

Sample Sizes for Usability Studies: Additional Considerations

JAMES R. LEWIS,¹ *International Business Machines, Inc., Boca Raton, Florida*

Recently, Virzi (1992) presented data that support three claims regarding sample sizes for usability studies: (1) observing four or five participants will allow a usability practitioner to discover 80% of a product's usability problems, (2) observing additional participants will reveal fewer and fewer new usability problems, and (3) more severe usability problems are easier to detect with the first few participants. Results from an independent usability study clearly support the second claim, partially support the first, but fail to support the third. Problem discovery shows diminishing returns as a function of sample size. Observing four to five participants will uncover about 80% of a product's usability problems as long as the average likelihood of problem detection ranges between 0.32 and 0.42, as in Virzi. If the average likelihood of problem detection is lower, then a practitioner will need to observe more than five participants to discover 80% of the problems. Using behavioral categories for problem severity (or impact), these data showed no correlation between problem severity (impact) and rate of discovery. The data provided evidence that the binomial probability formula may provide a good model for predicting problem discovery curves, given an estimate of the average likelihood of problem detection. Finally, data from economic simulations that estimated return on investment (ROI) under a variety of settings showed that only the average likelihood of problem detection strongly influenced the range of sample sizes for maximum ROI.

INTRODUCTION

The goal of many usability studies is to identify design problems and recommend product changes (to either the current product or future products) based on the design problems (Gould, 1988; Grice and Ridgway, 1989; Karat, Campbell, and Fiegel, 1992; Whitefield and Sutcliffe, 1992; Wright and Monk, 1991). During a usability study, an ob-

server watches representative participants perform representative tasks to understand when and how they have problems using a product. The problems provide clues about how to redesign the product to either eliminate the problem or provide easy recovery from it (Lewis and Norman, 1986; Norman, 1983).

Human factors engineers who conduct industrial usability evaluations need to understand their sample size requirements. If they collect a larger sample than necessary, they

¹ Requests for reprints should be sent to James R. Lewis, IBM Corp., P.O. Box 1328, Boca Raton, FL 33429-1328.

might increase product cost and development time. If they collect too small a sample, they might fail to detect problems that, uncorrected, would reduce the usability of the product. Discussing usability testing, Keeler and Denning (1991) showed a common negative attitude toward small-sample usability studies when they stated, "actual [usability] test procedures cut corners in a manner that would be unacceptable to true empirical investigations. Test groups are small (between 6 and 20 subjects per test)" (p. 290). Yet, in any setting, not just an industrial one, the appropriate sample size accomplishes the goals of the study as efficiently as possible (Kraemer and Thiemann, 1987).

Virzi (1992) investigated sample size requirements for usability evaluations. He reported three experiments in which he measured the rate at which trained usability experts identified problems as a function of the number of naïve participants they had observed. For each experiment, he ran a Monte Carlo simulation to permute participant orders 500 times and measured the cumulative percentage of problems discovered for each sample size. In the second experiment, the observers provided ratings of problem severity (using a seven-point scale). In addition to having the observers provide problem severity ratings (this time using a three-point scale), an independent set of usability experts provided estimates of problem severity based on brief, one-paragraph descriptions of the problems discovered in the third experiment.

This helped to control the effect of knowledge of problem frequency on the estimation of problem severity. The correlation between problem frequency and test observers' judgment of severity in the second experiment was 0.463 ($p < 0.01$). In the third experiment, agreement among the test observers and the independent set of judges was significant, $W(16) = 0.471$, $p < 0.001$, for the rank order of 17 problems in terms of how disruptive they were likely to be to the usability of the system. (Virzi did not report the magnitude of the correlation between problem frequency and either the test observers' or independent judges' estimates of problem severity for the third experiment.) Table 1 shows some of the key features of the three experiments.

