Behavior Research Methods, Instruments, & Computers
2003, 35 (3), 379-383

Beyond the five-user assumption: Benefits of increased sample sizes in usability testing

LAURA FAULKNER
University of Texas, Austin, Texas

It is widely assumed that 5 participants suffice for usability testing. In this study, 60 users were tested and random sets of 5 or more were sampled from the whole, to demonstrate the risks of using only 5 participants and the benefits of using more. Some of the randomly selected sets of 5 participants found 99% of the problems; other sets found only 55%. With 10 users, the lowest percentage of problems revealed by any one set was increased to 80%, and with 20 users, to 95%.

Many usability professionals struggling with limited budgets and recognition use only 5 participants for usability testing, rather than larger samples typically required for empirical research. Large numbers of test sessions call for resources not readily available to usability practitioners, who are frequently solo venturers within a development group or company. Despite its attractiveness for usability professionals' efforts to gain acceptance for themselves and their practices, other practitioners wondered if what this author has termed the *5-user assumption* was appropriate and representative of best practices for the field. Articles with titles such as *Why Five Users Aren't Enough* (Woolrych & Cockton, 2001) and *Eight is Not Enough* (Perfetti & Landesman, 2002) critique the assumption, highlight issues of reliability with small user sets, and express concern over the impact of usability problems that may be missed when only 5 users are tested.

Early studies supporting the assumption argued that just 5 participants could reveal about 80% of all usability problems that exist in a product (Nielsen, 1993; Virzi, 1992). This figure indicates a probability of the percentage of problems missed; there is, currently, no way to determine with reasonable certainty that any set of five tests matched those percentages, or which particular problems were revealed or missed (Woolrych & Cockton, 2001). Furthermore, if, for example, only novice users were tested, a large number of usability problems may have been revealed, but the test would not show which are the most severe and deserve the highest priority fixes. Expert results may highlight severe or unusual problems but miss problems that are fatal for novice users. Finally, the abstract argument in favor of the assumption depends on the independence of the problems encountered—that is, that encountering one of them will not increase or decrease the likelihood of encountering any other problem.

This author envisioned a way to address these issues: Conduct usability tests in which data are collected from larger samples of multilevel users. If the resulting data could then be presented in an accessible manner for usability professionals who are not well versed in the complexities of statistics, the practitioners could come to recognize the risks of the 5-user assumption and the benefits of additional users.

Background

The 5-user assumption arose from two sources: (1) secondary analyses of other testers' data by Nielsen (1993) and (2) "law-of-diminishing-returns" arguments made by Virzi (1992). Both Nielsen (1993) and Virzi sought to demonstrate that statistical rigor can be relaxed considerably in real-world usability testing. However, in applying the assumption, usability practitioners have experienced limitations. For example, in one study (Spool & Schroeder, 2001) the first 5 users revealed only 35% of the usability problems. Furthermore, both the 13th and 15th users tested revealed at least one new *severe* problem that would have been missed if the study team had stopped after the first five test sessions. Another study team tested 18 users; each new user, including those in test sessions 6–18, found "more than five new obstacles" (Perfetti & Landesman, 2002). Again, the problems in those later sessions would have been missed had the study team stopped after the first five user-test sessions. The usability problems found by these teams and others, beyond the 5-user barrier, indicate the need to move usability practices toward increasing maturity to account for usability problems missed by the first 5 users. Building on the foundational work of Nielsen (1993) and Virzi, it is appropriate to revisit the calculations and data on which the assumption was based.

Virzi's (1992) essential finding was that 5 users would uncover approximately 80% of the usability problems in a product. In cases of the most severe errors, he indicated

Correspondence concerning this article should be addressed to L. Faulkner, Applied Research Laboratories, University of Texas at Austin, P. O. Box 8029, Austin, TX 78713-8029 (e-mail: laura@arlut.utexas.edu).

that only 3 users would reveal most of the problems. He calculated these various sample sizes against the number of errors revealed by 12 users in the first study and by 20 in the second and third studies.

