems. This finding is consistent with Virzi's earlier recommendations. However, the worst performing group revealed just 55%. With groups of 10 participants, the average increases to 95% and the minimum to 82%. This study provides some support for the "five-user" guideline, but also suggests it may be better to have more participants to ensure you find more of the problems.

Over time, user experience professionals have explored some of the factors that may impact the number of participants needed for a usability test. The table below identifies some factors that influence the number of participants recommended for a usability test.

Factor	You need more participants when...	Why?
Number of user groups	There are more user groups (i.e., distinct categories of people who might use a product in different ways, such as managers versus employees)	Each user group will likely have their own goals and approaches for working with the system, so it is important to include all the groups in usability testing.
Heterogeneity of the user groups	A user group is more diverse	With a diverse group, people are more likely to have different experiences and expectations and to use the system in different ways, and therefore more likely to uncover different kinds of problems.
Complexity of the system	The system is more complex	There may be more tasks or components of the system to test than each participant can complete in one session. Also, there may be more ways for users to get lost or make mistakes.

Factor	You need more participants when...	Why?
Opportunities for additional testing	There is only one opportunity to conduct usability testing	If there are future opportunities for testing, and therefore for uncovering additional problems, it is not so critical to find as many problems as possible in one round. With just one round, it is more important to identify as many problems as possible.
Purpose of the test	You want to use statistics (e.g., comparing task times between two designs) or other quantitative measures (e.g., eye tracking) to evaluate the findings	There needs to be a sufficient number of participants in order to conduct statistics on the quantitative measures.
Amount of previous user experience work on the system	The development process incorporated less user experience work.	Any previous user experience work on the product (in research, design, or evaluation) could have identified and resolved potential usability problems early, so there may not be so many to uncover during one usability test.

In addition, practical issues such as budget, time, and availability of participants will also influence the number of participants to include in a study. Test administrators should consider all these factors when determining the number of participants for a usability test. A good general guideline is 5-10 participants per user group for qualitative tests (i.e., those focusing on finding problems or other qualitative data) and 20-30 per user group for quantitative tests (i.e., those where statistical comparisons or quantitative measures are more important).

Macefield (2009) has a good exploration of the research findings and the reasons for the differences. He also notes that all these studies focus on conducting usability tests to find usability problems. This is a common goal of usability testing, but not the only goal. For example, the research doesn't cover how the number of participants affects the results in A/B testing, where the goal is to compare quantitative metrics, such as task time or success rate, across multiple designs.

One additional consideration for usability testing at government agencies is the requirement from the Office of Management and Budget (OMB) to get clearance for testing with ten or more participants. Because the process is complex and lengthy, agencies sometimes conduct tests with nine or fewer participants, even in situations that might benefit from a larger sample size. Although some testing is better than none, agencies should consider the limitations of small sample sizes as they are evaluating their test results.

Because of this requirement, agencies may not be able to include participants from all user groups. It may be possible to get a fairly diverse group of nine participants, but it may be impossible to determine how specific users groups would interact with the system as a whole. In addition, agencies will not find all, or maybe even most, of the usability problems. However, they will probably find enough to act on. Also, tests with smaller sample sizes should focus on qualitative findings (e.g., usability problems and recommendations) rather than quantitative measures. Agencies should be careful about how they report quantitative data from studies with smaller sample sizes, including measures such as task time and success rates, and other metrics such as eye tracking. Finally, in situations where a system is critical or where there are safety concerns, agencies should consider completing the OMB process to be able to test a sufficient number of participants.

Evaluator Effect

Just as it is important to think about how many participants to have, it is also important to think about how many test observers to have. Several studies have explored a phenomenon called the "Evaluator Effect," where the problems identified in usability testing depend significantly on the number of trained professionals (test administrators plus additional observers) who are observing the sessions and reporting problems. In an early study of the evaluator effect, Jacobsen, Hertzum, and John (1998a, 1998b) had four evaluators review videos from usability test sessions.

Of the 93 problems identified across all evaluators, only 20% were found by all four, and 46% were found by only one evaluator.

Jacobsen et al. concluded the impact of adding more evaluators may be similar to the impact of adding more participants. In fact, adding more evaluators can be an effective strategy for usability testing, especially when participants may be hard to find or there is not much time for testing.

