HUMAN FACTORS, 1992, 34(4), 457–468

Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?

ROBERT A. VIRZI,[1] *GTE Laboratories Inc., Waltham, Massachusetts*

Attention has been given to making user interface design and testing less costly so that it might be more easily incorporated into the product development life cycle. Three experiments are reported in this paper that relate the proportion of usability problems identified in an evaluation to the number of subjects participating in that study. The basic findings are that (a) 80% of the usability problems are detected with four or five subjects, (b) additional subjects are less and less likely to reveal new information, and (c) the most severe usability problems are likely to have been detected in the first few subjects. Ramifications for the practice of human factors are discussed as they relate to the type of usability test cycle employed and the goals of the usability test.

INTRODUCTION

A need exists to reduce the cost of applying good design practices, such as iterative design or user testing, to the development of user interfaces (see Meister, 1989, chap. 7). The application of traditional human factors techniques is considered too expensive (Bellotti, 1988; Nielsen, 1989), too cumbersome (Melkus and Torres, 1988), or too time consuming (Denning, Hoiem, Simpson, and Sullivan, 1990; Nielsen and Molich, 1990). Several approaches have been taken toward limiting the cost of design. For example, ways to reduce the cost of prototyping were suggested by Nielsen (1989), who advocated limiting the scope of the prototype by eliminating application functionality or by limiting the number of the features represented. Virzi (1989) discussed reducing the overall level of fidelity of the prototype. Jeffries, Miller, Wharton, and Uyeda (1991) compared the cost-effectiveness of four methods for conducting a usability analysis. Denning et al. (1990) reported a method for reducing the time spent analyzing verbal protocols. The focus of this paper is on another alternative for reducing the cost of designing user interfaces: running fewer subjects in any iteration of a usability test. This work extends and refines work I reported earlier (Virzi, 1990).

The initial motivation for the current experiments came from Nielsen (1990), who reported an experiment that was designed to measure the percentage of usability problems computer scientists would find using the think-aloud technique (see Lewis, 1972, for a discussion of the technique). In this study, 20 groups of minimally trained experimenters (computer science students who had taken a course in usability testing) independently conducted usability tests of a paint program. Their task was to find as many of the usability

[1] Requests for reprints should be sent to Robert A. Virzi, GTE Laboratories Inc., 40 Sylvan Rd., MS 38, Waltham, MA 02254.

problems Nielsen had defined a priori as "major usability problems" as they could.

A surprising result, though tangential to the main thrust of the work, was how good the student experimenters were at finding the major usability problems Nielsen (1990) had identified. The students were minimally trained in human factors and ran an average of 2.8 subjects per evaluation. Still, for any given evaluation the mean probability of detecting a major usability problem was 0.40, with an average of 49% of all major usability problems detected.

Three experiments are reported in this paper that were developed to extend and generalize Nielsen's (1990) results. In these experiments I examined the rate at which usability problems were identified as a function of the number of naive subjects run in a single usability evaluation when the evaluation was conducted by trained usability experts. In addition, all problems were considered, not merely those defined a priori as the important problems.

EXPERIMENT 1

The data reported in this study were collected as part of the evaluation of a voice mail system conducted at GTE Laboratories Incorporated. Both the user manual and the interactive system were studied, but only the data from the analysis of the manual are reported here. The primary goal of this experiment is to show that the likelihood of uncovering a new usability problem decreases as more and more subjects participate in a usability evaluation; a formula is introduced for estimating this function.

Method

Subjects. Twelve subjects (aged 18 through 65) who had no prior experience with voice mail systems were recruited from surrounding communities through an advertisement in a local newspaper. They were paid a small honorarium for their participation in the study.

Tasks. Each subject was given the manual to the voice mail system being evaluated and was given an opportunity to study the manual, if desired. Later the subjects were asked to describe how the system would respond in three specific situations based on the information contained in the manual alone. Although they could refer to the manual, subjects never actually interacted with the voice mail system.

Procedure. Subjects were escorted into a testing room and seated at a desk with the manual. Before they were allowed to start the tasks, the think-aloud procedure was explained. Subjects were asked to maintain a running commentary as they interacted with the system (Lewis, 1972). The experimenter prompted the subjects to speak out loud if they fell silent during the course of completing tasks. The experimenter remained in the room during the evaluation and recorded the problems that subjects encountered and comments they made for later analysis.