For some time, Nielsen has been writing in support of the idea that 5 test users are sufficient in usability testing (Landauer & Nielsen, 1993; Nielsen, 1993) and remains a strong proponent of the assumption (Nielsen, 2000). He based his initial calculations of user error rates on data from 13 studies. In calculating the confidence intervals, he uses the *z* distribution, which is appropriate for large sample sizes, rather than the *t* distribution, which is appropriate for small sample sizes. Using *z* inflates the power of his predictions; for instance, what he calculates as a confidence interval of ±24% would actually be ±32% (Grosvenor, 1999). Woolrych and Cockton (2001), in their detailed deconstruction of Landauer and Nielsen's (1993) formula, confirmed the potential for overpredicting the reliability of small-sample usability test results, demonstrating the inflated fixed value recommended by Landauer and Nielsen for the probability that any user will find any problem.

Nielsen (1993) and Virzi (1992) both made attempts to describe the limitations of their 5-user recommendations. Virzi indicated that "[s]ubjects should be run until the number of new problems uncovered drops to an acceptable level" (p. 467), leaving the practitioner to define "acceptable level." Nielsen (1993) included explanations of "confidence intervals" and what the calculations actually indicated; however, practitioners tend to adopt a minimal actual number of test users, specifically, 5. Grateful practitioners overlooked the qualifying statement in Nielsen's (1993) text that indicated that 5 users "might be enough for many projects" (p. 169). They then shared the information through mentoring relationships, thereby propagating the 5-user assumption (Grosvenor, 1999).

The assumption was examined by Nielsen (1993) via data from other professionals' usability tests, by Virzi (1992) in direct tests but by extrapolating results from small numbers of users overall, by Woolrych and Cockton (2001) in a secondary analysis of data from a heuristic evaluation study, and by Spool and Schroeder (2001) in an unstructured goal-oriented test.

The present study was designed to test the 5-user assumption in a direct and structured manner. Data generated by the 60 users in the test allowed for sampling the results in sets of 5 or more, comparing the problems identified by each set against the total problems identified by the entire group. This process was used to measure the effect that sets of different sizes would have on the number of usability problems found, data reliability, and confidence.

Method

The study was a structured usability test of a web-based employee time sheet application. Sixty user-participants were given a single task of completing a weekly time sheet and were provided with the specific data to be entered. Rather than focus only on novices or experts, this study was designed to capture a full range of user data in a single usability test. The 60 participants, then, were sampled from three levels of user experience and given the following designations: (1) *novice/novice* (inexperienced computer users who had never used the application); (2) *expert/novice* (experienced computer users who had never used the application); and (3) *expert/expert* (experienced computer users who were also experienced with the application).

All test sessions were conducted in a single location on the same computer to control for computer performance and environmental variation. Two types of data were collected: (1) time, measured in minutes to complete the test task, and (2) user deviations, measured on a tabular data collection sheet devised to ensure that the same types of data were collected from each session. The primary characteristic of the data sheet was a detailed list of user actions and the names of the specific windows and elements with which the users would interact to perform each action. The action list was derived by determining the optimal path to completion of the task—specifically, a set of steps that would allow the user to complete the given task with the simplest and fastest set of actions. Actual user behavior was logged by simple tick marks next to each optimal path step whenever the participant deviated from that step.[1] Multiple deviations on a single step were noted with multiple tick marks. The basic measure analyzed for this study was total number of deviations committed by each user on all elements.

Results and Analyses

The primary results were straightforward and predictable, with user deviations and time to completion being higher for those with the least experience and lower for those with more experience, as is shown in Table 1. Standard deviations (*SD*s) were large, as is common in usability studies (Nielsen, 1993), with the novice/novice group having the largest on both measures. Variances within the groups were smaller at the higher experience levels. Post hoc tests indicated that each of the three groups differed significantly from the others in user deviations [$F(2,57) = 70.213$, $p < .01$] and time to complete [$F(2,57) = 63.739$, $p < .01$].