In addition to the work of Jacobsen et al., Rolf Molich has conducted a series of nine studies looking at the results of different labs conducting the same usability test, which he calls the [Comparative Usability Evaluation \(CUE\)](#) studies. In the first study, four usability labs conducted a usability test on a calendar program (Molich, et al., 1998). Of the 162 problems reported across all four labs, only 13 (8%) were reported by more than one lab. Only one problem was reported by all four labs. The authors suggest that these large differences may have been due to the large number of problems in the system overall, as well as the differences in approaches that the labs took to test the system.

In subsequent studies, Molich refined the research protocol to control for various aspects of the testing process, to try to identify sources of the inconsistencies. These efforts included (1) providing more direction on the components of the system to evaluate (CUE-2, Molich, Ede, Kaasgaard, and Karyukin, 2004), (2) using highly experienced evaluators (CUE-4, Molich and Dumas, 2008), and finally, (3) repeating the Jacobsen, Hertzum, and John protocol, where the evaluators reviewed recordings of test sessions, rather than conducting the tests themselves (CUE-9, Molich, McGinn, and Bevan, 2013). All of the studies, including the last one, continued to find significant differences in the problems reported by the participating evaluators.

It is helpful to understand why this might happen. Usability tests require interpretation and judgment, which are very individual activities. Differences in training, experience, expertise, and judgment process all impact how the evaluators interpret the results. In addition, for the CUE studies, differences in test protocols can also lead to different findings.

In light of these findings, it can be tempting to discount usability testing as an inaccurate method. However, because it is so helpful for identifying and addressing usability problems, a better approach may be to understand its strengths and weaknesses in the planning, conducting, and analysis of usability test sessions. The most important take away from these studies is probably the significant benefit of having additional evaluators observe and analyze usability tests. It may be tempting to use just one test administrator to save money, but that limits the effectiveness of the method.

It is important to recognize this characteristic of usability tests. One test administrator working alone will be able to identify usability problems and solutions, but will be limited in what he or she will find. Instead, if there are opportunities for multiple evaluators to participate, the diversity will have a greater impact than if they worked alone (Hertzum and Jacobsen, 2001).

It may also be useful to reconsider the role of usability testing. For a long time, usability experts thought of it as the “gold standard” of usability methods, the one that would find ALL the usability problems. Reports often listed pages and pages of problems, giving the impression that the results were thorough and complete. Instead, it may be helpful to think of usability testing as a good way to identify some usability problems, which, when corrected, will improve the overall usability of the system. With an iterative testing cycle, this can be very effective. In addition, there may be some benefit in limiting the findings to the top five or ten problems, as the development team may find a shorter report to be more manageable and realistic than a long list of recommendations. This can be an effective approach for many kinds of products and systems. However, critical systems that may impact people’s health and safety should still follow a more rigorous approach.

Think-Aloud Methods

One standard strategy for usability tests is to ask the participants to “think aloud” during the test session. These comments form the basis for much of the qualitative findings from test sessions. Therefore, it is important to understand how the method can impact findings.

The think-aloud method evolved from the Verbal Protocol Analysis method originally developed by Ericsson and Simon (1993, originally published in 1984). Their goal was to focus on verbalizations of short-term memory activity as participants completed tasks. They recommended providing specific instructions about how to think aloud and having the participant practice before the data collection begins. During the test session, the test administrator should not interact with the participant at all, except to say “keep talking” when the participant is quiet for a period of time.

Their approach does not include probes requiring participants to access long-term memory, which are common in usability tests. These usability test probes include questions such as “Why did you do that?” and “What did you expect to happen?” Further, they believed that interrupting the participant’s natural process for any reason compromised the validity of the remainder of the test session. However, as usability testing evolved, test administrators were encouraged to use these more specific probes as a way to get more qualitative information from a test session (e.g., Dumas and Redish, 1999, originally published in 1993). In addition, Boren and Ramey (2000) suggest that a conversational approach to interacting with participants might be more natural and comfortable for participants than repeated probes to “keep talking.”