Results

A total of 13 usability problems were identified based on an analysis of the verbal protocols of the 12 subjects. The analyses reported here are based on this input.

Subjects differed in the number of usability problems that they uncovered. One subject uncovered 8 of 13 usability problems (62%), whereas another subject uncovered only 2 of the 13 problems (15%). Problems also varied, in that some were uncovered by many subjects (10 of 12 subjects, or 83%), whereas one problem was uncovered by a single subject (8%). Table 1 presents these results.

A Monte Carlo procedure was applied to the data to derive the general form of the curve relating the proportion of usability problems uncovered to the number of subjects

TABLE 1

Voice Mail Manual Evaluation: Subjects' Ability to Detect Problems and Probability of Detecting Any Given Problem (percentages)

	M	*SD*	*Min.*	*Max.*
Problems identified per subject	32.0	0.14	15.0	62.0
Subjects uncovering each problem	32.0	0.20	8.0	83.0

participating in the evaluation (see Diaconis and Efron, 1983). A computer program was used to generate 500 permutations of the subject order and to calculate the mean number of unique problems identified at each sample size (1–12). The resultant curve is shown as the solid line in Figure 1.

Probability theory indicates that the proportion of problems one would expect to find at a given sample size is given by the formula $1 - (1 - p)^n$, where p is the probability of detecting a given problem and n is the sample size. If the probability of problem detection is taken as 0.32, the mean probability in the current sample, and plot that as a function of the number of subjects, the broken line shown in Figure 1 is obtained.

Discussion

The current data suggest that Nielsen's (1990) finding was not far from the mark. When only three subjects were run, the experimenter was likely to uncover about 65% of all the usability problems in the current study versus 49% in Nielsen and Molich's (1990) study. Although there were many differences between the two studies (expert vs. naive experimenters, searching for all problems vs. only the major problems, etc.), the basic and surprising finding still holds: the first few subjects run in an evaluation are likely to let the experimenter uncover the majority of the usability problems.

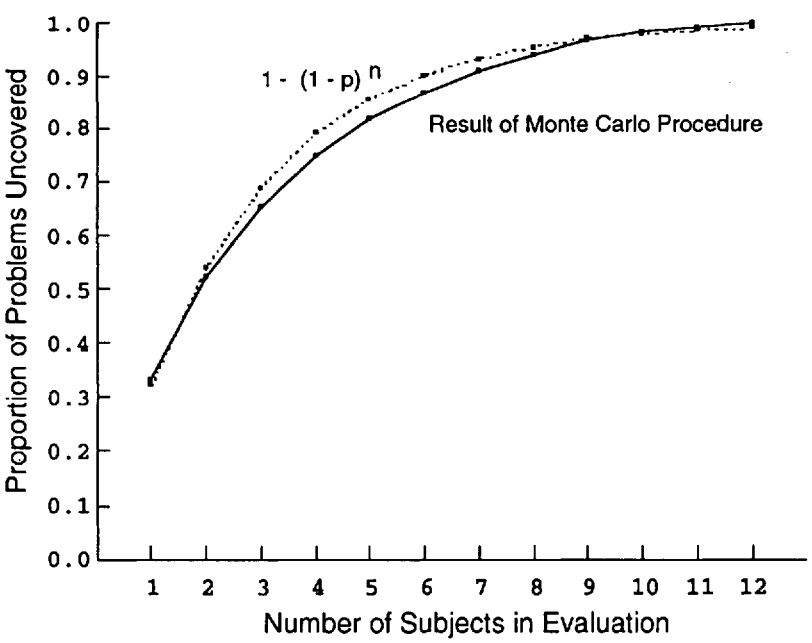

Figure 1. *The proportion of usability problems uncovered is shown as a function of the number of subjects who participated in an evaluation (solid line). The values expected when* p = *0.321 are shown as the broken line.*

Further, the results suggest that there are diminishing returns: with each additional subject, it is increasingly unlikely that new usability problems will be uncovered. Only three subjects are needed to uncover 65% of the problems, five are needed to find 80%, and nine are needed to find 95%, on average. Clearly, later subjects are not as likely to uncover new usability problems as are earlier ones.

Note that the proportion of problems identified across subjects using the Monte Carlo procedure is somewhat different from the curve predicted by probability theory. This is a consistent result, obtaining in the three experiments reported here as well as in other evaluations I have conducted. I will return to the reason for the overprediction in the general discussion.