Based on these experiments, Virzi (1992) made three claims regarding sample size for usability studies: (1) Observing four or five participants will allow a practitioner to discover 80% of a product's usability problems, (2) observing additional participants will reveal fewer and fewer new usability problems, and (3) more severe usability problems are easier to detect with the first few participants. These important findings are in need of replication. One purpose of this paper is to report the results of an independent usability study that clearly support the second claim, partially support the first, and fail to support the third. Another purpose is to develop a mathematical model of problem discovery based on the binomial probability formula and examine its extension into economic

TABLE 1

Key Features of Virzi's (1992) Three Experiments

| <i>Experiment</i> | <i>Sample Size</i> | <i>Number of Tasks</i> | <i>Number of Problems</i> | <i>Average Likelihood of Problem Detection</i> |
|-------------------|--------------------|------------------------|---------------------------|--|
| 1 | 12 | 3 | 13 | 0.32 |
| 2 | 20 | 21 | 40 | 0.36 |
| 3 | 20 | 7 | 17 | 0.42 |

simulations that estimate return on investment (ROI) for a usability study as a function of several independent variables.

THE OFFICE APPLICATIONS USABILITY STUDY

Lewis, Henry, and Mack (1990) conducted a series of usability studies to develop usability benchmark values for integrated office systems. The following method and results are from one of these studies (the only one for which we kept a careful record of which participants experienced which problems). A set of 11 scenarios served as stimuli in the evaluation.

Method

Participants. Fifteen employees of a temporary help agency participated in the study. All participants had at least three months' experience with a computer system but had no programming training or experience. Five participants were clerks or secretaries with no experience in the use of a mouse device, five were business professionals with no mouse experience, and five were business professionals who did report mouse experience.

Apparatus. The office system had a word processor, a mail application, a calendar application, and a spreadsheet on an operating system that allowed a certain amount of integration among the applications.

Procedure. A participant began with a brief tour of the lab, read a description of the purpose of the study, and completed a background questionnaire. After a short tutorial on the operating system, the participant began working on the scenarios. It usually took a participant about 6 h to complete the scenarios. Observers watched participants, one at a time, by closed-circuit television. In addition to several performance measures, observers carefully recorded the problems that

participants experienced during the study. They classified the problems in decreasing level of impact according to four behavioral definitions:

1. *Scenario failure.* The problem caused the participant to fail to complete a scenario by either requiring assistance to recover from the problem or producing an incorrect output (excluding minor typographical errors).
2. *Considerable recovery effort.* The participant either worked on recovery from the problem for more than a minute or experienced the problem multiple times within a scenario.
3. *Minor recovery effort.* The participant experienced the problem only once within a scenario and required less than a minute to recover.
4. *Inefficiency.* The participant worked toward the scenario's goal but deviated from the most efficient path.

Results

Participants experienced 145 different problems during the usability study. The average likelihood of problem detection was 0.16. Figure 1 shows the results of applying a Monte Carlo procedure to calculate the mean of 500 permutations of participant orders, revealing the general form of the cumulative problem discovery curve. Figure 1 also shows the predicted cumulative problem discovery curve using the formula $1 - (1 - p)^n$ (Virzi, 1990, 1992; Wright and Monk, 1991), where p is the probability of detecting a given problem and n is the sample size. The predicted curve shows an excellent fit to the Monte

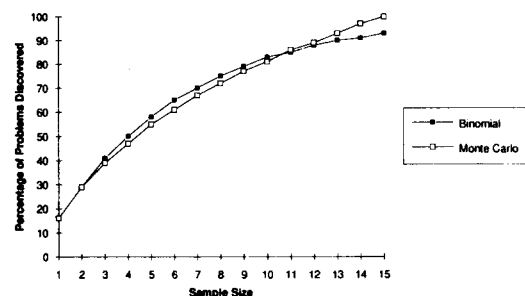


Figure 1. Predicted problem discovery as a function of sample size.

Carlo curve, Kolmogorov-Smirnov $J'_3 = 0.73$, $p = 0.66$ (Hollander and Wolfe, 1973). For this study, observing five participants would uncover only 55% of the problems. To uncover 80% of the problems would require 10 participants.

Different participants might experience the same problem but might not experience the same impact. For subsequent data analyses, the impact rating for each problem was the modal impact level across the participants who experienced the problem. (If the distribution was bimodal, then the problem received the more severe classification.) Figure 2 shows the results of applying the same Monte Carlo procedure to problems for each of the four impact levels. The curves overlap considerably, and the Pearson correlation between problem frequency and impact level was not significant, $r(143) = 0.06$, $p = 0.48$.