To draw random samples of user data from the complete data set, the author wrote a program in MATLAB that allowed for the drawing of any sample size from the total of 60 users. The program ran 100 trials each, sampling 5, 10, 20, 30, 40, 50, and all 60 users. The full group of 60 users identified 45 problems. In agreement

Table 1
Group Means for User Deviations Logged and Time to Complete Task

Experience Level*	*M*	*SD*
User Deviations		
Novice/novice	65.60	14.78
Expert/novice	43.70	14.16
Expert/expert	19.20	6.43
Time to Complete		
Novice/novice	18.15	4.46
Expert/novice	10.30	2.74
Expert/expert	7.00	1.86

Note—Means differed significantly in the Tukey honestly significant difference comparison for user deviations [$F(2,57) = 7.213$, $p < .01$] and for time to complete [$F(2,57) = 63.739$, $p < .01$]. *$n = 20$ for each group.

with the observations of Nielsen (1993) and Virzi (1992), the average percentage of problem areas found in 100 trials of 5 users was 85%, with an *SD* of 9.3 and a 95% confidence interval of ±18.5%. The percentage of problem areas found by any one set of 5 users ranged from 55% to nearly 100%. Thus, there was large variation between trials of small samples.

Adding users increased the minimum percentage of problems identified. Groups of 10 found 95% of the problems (*SD* = 3.2; 95% confidence interval = ±6.4). Table 2 shows that groups of 5 found as few as 55% of the problems, whereas no group of 20 found fewer than 95%. Even more dramatic was the reduction in variance when users were added. Figure 1 illustrates the increased reliability of the results when 5, 10, and 15 users were added to the original sets of 5.

To summarize, the risk of relying on any one set of 5 users was that nearly half of the identified problems could have been missed; however, each addition of users markedly increased the odds of finding the problems.

Discussion

This study supports the basic claims of Nielsen (1993) and Virzi (1992), but not the assumption that usability practitioners have built around those claims—namely, that 5 users are a sufficient sample for any usability test. Merely by chance, a practitioner could encounter a 5-user sample that would reveal only 55% of the problems or perhaps fewer, but, on the basis of the 5-user assumption, still believe that the users found 85%. Furthermore, this study provided a visual reference for practitioners to apply the concept of variability and to readily grasp the increasing reliability of data with each set of participants added to a usability test.

Hudson (2001) indicated that small numbers of participants may be used in "detailed and well-focused tests." The high *SD*s in the present study occurred even within the well-defined controls and structured nature of the experiment. Variability has become a more prevalent issue as usability testing has extended to unstructured testing of websites (Spool & Schroeder, 2001). Furthermore, the lack of controls in real-world usability testing provides more opportunities for unequal *n*s between different user groups, thereby creating a greater risk of violating the homogeneity-of-variance assumption.

A problem with relying on probability theories and prediction models to drive usability testing, as suggested by Nielsen (1993) and Virzi (1992), is that in an applied situation it is difficult to accurately calculate the probability of finding a given usability problem (Woolrych & Cockton, 2001). Each usability problem has its own probability of being found, due to factors such as severity, user traits, product type, level of structure in the test, and the number of users tested (Grosvenor, 1999; Woolrych & Cockton, 2001). In terms of severity, for one, a glaring problem has a high probability of being found, but a subtle problem has a lower one; more test users are required to find low-severity problems than high-severity problems (Virzi, 1992). Unfortunately, the subtle problem may have the more serious implications, as was the case in the 1995 crash of American Airlines flight 965 near Cali, Colombia, the Three-Mile Island accident in 1979, and similar events, in which one or more subtle usability problems significantly contributed to the disasters (Reason, 1997; Wentworth, 1996). The subtle problems in those cases were missed by numerous previous users of the system and, accordingly, would have been missed by small usability test groups.

Table 2
Percentage of Total Known Usability Problems Found in 100 Analysis Samples

No. Users	Minimum % Found	Mean % Found	*SD*	*SE*
5	55	85.55	9.2957	.9295
10	82	94.686	3.2187	.3218
15	90	97.050	2.1207	.2121
20	95	98.4	1.6080	.1608
30	97	99.0	1.1343	.1464
40	98	99.6	0.8141	.1051
50	98	100	0	0

Conclusion

Perhaps the most disturbing aspect of the 5-user assumption is that practitioners have so readily and widely embraced it without fully understanding its origins and the implications (e.g., in this study, that any given set of 5 users may reveal only 55% of the usability problems). Contentment with an 80% accuracy rate for finding usability errors demonstrates the belief that the 5 users' actions will always fall within the average and that 80% of the usability problems have actually been revealed.