EXPERIMENT 2

Experiment 1 demonstrated the basic effect of diminishing returns. Experiment 2 was conducted to determine whether the proportion of problems detected varies as a function of problem severity in addition to the number of subjects participating in the evaluation. This is a matter of great practical import because an evaluator may choose to run a small number of subjects only if he or she is convinced that there is not an inordinate amount of risk that the evaluation will fail to detect major usability problems. If the probability of detecting the major usability problems after running a few subjects is high, the evaluator might be willing to forgo detection of some less important problems, which could be detected in subsequent iterations, for example. However, if some severe usability problems are likely to be missed when the sample size is small, an evaluator may be less likely to restrict the sample size for fear that a critical problem will evade detection.

Under my supervision, three human factors undergraduate students conducted a usability evaluation of a computer-based appointment calendar as part of a design course offered at Tufts University. We believed an appointment calendar would be a good application for the naive subjects because we expected them to be familiar with the type of functionality such a program might deliver. The program we selected for evaluation was chosen because it was an early version of a shareware product that clearly had some usability problems. Pilot testing indicated that the subjects would be able to exercise the full functionality of the program within one hour, which is about as long as subjects could be expected to participate in this study.

Method

Subjects. Twenty undergraduates were recruited from an introductory psychology course at a private university. Only those applicants who reported little or no computer experience and no experience with computer-based appointment calendars were included in the study.

Tasks. We devised 21 tasks that exercised virtually all of the functions in the program. An appointment calendar was created that ostensibly planned the activities of an executive at a chocolate factory. Users were asked to imagine that they were that executive and were given tasks that required them to manipulate the calendar. The tasks were presented as short descriptions of actions that the executive wanted to perform on the calendar. For example, one task was, "Mr. Waters has called and would like to change his Monday appointment to Wednesday at 1:00 p.m. Please reschedule him."

Procedure. The subjects, who were tested individually, were seated in front of a computer with the appointment calendar running and visible on the screen. One of the experimenters interacted with the subject while a

second experimenter ran a video camera situated behind the subject which captured the subject's comments and actions for later analysis.

Each subject carried out the 21 tasks in a fixed order using the computer-based appointment calendar. Prior to the start of the tasks, the think-aloud procedure was explained. The experimenter prompted the subjects to speak out loud if they fell silent during the course of completing tasks.

Preanalysis. After all 20 subjects had been run—but prior to detailed examination of the videotape records—the experimenters jointly identified a set of 40 potential usability problems. For this analysis a usability problem was operationally defined as a change needed in the user interface of the calendar program. If the experimenters determined that a single change in the program might alleviate several potential errors, these were considered the same usability problem.

The 40 usability problems were rated by the experimenters on a seven-point scale (1 = *not likely to have a large impact on the usability of the calendar program* to 7 = *likely to be a severe obstacle to successful use of the program*). The problems were independently rated by each of the experimenters, and differences were resolved by consensus.

Subsequently the videotape of each subject was reviewed and problem episodes were identified. A problem episode could occur when (a) the subject made an error or became confused while interacting with the program and was aware of it, (b) the subject mentioned a potential problem with or inadequacy of the program but did not actually make an error, or (c) the experimenter identified a mistake or misuse of the program that the subject made but was not aware of. The usability problem (from the list of 40 problems) underlying each of the episodes was identified by at least two of the experimenters independently. Again, any discrepancies in the identification of the underlying problem among the experimenters were resolved by consensus.

These steps allowed us to identify the particular usability problems that each subject had noted and to gauge their relative severity. The raw data are presented in Virzi (1990).

Results

Subjects differed in the number of usability problems that they uncovered. The most prolific subject uncovered almost half of the usability problems (19 of 40), whereas the least informative subject uncovered only 8 (20%). Problems also varied in that some were uncovered by almost all of the subjects (19 of 20), whereas others were uncovered by a single subject (Table 2).

A Monte Carlo procedure was applied to the data to derive the general form of the curve relating the proportion of usability problems uncovered to the number of subjects participating in a usability evaluation. We took 500 permutations of the subject order and calculated the mean number of unique problems identified at each sample size. The resultant curve is shown as the solid line in Figure 2.