Several studies have evaluated the impact of thinking aloud on usability test results. In an early study, Wright and Converse (1992) explored a think-aloud method where participants explained every step they took during the test session. If they stopped talking for a period of time, the experimenter asked, “What are you thinking about?” The researchers found that thinking aloud led to fewer errors and shorter task times than a silent control condition. This may seem counter-intuitive, since adding an extra task (thinking aloud) might add an additional burden. However, thinking aloud may make participants more aware of their thought processes and allow them to work out a solution faster (Dumas and Redish, 1999).

Later, Hertzum, Hansen, and Andersen (2009) conducted another study of the impact of thinking aloud on usability test results. They found that the traditional method (i.e., Ericsson and Simon’s method) did not influence performance, but a more relaxed method, like the method common in usability testing, led to longer task times and higher mental workload, as measured with the NASA-TLX tool (Hart and Staveland, 1988). They found that task success was not affected. These findings were different from Wright and Converse’s earlier work, but they still show that thinking aloud affects performance. Additionally, Olmsted-Hawala, Murphy, Hawala, and Ashenfelter (2010) found that the more relaxed method (what they called “coaching”) led to higher task success and higher satisfaction, but did not affect task time.

Because of the variety of approaches to thinking aloud, Nørgaard and Hornbæk (2006) studied how people are actually using the think aloud in usability tests. The practitioners they studied tended to use the more relaxed think aloud method in their usability tests. The most significant finding of the study is that, without standard probes such as “keep talking,” the test administrators often asked questions that were not based on their observations of that specific test session, but rather were intended to obtain support for their pre-existing hypotheses about potential usability problems. In addition, many questions were hypothetical (e.g., “What would you do if...”) rather than based on the existing product and scenarios they were testing.

The results described above suggest that it is important to carefully consider how to implement thinking aloud in usability tests. User experience professionals find it to be a powerful tool for identifying usability problems and solutions, but it can impact the participant’s performance. Because thinking aloud can affect some test metrics (e.g., task time, success rate), it may be better not to use it for tests focusing on quantitative measures. Further, test administrators should carefully consider the types and content of the probes they want to use to encourage participants to speak.

Conclusion

The studies described in this paper shed some light on the science of usability testing, in terms of the number of participants, the number of observers, and the impact of the think aloud on test findings. There is more to be studied, but researchers have made significant progress in understanding how the nuances of the methods of usability testing can impact the results. These lines of research are critical to better understand the strengths and weaknesses of usability testing. They also help user experience professionals as they plan and conduct tests, as well as how they interpret and present the results.

References

- Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261-278.
- Dumas, J. S., & Redish, J. (1999). *A Practical Guide to Usability Testing*. Intellect Books.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Using Verbal Reports as Data*. Cambridge, MA: The MIT Press.
- Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3), 379-383.
- Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press. Retrieved from http://humansystems.arc.nasa.gov/groups/tlx/downloads/Hart_Staveland.PDF.
- Hertzum, M., Hansen, K. D., & Andersen, H. H. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2), 165-181.
- Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4), 421-443.
- Jacobsen, N. E., Hertzum, M., & John, B. E. (1998b). The evaluator effect in usability studies: Problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (336-1340).
- Jacobsen, N. E., Hertzum, M., & John, B. E. (1998b). The evaluator effect in usability tests. In *CHI 98 Conference Summary on Human Factors in Computing Systems* (255-256).
- 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. Retrieved 9/17/2015 from <http://uxpajournal.org/how-to-specify-the-participant-group-size-for-usability-studies-a-practitioners-guide/>.
- Molich, R. (n.d.). CUE (Comparative Usability Evaluation). Retrieved October 27, 2015, from <http://www.dialogdesign.dk/CUE.html>.
- Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *The Proceedings of the Usability Professionals Association Conference*.
- Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, 27(3), 263-281.
- Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology*, 23(1), 65-74.
- Molich, R., McGinn, J., & Bevan, N. (2013). You say disaster, I say no problem: Unusable problem rating scales. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 301-306).
- Nielsen, J., & Landauer, T. K. (1993, May). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems* (206-213).
- Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the 6th Conference on Designing Interactive Systems* (pp. 209-218).

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In the *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2381-2390).

Spool, J., & Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (pp. 285-286).

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (291-294).

Virzi, R. A. (1992). Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(4), 457-468.

Wright, R. B., & Converse, S. A. (1992). Method bias and concurrent verbal protocol in software usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 36(16), (pp. 1220-1224).