Discussion

These results are completely consistent with the earlier finding that additional participants discover fewer and fewer problems (Virzi, 1992). If the average likelihood of problem detection had been in the range of 0.32 to 0.42, then five participants would have been enough to uncover 80% of the problems. However, because the average likelihood of problem detection was considerably lower in this study than in the three Virzi

studies, usability studies such as this would need 10 participants to discover 80% of the problems. This shows that it is important for usability evaluators to have an idea about the average likelihood of problem detection for their types of products and usability studies before they estimate sample size requirements. If a product has poor usability (has a high average likelihood of problem detection), it is easy to improve the product (or at least discover a large percentage of its problems) with a small sample. However, if a product has good usability (has a low average likelihood of problem detection), it will require a larger sample to discover the remaining problems.

These results showed no significant relationship between problem frequency and problem impact. This outcome failed to support the claim that observers would find severe usability problems faster than they would less severe problems (Virzi, 1992). Virzi used the term *problem severity*, and my colleagues and I (Lewis et al., 1990) described the dimension as *problem impact*. Our conception of problem severity was that it is the combination of the effects of problem impact and problem frequency. Because the impact for a problem in this usability study was assignment to behaviorally defined categories, this impact classification should be independent of problem frequency.

In his third experiment, Virzi attempted to control the confounding caused by having the observers, who have problem frequency knowledge, also rate severity (using a three-point scale). The observers and an independent group of usability experts ranked 17 of the problems along the dimension of disruptiveness to system usability, and the results indicated significant agreement, $W(16) = 0.471$, $p < 0.001$. However, this procedure (providing one-paragraph problem descriptions to the independent group of experts) might not have successfully removed the

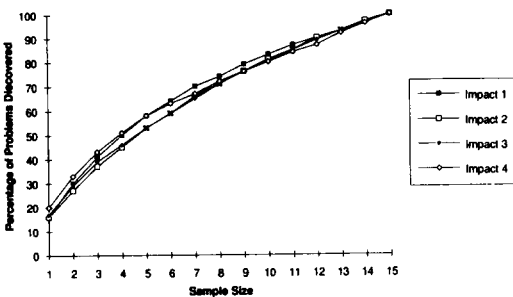


Figure 2. Problem discovery rate as a function of problem impact.

influence of problem frequency from severity estimation.

It is unfortunate that Virzi did not report the magnitude of the correlation between problem frequency and the severity judgments of the usability experts who did not have any knowledge of problem frequency. Given these conflicting results and the logical independence of problem impact (or severity) and frequency, human factors engineers and others who conduct usability evaluations should take the conservative approach of assuming no relationship between problem impact and frequency until future research resolves the different outcomes.

PROBLEM DISCOVERY CURVES AND THE BINOMIAL PROBABILITY FORMULA

Several researchers have suggested that the formula $1 - (1 - p)^n$ predicts the rate of problem discovery in usability studies (Virzi, 1990, 1992; Wright and Monk, 1991). However, none of these researchers has offered an explanation of the basis for that formula. In an earlier paper (Lewis, 1982), I proposed that the binomial probability theorem could provide a statistical model for determining the likelihood of detecting a problem of probability p , r times, in a study with n participants.

The binomial probability formula is $P(r) = \binom{n}{r} p^r (1 - p)^{n-r}$ (Bradley, 1976), where $P(r)$ is the likelihood that an event will occur r times, given a sample size of n and the probability p that the event will occur in the population at large. The conditions under which the binomial probability formula applies are random sampling, independent observations, two mutually exclusive and exhaustive categories of events, and sample observations that do not deplete the source. Problem discovery usability studies usually meet these conditions. Usability practitioners should attempt to sample participants randomly. (Although circumstances rarely allow true

random sampling in usability studies, experimenters do not usually exert any influence on precisely who participates in the study, resulting in a quasi-random sampling.)

Observations among participants are independent, because the problems experienced by one participant cannot have an effect on those experienced by another participant. (Note that this model does not require independence among the different types of problems that occur.) The two mutually exclusive and exhaustive problem detection categories are (1) the participant encountered the problem and (2) the participant did not experience the problem.