If the probability of problem detection is taken as 0.36, the mean probability in this sample, and plot that as a function of the number of subjects, the broken line as shown in Figure 1 is obtained. Again, the fit is close

TABLE 2

Appointment Calendar Evaluation: Subjects' Ability to Detect Problems and Probability of Detecting Any Given Problem (percentages)

	M	*SD*	*Min.*	*Max.*
Problems identified per subject	36.0	0.06	20.0	48.0
Subjects uncovering each problem	36.0	0.24	5.0	95.0

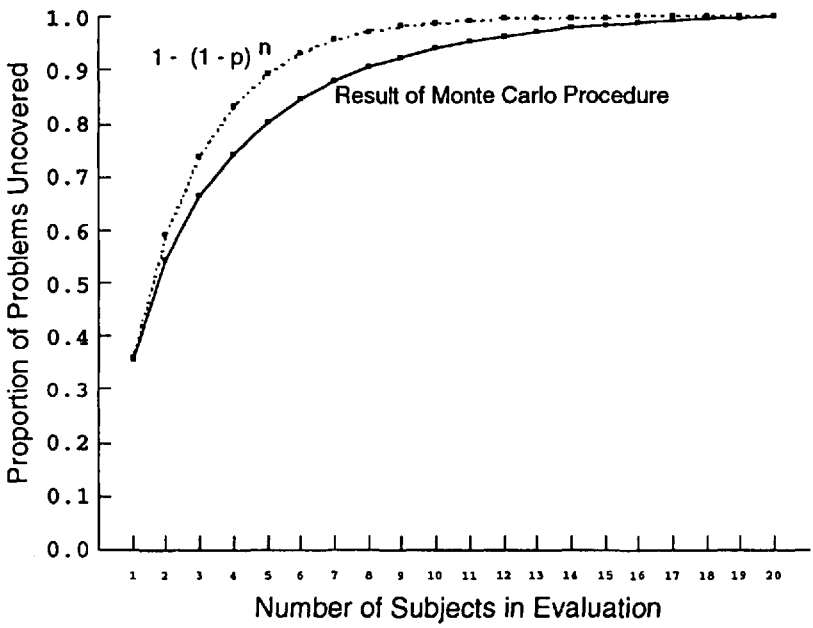

Figure 2. *The proportion of usability problems uncovered is shown as a function of the number of subjects who participated in an evaluation (solid line). The values expected when* p = *0.36 are shown as the broken line.*

but tends to overpredict the amount of information one would have at a given sample size.

In Figure 3 separate curves are shown for the problems at each level of severity. These curves were generated by permuting the subject order 100 times and plotting the mean proportion of problems that were uncovered at each level of severity as a function of the number of subjects run in the evaluation. The figure reveals that as problem severity increases, the likelihood that the problem will be detected within the first few subjects also increases. Put another way, judgments of problem severity and problem frequency are significantly correlated ($r = 0.463, p < 0.01$), suggesting that the more severe a problem is, the more likely it will be uncovered within the first few subjects.

Discussion

The results of this study replicated those of the first experiment. Approximately 80% of all of the usability problems were found after five subjects were run. The law of diminishing returns thus applies, and additional subjects are less likely to identify a new usability problem.

EXPERIMENT 3

Experiment 2 demonstrated the basic effect of diminishing returns over subjects and extended this to show that important usability problems tend to be found first. One flaw in the design of Experiment 2, however, was that the experimenters who conducted the evaluation also judged problem severity. They may have been influenced by exposure to the subjects and knowledge of problem frequency. To circumvent this problem, a third usability evaluation was conducted in which both the experimenters and a panel of six "double usability experts" (Nielsen and Molich, 1990) were asked to rank the problems in terms of how disruptive they were likely to be to the usability of the system. (A