Both Nielsen (1993) and Virzi (1992) were writing in a climate in which the concepts of usability were still being introduced into the software development field, as they still are in many organizations. They were striving to lighten the requirements of usability testing in order to make usability practices more attractive to those working with the strained budgets and in the developer-driven environment of the software industry. Nielsen himself is credited as being the inventor of "discount usability." However, as usability is more fully recognized as essential to the development effort, software organizations may be forced to rethink the sufficiency of a 70% chance of finding 80% of the usability problems in a given product (Nielsen, 1993).

Despite practitioners' love of the 5-user assumption, the answer to the question, "How many users does it take to test the usability of an interface?" remains a vague, "It depends." Variables over which the practitioners have varying levels of control, such as types of test users available or accessible to the practitioner, the mission criticality of a system, or the potential consequences of any particular usability problem, can have a profound impact on the number of test users required to obtain accurate and valid results. The assumptions inherent in the mathematical formulas and models attempted by Nielsen (1993, 2000) and others, and the information required to

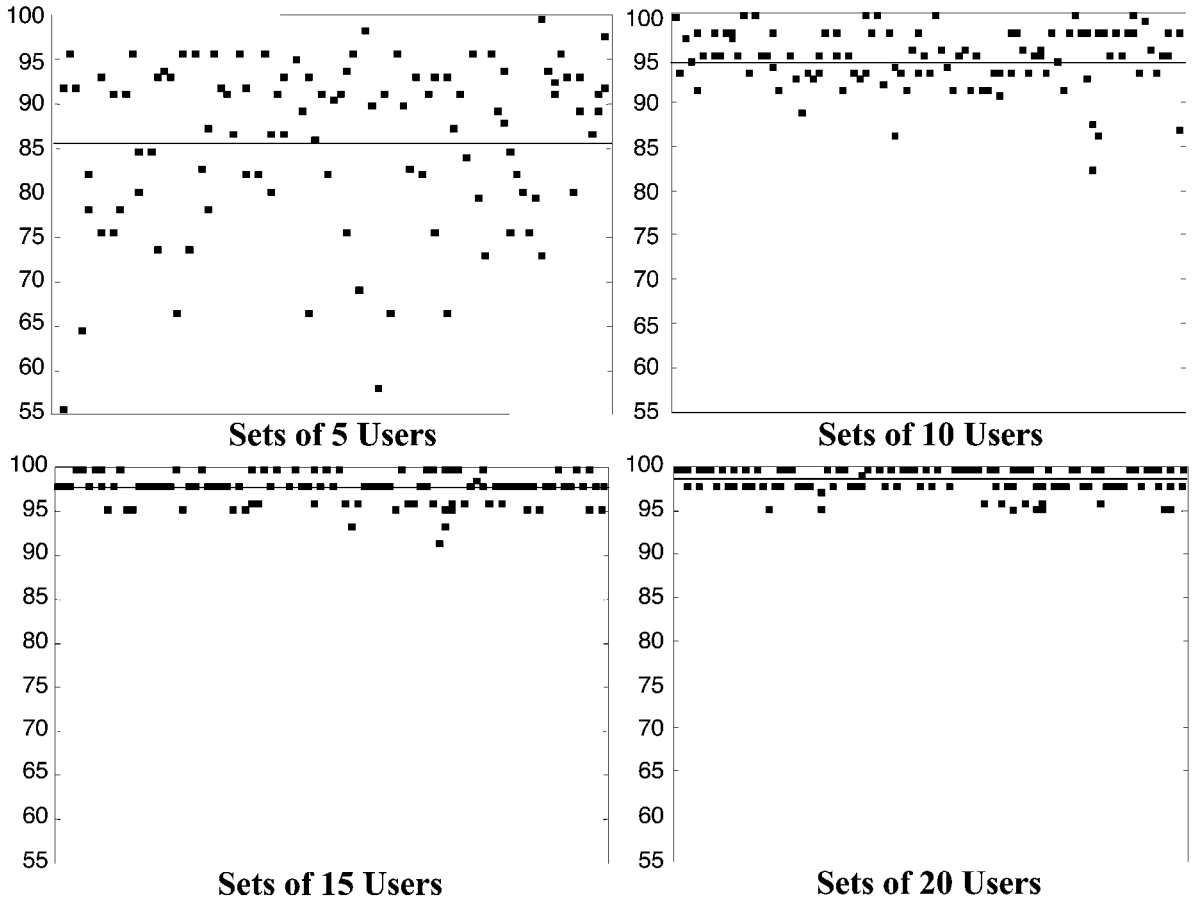

Figure 1. The effect of adding users on reducing variance in the percentage of known usability problems found. Each point represents a single set of randomly sampled users. The horizontal lines show the mean for each group of 100.