Finally, the sampled observations in a usability study do not deplete the source. The probability that a given sample size will produce at least one instance of problem detection is 1 minus the probability of no detections, or $1 - P(0)$. When $r = 0$, $P(0) = \binom{n}{0} p^0 (1 - p)^{n-0}$, which reduces to $P(0) = (1 - p)^n$. Thus the cumulative binomial probability for the likelihood that a problem of probability p will occur at least once is $1 - (1 - p)^n$.

Problem Discovery Curves for Specific Problems of Varying Probabilities

As shown in Figure 1, the formula $1 - (1 - p)^n$ provides a good fit to Monte Carlo estimations where p is the average likelihood of problem detection for a set of problems. Another approach is to select specific problems of varying probabilities of detection and compare Monte Carlo problem discovery curves with curves predicted with the cumulative binomial probability formula. Figure 3 shows the problem discovery likelihoods (each based on 500 Monte Carlo participant-order permutations) for five specific problems, with problem probabilities ranging from 0.14 to 0.74. The figure also shows the predicted cumulative problem discovery curves.

As seen in Figure 3, $1 - (1 - p)^n$ provides an excellent fit to the results of Monte Carlo

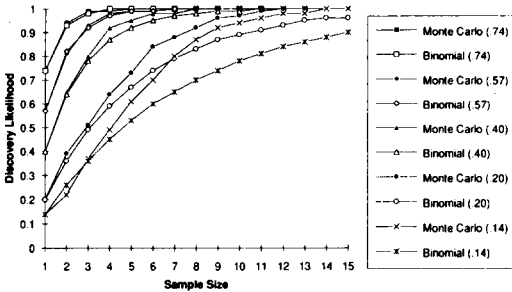


Figure 3. Predicted problem discovery rates as a function of individual problem likelihood.

permutations. For low-probability problems (0.14, 0.20, and 0.40) the Kolmogorov-Smirnov J'_3 was 0.54 ($p = 0.93$), and for high-probability problems (0.57 and 0.74) J'_3 was 0.18 ($p = 1.00$). Figure 3 shows that the binomial and Monte Carlo curves deviated more when problem probability was low. This is probably because the curves based on the Monte Carlo simulations must end at a discovery likelihood of 1.0, but the binomial curves do not have this constraint. A sample size of 15 was probably not sufficient to demonstrate perfectly accurate problem discovery curves for low-probability problems using Monte Carlo permutations.

Note also that the binomial curves for high-probability problems (0.57 and 0.74) matched the Monte Carlo curve very closely, but low-probability problems (0.14, 0.20, and 0.40) underpredicted the Monte Carlo curve. This lends support to Virzi's (1992) suggestion that the tendency of the formula $1 - (1 - p)^n$ to overpredict problem discovery for sets of problems is a Jensen's Inequality artifact.

Jensen's Inequality is a general inequality satisfied by a convex function:

$$f\left(\sum_{i=1}^n a_i x_i\right) \leq \sum_{i=1}^n a_i f(x_i),$$

where x_i is any number in the region where f is convex, and a_i is nonnegative and sums to

one (Parker, 1984). Because any a_i can equal $1/n$, the formula applies to the arithmetic mean. Applied to the data in this paper, the function of a mean (such as the average of a series of Monte Carlo trials) will be less than or equal to the mean of a function (such as p averaged over a set of problems, then placed into the binomial probability formula).

Detecting Problems at Least Twice

If the binomial probability formula is a reasonable model for problem discovery in usability studies, then it should also predict the likelihood that a problem will occur at least twice. (In practice, some usability practitioners use this criterion to avoid reporting problems that might be idiosyncratic to a single participant.) The cumulative binomial probability for P (at least two detections) is $1 - [P(0) + P(1)]$. Because $P(1) = np(1 - p)^{n-1}$, P (at least two detections) = $1 - [(1 - p)^n + np(1 - p)^{n-1}]$.