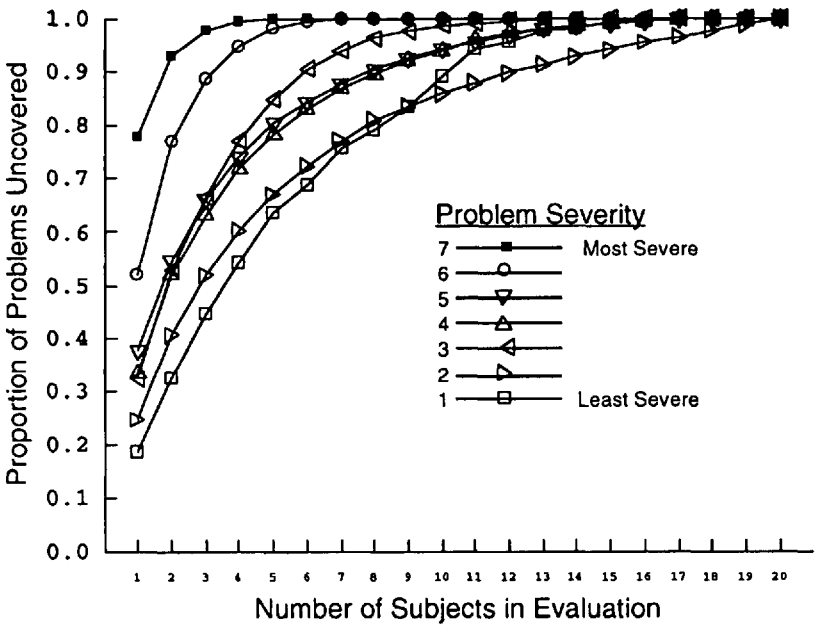

Figure 3. *The proportion of usability problems uncovered is shown as a function of the number of subjects for all problems at a given level of severity. More severe problems tend to be uncovered within the first few subjects.*

double usability expert has experience in both usability testing and the domain of the application.) The rankings made by the panel of experts were then compared with the rankings generated by the experimenters. The panel of experts was not given direct access to either the subjects or the subjects' data and so could not have been influenced by them. We expected the panel of experts' judgments of problem severity to coincide with those of the experimenters, supporting the notion that the experimenters were not unduly influenced by having experience with the subjects.

In addition, in the preceding discussion it was assumed that the goal of a usability evaluation is to *identify* problems, not to *quantify* them. In fact, an evaluator must not only identify usability problems, he or she must decide how severely a problem will affect usability. This, along with cost considerations, forms the basis for deciding if and when a usability problem will be addressed.

When an evaluation is conducted with a small number of subjects, the evaluator cannot rely on frequency data in making judgments of problem severity. We wanted to test whether or not experts could make judgments of problem severity without access to frequency information. The current experiment provides two lines of evidence. First, to the extent that the panel of experts (who do not have access to frequency information) agreed with the experimenters, we would have evidence supporting the claim that experts can judge problem severity as well with frequency information as without. We would not expect perfect agreement, as the panel lacked more than merely frequency information. For example, they were not aware of how severely a particular problem affected any given subject. Also, to the extent that the panel of experts agreed among themselves, we would have evidence that the task provided enough information to be completed

reliably. Although this is a weaker claim, it provides converging evidence if the panel also agreed with the experimenters.

The system evaluated in this experiment was a voice response system in the field that handled several hundred calls per day. Callers interacted with it by pressing the keys on a touch-tone telephone. In response to key presses, callers heard prerecorded messages on a variety of topics such as news, sports, and weather. The evaluation considered the interactive system as well as the design of the supporting documentation.

Method for the Usability Evaluation

Subjects. Twenty subjects were recruited from surrounding communities through an advertisement in a local newspaper. Subjects were paid a small honorarium for their participation in the study.

Tasks. Subjects were asked to complete seven tasks that exercised the full functionality of the system. The tasks included interacting with the voice response system and locating and acting on information in the supporting documentation. Presentation of the tasks was randomized over subjects.

Procedure. Subjects were run individually in the laboratory, seated on a couch in front of a coffee table with a standard desk telephone and the supporting documentation on it. The think-aloud procedure was explained to them prior to the start of the experimental tasks.

Preanalysis. After the 20 subjects had been run, the two experimenters (myself and a coworker) identified 17 distinct usability problems. The problems were assigned ratings of either high, medium, or low in terms of their impact on usability. A three-point rating scale was used instead of the seven-point scale from the previous study.

Subsequently, the videotapes were scored. Problem episodes, as defined in Experiment 2, were identified for each subject, and determination was made as to the usability problem that lay behind it. Thus for each subject we had a complete list of the usability problems that that subject had uncovered.

Method for the Problem Ranking Task

Subjects. Six experts were recruited from GTE who had experience both in the design of interactive voice response systems and in usability evaluation.