use those formulas, such as probabilities, make them impractical and misleading for ordinary usability practitioners. Although practitioners like simple directive answers such as the 5-user assumption, the only clear answer to valid usability testing is that the test users must be representative of the target population. The important and often complex issue, then, becomes defining the target population. There are strategies that a practitioner can employ to attain a higher accuracy rate in usability testing. One would be to focus testing on users with goals and abilities representative of the expected user population. When fielding a product to a general population, one should run as many users of varying experience levels and abilities as possible. Designing for a diverse user population and testing usability are complex tasks. It is advisable to run the maximum number of participants that schedules, budgets, and availability allow. The mathematical benefits of adding test users should be cited. More test users means greater confidence that the problems that need to be fixed will be found; as is shown in the analysis for this study, increasing the number from 5 to 10 can result in a dramatic improvement in data confidence. Increasing the number tested to 20 can allow the practitioner to approach increasing levels of certainty that high percentages of existing usability problems have been found in testing. In a mission-critical system, large user sets at all experience levels should be tested. Multiple usability strategies should be applied to complement and supplement testing.

Usability test results make for strong arguments with design teams and can have a significant impact on fielded products. For example, in the complex intertwining of systems and with the common practice of integrating commercial, off-the-shelf software products into newly developed systems, implications of software usability problems cannot always be anticipated, even in seemingly simple programs. The more powerful argument for implementing software usability testing, then, is not that it can be done cheaply with, say, 5 test users, but that the implications of missing usability problems are severe enough to warrant investment in fully valid test practices.

REFERENCES

GROSVENOR, L. (1999). *Software usability: Challenging the myths and assumptions in an emerging field.* Unpublished master's thesis, University of Texas, Austin.

HUDSON, W. (2001). How many users does it take to change a website? *SIGCHI Bulletin May/June 2001*. Retrieved April 15, 2003 from http://www.acm.org/sigchi/bulletin/2001.3/mayjun01.pdf.

LANDAUER, T. K., & NIELSEN, J. (1993). A mathematical model of the finding of usability problems. *Interchi '93*, ACM Computer–Human Interface Special Interest Group.

NIELSEN, J. (1993). *Usability engineering*. Boston: AP Professional.

NIELSEN, J. (2000, March). *Why you only need to test with 5 users: Alertbox*. Retrieved April 15, 2003 from http://www.useit.com/alertbox/20000319.html.

PERFETTI, C., & LANDESMAN, L. (2002). *Eight is not enough*. Retrieved April 14, 2003 from http://world.std.com/~uieweb/Articles/eight_is_not_enough.htm.

REASON, J. (1997). *Managing the risks of organizational accidents*. Brookfield, VT: Ashgate.

SPOOL, J., & SCHROEDER, W. (2001). Testing web sites: Five users is nowhere near enough. In *CHI 2001 Extended Abstracts* (pp. 285-286). New York: ACM Press.

VIRZI, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, **34**, 457-468.

WENTWORTH, R. J. (1996, March). *Group chairman's factual report (American Airlines Flight 965)*. Washington, DC: National Transportation Safety Board, Office of Aviation Safety.

WOOLRYCH, A., & COCKTON, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), *Proceedings of IHM-HCI 2001 Conference: Vol. 2* (pp. 105-108). Toulouse, France: Cépadèus.

NOTE

1. At the time of the study, this data notation technique was simply shorthand for the usual note taking employed by usability professionals for many years. The concept of the approach as a possible unique data collection method is a recent development. Examination of it as such is planned as a full, independent study. The data in the 5-user study is consistent with that in other publications, such as the ones cited throughout this article.

(Manuscript received September 19, 2002;
revision accepted for publication May 18, 2003.)