Figure 4 shows the Monte Carlo (based on 500 participant-order permutations) and binomial problem discovery curves for the likelihood that a problem will occur at least twice (Lewis et al., 1990). A Kolmogorov-Smirnov goodness-of-fit test revealed that the binomial probability formula did not provide an adequate fit to the Monte Carlo data, $J'_3 = 1.27$, $p = 0.08$. However, the average likelihood of detecting a problem at least twice was quite low in this study (0.03), so it

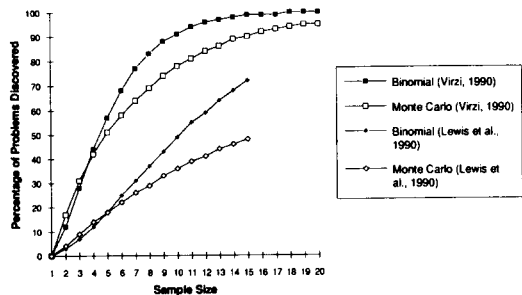


Figure 4. Predicted problem discovery rates for detecting problems at least twice.

was possible that a sample size of 15 might not be adequate to model this problem discovery situation. With data reported by Virzi (1990), Figure 4 also shows the Monte Carlo (based on 500 participant-order permutations) and binomial problem discovery curves for a usability study in which the average likelihood of detecting a problem at least twice was 0.12 and there were 20 participants. In that situation, the Kolmogorov-Smirnov goodness-of-fit test provided strong support for prediction based on the binomial probability formula, $J'_3 = 0.47$, $p = 0.98$.

Discussion

These data provide support for the hypothesis that the cumulative binomial probability formula is a reasonable model for problem discovery in usability studies. To help human factors engineers select an appropriate sample size for a usability study, Table 2 shows the expected proportion of detected problems

(at least once) for various problem detection probabilities through a sample size of 20. Table 3 shows the minimum sample size required to detect problems of varying probabilities at least once (or, as shown in parentheses, at least twice). When the problem probability is the average across a set of problems, then the cumulative likelihood that the problem will occur is also the expected proportion of discovered problems.

For example, if a practitioner planned to discover problems from a set with an average probability of detection of 0.25, was willing to treat a single detection of a problem seriously, and planned to discover 90% of the problems, the study would require 8 participants. If the practitioner planned to see a problem at least twice before taking it seriously, the sample size requirement would be 14. If a practitioner planned to discover problems at least once with probabilities as low as 0.01 and with a cumulative likelihood

TABLE 2

Expected Proportion of Detected Problems (at Least Once) for Various Problem Detection Probabilities and Sample Sizes

| Sample Size | $p = 0.01$ | $p = 0.05$ | $p = 0.10$ | $p = 0.15$ | $p = 0.25$ | $p = 0.50$ | $p = 0.90$ |
|-------------|------------|------------|------------|------------|------------|------------|------------|
| 1 | 0.01 | 0.05 | 0.10 | 0.15 | 0.25 | 0.50 | 0.90 |
| 2 | 0.02 | 0.10 | 0.19 | 0.28 | 0.44 | 0.75 | 0.99 |
| 3 | 0.03 | 0.14 | 0.27 | 0.39 | 0.58 | 0.88 | 1.00 |
| 4 | 0.04 | 0.19 | 0.34 | 0.48 | 0.68 | 0.94 | 1.00 |
| 5 | 0.05 | 0.23 | 0.41 | 0.56 | 0.76 | 0.97 | 1.00 |
| 6 | 0.06 | 0.26 | 0.47 | 0.62 | 0.82 | 0.98 | 1.00 |
| 7 | 0.07 | 0.30 | 0.52 | 0.68 | 0.87 | 0.99 | 1.00 |
| 8 | 0.08 | 0.34 | 0.57 | 0.73 | 0.90 | 1.00 | 1.00 |
| 9 | 0.09 | 0.37 | 0.61 | 0.77 | 0.92 | 1.00 | 1.00 |
| 10 | 0.10 | 0.40 | 0.65 | 0.80 | 0.94 | 1.00 | 1.00 |
| 11 | 0.10 | 0.43 | 0.69 | 0.83 | 0.96 | 1.00 | 1.00 |
| 12 | 0.11 | 0.46 | 0.72 | 0.86 | 0.97 | 1.00 | 1.00 |
| 13 | 0.12 | 0.49 | 0.75 | 0.88 | 0.98 | 1.00 | 1.00 |
| 14 | 0.13 | 0.51 | 0.77 | 0.90 | 0.98 | 1.00 | 1.00 |
| 15 | 0.14 | 0.54 | 0.79 | 0.91 | 0.99 | 1.00 | 1.00 |
| 16 | 0.15 | 0.56 | 0.81 | 0.93 | 0.99 | 1.00 | 1.00 |
| 17 | 0.16 | 0.58 | 0.83 | 0.94 | 0.99 | 1.00 | 1.00 |
| 18 | 0.17 | 0.60 | 0.85 | 0.95 | 0.99 | 1.00 | 1.00 |
| 19 | 0.17 | 0.62 | 0.86 | 0.95 | 1.00 | 1.00 | 1.00 |
| 20 | 0.18 | 0.64 | 0.88 | 0.96 | 1.00 | 1.00 | 1.00 |