Procedure. Brief, one-paragraph descriptions of the 17 usability problems were prepared. A random order for presenting the tasks to the experts was generated. The task descriptions were then distributed to the panel of experts with instructions to rank the problems in terms of how disruptive they were likely to be to the usability of the system. The experts were given a copy of the documentation and had unrestricted access to the system. They were not given any other information. Specifically, they were given no data regarding the frequency with which the problems had occurred, nor were they aware of any problems that had been particularly disruptive for a given subject. Independently, the experimenters also ranked the 17 usability problems on the same basis.

Results

Subjects differed in the number of usability problems that they uncovered. Subjects uncovered between 3 and 12 of the 17 problems (18% and 71%, respectively). Problem detection varied from a low of 1 subject finding a problem to as many as 17 subjects (5% and 85%, respectively). These data are shown in Table 3.

The same Monte Carlo procedure that was used in the previous two studies was applied to the data from the current experiment to generate the general curve relating the number of subjects participating in an evaluation to the proportion of problems revealed. The

TABLE 3

Voice Response System Evaluation: Subjects' Ability to Detect Problems and Probability of Detecting Any Given Problem (percentages)

	M	*SD*	*Min.*	*Max.*
Problems identified per subject	42.0	0.15	18.0	71.0
Subjects uncovering each problem	42.0	0.21	5.0	85.0

resultant curve is shown as the solid line in Figure 4. The dashed line in the figure presents the number of problems expected when the probability of detecting a given problem is taken to be 0.42.

In Figure 5 separate curves are shown for the problems at each of three levels of severity. The figure confirms the effect demonstrated in the previous experiment: as problem severity increases, the likelihood that the problem will be detected within the first few subjects also increases.

I now turn to the results of the expert ranking procedure. The overall goal of this analysis was to measure the level of agreement among the panel of experts and the experimenters. The experimenters may have been influenced through their exposure to the subjects performing the tasks, whereas the experts could not have been influenced. Table 4 presents the matrix of Spearman rank-order correlations for the experimenters' ranking and the rankings of the six independent judges.

Kendall's coefficient of concordance (Hays, 1973) was calculated to assess the overall degree of agreement among the experimenters' rankings and those of the six expert judges. A relatively high degree of concordance was obtained, $W(16) = 0.471$, $p < 0.001$, indicating that all the judges, with or without exposure to the subjects, viewed the problems at comparable levels of severity.

Looking further into the pattern of agreement, one can assess the degree to which the

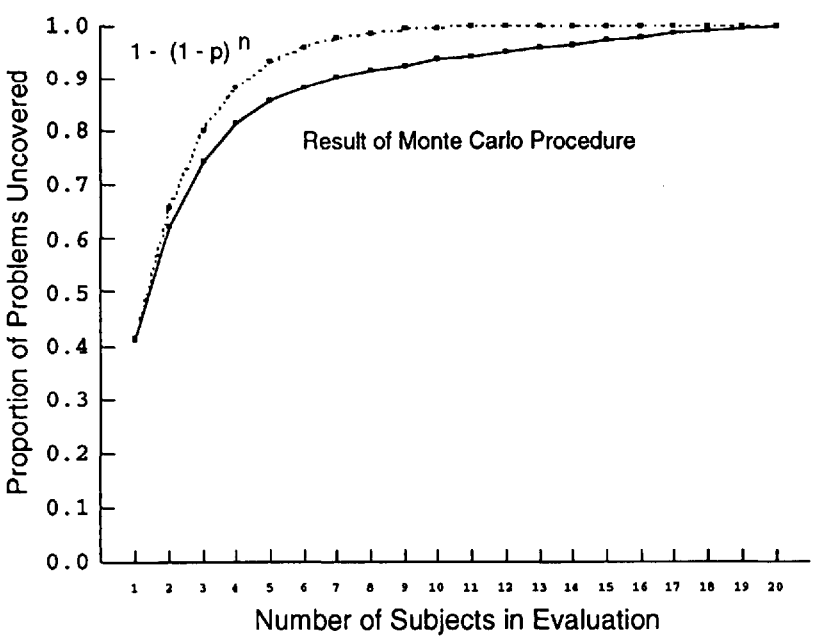

Figure 4. *The proportion of usability problems uncovered is shown as a function of the number of subjects who participated in an evaluation (solid line). The values expected when* p = *0.42 are shown as a broken line.*