TABLE 3

Sample Size Requirements as a Function of Problem Detection Probability and the Cumulative Likelihood of Detecting the Problem at Least Once (Twice)

| Problem Detection Probability | Cumulative Likelihood of Detecting the Problem at Least Once (Twice) | | | | | |
|-------------------------------|--|-----------|-----------|-----------|-----------|-----------|
| | 0.50 | 0.75 | 0.85 | 0.90 | 0.95 | 0.99 |
| 0.01 | 68 (166) | 136 (266) | 186 (332) | 225 (382) | 289 (462) | 418 (615) |
| 0.05 | 14 (33) | 27 (53) | 37 (66) | 44 (76) | 57 (91) | 82 (121) |
| 0.10 | 7 (17) | 13 (26) | 18 (33) | 22 (37) | 28 (45) | 40 (60) |
| 0.15 | 5 (11) | 9 (17) | 12 (22) | 14 (25) | 18 (29) | 26 (39) |
| 0.25 | 3 (7) | 5 (10) | 7 (13) | 8 (14) | 11 (17) | 15 (22) |
| 0.50 | 1 (3) | 2 (5) | 3 (6) | 4 (7) | 5 (8) | 7 (10) |
| 0.90 | 1 (2) | 1 (2) | 1 (3) | 1 (3) | 2 (3) | 2 (4) |

Note. These are the minimum sample sizes that result after rounding cumulative likelihoods to two decimal places. Strictly speaking, therefore, the cumulative probability for the 0.50 column is 0.495, and that for the 0.75 column is 0.745, and so on. If a practitioner requires greater precision, the method described in the paper will allow the calculation of a revised sample size, which will always be equal to or greater than the sample sizes in this table. The discrepancy will increase as problem probability decreases, cumulative probability increases, and the number of times a problem must be detected increases.

of discovery of 0.99, the study would require 418 participants (an unrealistic requirement in most settings, implying unrealistic study goals).

A RETURN-ON-INVESTMENT MODEL FOR USABILITY STUDIES

The preceding analyses show that, given an estimate of the average likelihood of problem detection, it is possible to generate problem discovery curves with the cumulative binomial probability distribution. These curves provide a basis for selecting an appropriate sample size for usability studies. However, a more complete analysis should address the costs associated with running additional participants, fixing problems, and failing to discover problems. Such an analysis should allow usability practitioners to specify the relationship between sample size and return on investment.

Method

Six variables were manipulated in ROI simulations to determine those variables that exert influence on (1) the sample size at the maximum ROI, (2) the magnitude of the maximum ROI, and (3) the percentage of prob-

lems discovered at the maximum ROI. (Table 4 lists the variables and their values.) The equation for the simulations was $ROI = Savings/Costs$, where Savings is the cost of the discovered problems had they remained undiscovered, minus the cost of fixing the discovered problems, and Costs is the sum of the daily cost to run a study, plus the costs associated with problems that remain undiscovered. Thus a better ROI will have a higher numerical value.

The simulations included cumulative binomial problem discovery curves for sample sizes from 1 to 20 for three average likelihoods of problem discovery (0.10, 0.25, and 0.50). For each sample size and average likelihood of problem discovery, a BASIC program provided the expected number of discovered and undiscovered problems. The program then crossed the discovered problem cost of \$100 with undiscovered problem costs of \$200, \$500, and \$1000 (low set), and the discovered problem cost of \$1000 with undiscovered problem costs of \$2000, \$5000, and \$10,000 (high set) to calculate ROIs.