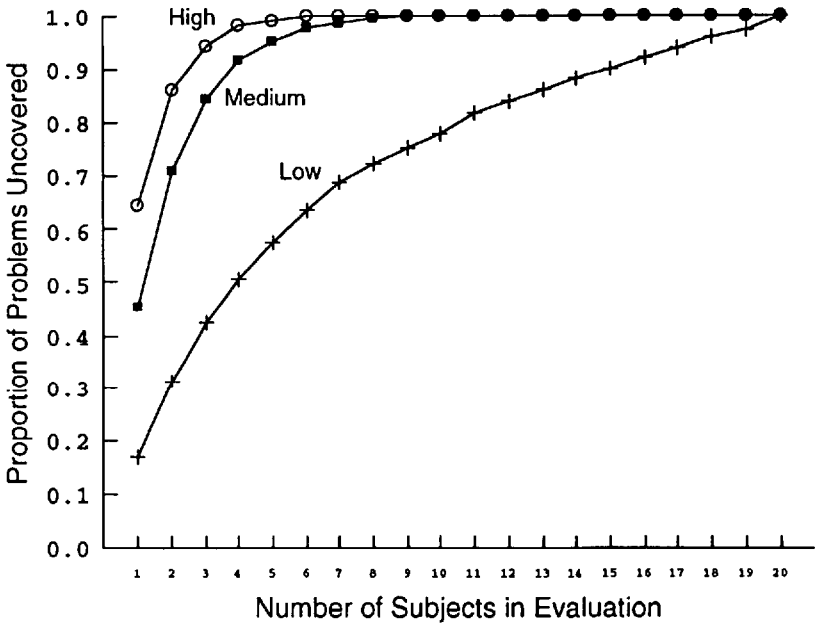

Figure 5. *The proportion of usability problems uncovered is shown as a function of the number of subjects for all problems at a given level of severity. More severe problems tend to be uncovered within the first few subjects.*

panel of experts agreed with the experimenters. The mean r_s correlation between the experimenters' rankings and those of the independent judges was 0.455, indicating that the experts generally agreed with the ranking given by the experimenters. The degree of agreement among the experts can be approximated by examining the mean correlation among all pairs of experts. This mean r_s among the experts was 0.328, confirming a moderate amount of reliability in the rankings.

Discussion

Experiment 3 replicated and extended the results of the previous two studies. Examination of curves relating the proportion of usability problems uncovered to the number of subjects in the evaluation reveals that the law of diminishing returns applies. Most usability problems are discovered with the first few subjects. In the current study the first 5 subjects tended to find about 85% of the usability problems uncovered by running 20 subjects.

TABLE 4

Matrix of Rank-Order Correlations among the Seven Judges (Experimenter and Six Usability Experts)

	Experimenter	*Judge 1*	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*Judge 6*
Experimenter	1.000						
Judge 1	0.213	1.000					
Judge 2	0.723	0.569	1.000				
Judge 3	0.343	0.645	0.453	1.000			
Judge 4	0.534	0.417	0.267	0.289	1.000		
Judge 5	0.304	0.049	0.098	0.225	0.152	1.000	
Judge 6	0.618	0.174	0.483	0.419	0.336	0.341	1.000

The obtained rate of information extraction was, again, slightly below that predicted by probability theory.

When separate curves are plotted for the problems rated high, medium, and low in severity, it can be seen that the problems judged to be more severe are more likely to turn up early in the evaluation. On average, virtually all of the problems judged high in severity were discovered after running 5 subjects, whereas only 55% of the problems judged low in severity were found.

Independent assessment of problem severity was obtained in the current study by having a panel of double usability experts rank the problems in terms of severity. Overall, a high degree of agreement was found in the rankings; about 47% of the total variance possible in the rankings was accounted for. Additionally, agreement between the experimenters' ranking and the independent judges' rankings was relatively high, suggesting that the severity ratings presented in Figure 5 are representative.

The results of this study also support the claim that experts can judge problem severity without frequency information. The experts, who had no access to frequency information, generally agreed with the experimenters, who had conducted the study with more than 20 subjects. An assessment of how reliable the experts were in generating rankings showed moderate agreement, providing converging evidence that experts do not necessarily need frequency information to judge problem severity. This is important because evaluators who choose to run only a few subjects will be forced to make decisions regarding problem severity without frequency information.

GENERAL DISCUSSION

In all of the studies reported, approximately 80% of the usability problems identified would have been found after only five subjects. Important usability problems are more likely to be found with fewer subjects than are less important problems. A practitioner who chooses to run a small number of subjects will identify most of the major usability problems and some proportion of the less important problems. Experts were able to reach consensus regarding the relative severity of problems without benefit of frequency data. Usability experts can assess the severity of a problem without explicit knowledge of how frequent the error is likely to be.

For practitioners in an iterative design cycle, running fewer subjects in each iteration may be a valuable time- and money-saving approach. (Analyses of time spent in the laboratory suggest that up to one third of our development time was being devoted to the test phase.) On subsequent iterations, any major usability problems that were missed in previous test cycles and those that were introduced by ostensible improvements in the user interface will tend to come to the fore.

For the practitioner with only one chance to evaluate a user interface, the results suggest that the size of the evaluation should not be fixed prior to the test. Subjects should be run until the number of new problems uncovered drops to an acceptable level.

Curves relating the proportion of problems detected to the number of subjects in an evaluation are approximated by the formula $1 - (1 - p)^n$, where p is the mean probability of detecting a problem and n is the number of subjects run in the evaluation. This formula will tend to overpredict the amount of information obtained in an evaluation because p is an average probability and Jensen's Inequality states that, for convex functions, the average of some function is greater than or equal to the function of some averages (Royden, 1968). In practice, this difference has been small, and neglecting it will lead to predictions within a few percentage points of those obtained.

A practitioner may use this information in judging how many subjects he or she should run in a usability evaluation. For example, if an evaluator wanted to know how many subjects are required to identify any problem experienced by 10% or more of the population at the 90% confidence level, he or she could use the formula to determine that approximately 22 subjects would be required. If an 80% confidence level were acceptable, the evaluator would be able to determine that approximately 15 subjects are required.

ACKNOWLEDGMENTS

I am grateful to Anthony C. Salvador for making the data reported in Experiment 1 available to me. I would also like to thank David Fay and Greg Cermark for their help in analyzing the probability theory problem discussed in the Experiment 1 results section. Chris Arteberry, Chris Baglieri, and Jamie Katz conducted the study reported in Experiment 2 under my supervision. The usability evaluation in Experiment 3 was conducted by Tony Brown of the Omega Group and myself.

REFERENCES

Bellotti, V. (1988). Implications of current design practice for the use of HCI techniques. In D. Jones and R. Winder (Eds.), *People and computers IV* (pp. 13–34). Cambridge: Cambridge University Press.

Denning, S., Hoiem, D., Simpson, M., and Sullivan, K. (1990). The value of thinking-aloud protocols in industry: A case study at Microsoft. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1285–1289). Santa Monica, CA: Human Factors Society.

Diaconis, P., and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248,* 116–130.

Hays, W. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

Jeffries, R., Miller, J., Wharton, C., and Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the Association for Computing Machinery CHI '91* (pp. 119–124). New York: ACM.

Lewis, C. (1972). Using the "thinking-aloud" method in cognitive interface design (IBM Tech. Report RC 9265 [#40713], 2/17/82). Boca Raton, FL: IBM.

Meister, D. (1989). *Conceptual aspects of human factors.* Baltimore: Johns Hopkins University Press.

Melkus, L. A., and Torres, R. J. (1988). Guidelines for the use of a prototype in user interface design. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 370–374). Santa Monica, CA: Human Factors Society.

Nielsen, J. (1989). Usability engineering at a discount. In G. Salvendy and M. J. Smith (Eds.), *Designing and using human-computer interfaces and knowledge-based systems* (pp. 394–401). Amsterdam: Elsevier.

Nielsen, J. (1990). Evaluating the thinking-aloud technique for use by computer scientists. In H. Hartson and D. Hix (Eds.), *Advances in human-computer interaction* (Vol. 3, pp. 197–216). Norwood, NJ: Ablex.

Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the Association for Computing Machinery CHI '90* (pp. 249–256). New York: ACM.

Royden, H. (1968). *Real analysis* (2nd ed.). New York: Macmillan.

Virzi, R. (1989). What can you learn from a low-fidelity prototype? In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 224–228). Santa Monica, CA: Human Factors Society.

Virzi, R. (1990). Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 291–294). Santa Monica, CA: Human Factors Society.