For the simulations, the sample size variable of 20 participants covered a reasonable range and should result in the discovery of a

TABLE 4

Main Effects for the ROI Simulations

| Independent Variables | | Dependent Variables | | |
|--|--------|----------------------------|--------------------------|--|
| Variable | Value | Sample Size at Maximum ROI | Magnitude of Maximum ROI | Percentage of Problems Discovered at Maximum ROI |
| Average likelihood of problem discovery | 0.10 | 19.0 | 3.1 | 86 |
| | 0.25 | 14.6 | 22.7 | 97 |
| | 0.50 | 7.7 | 52.9 | 99 |
| | Range: | 11.3 | 49.8 | 13 |
| Number of problems available for discovery | 30 | 11.5 | 7.0 | 91 |
| | 150 | 14.4 | 26.0 | 95 |
| | 300 | 15.4 | 45.6 | 95 |
| | Range: | 3.9 | 38.6 | 4 |
| Daily cost to run study | 500 | 14.3 | 33.4 | 94 |
| | 1000 | 13.2 | 19.0 | 93 |
| | Range: | 1.1 | 14.4 | 1 |
| Cost to fix a discovered problem | 100 | 11.9 | 7.0 | 92 |
| | 1000 | 15.6 | 45.4 | 96 |
| | Range: | 3.7 | 38.4 | 4 |
| Cost of an undiscovered problem (low set) | 200 | 10.2 | 1.9 | 89 |
| | 500 | 12.0 | 6.4 | 93 |
| | 1000 | 13.5 | 12.6 | 94 |
| | Range: | 3.3 | 10.7 | 5 |
| Cost of an undiscovered problem (high set) | 2000 | 14.7 | 12.3 | 95 |
| | 5000 | 15.7 | 41.7 | 96 |
| | 10000 | 16.4 | 82.3 | 96 |
| | Range: | 1.7 | 70.0 | 1 |

large proportion of the problems that are available for discovery in many usability studies (Virzi, 1992). The values for the number of problems available for discovery are consistent with those reported in the literature (Lewis et al., 1990; Virzi, 1990, 1992), as are the values for the average likelihood of problem discovery. Assuming one participant per day, the values for the daily cost to run a study are consistent with current laboratory, observer, and participant costs. The ratio of the costs to fix a discovered problem to the costs of an undiscovered problem are congruent with software engineering indexes reported by Boehm (1981).

Results

Table 4 shows the results of the main effects of the independent variables in the simulations on the dependent variables of (1) the sample size at which the maximum ROI occurred, (2) the magnitude of the maximum

ROI, and (3) the percentage of problems discovered at the maximum ROI. The table shows the average value of each dependent variable for each level of all the independent variables, and the range of the average values for each independent variable. Across all the variables, the average percentage of discovered problems at the maximum ROI was 94%.

Discussion

All of the independent variables influenced the sample size at the maximum ROI, but the variable with the broadest influence (as indicated by the range) was the average likelihood of problem discovery (p). It also had the strongest influence on the percentage of problems discovered at the maximum ROI. Therefore, it is very important for usability practitioners to estimate the magnitude of this variable for their studies because it largely determines the appropriate sample size. If

the expected value of p is small (for example, 0.10), practitioners should plan to discover about 86% of the problems. If the expected value of p is larger (for example, 0.25 or 0.50), practitioners should plan to discover about 98% of the problems. If the value of p is between 0.10 and 0.25, practitioners should interpolate in Table 4 to determine an appropriate goal for the percentage of problems to discover.

Contrary to expectation, the cost of an undiscovered problem had a minor effect on sample size at maximum ROI, but, like all the other independent variables, it had a strong effect on the magnitude of the maximum ROI. Usability practitioners should be aware of these costs and their effect on ROI, but these costs have relatively little effect on the appropriate sample size for a usability study.

The definitions of the cost variables for the ROI simulations are purposely vague. Each practitioner needs to consider the potential elements of cost for a specific work setting. For example, the cost of an undiscovered problem in one setting might consist primarily of the cost to send personnel to user locations to repair the problem. In another setting the primary cost of an undiscovered problem might be the loss of future sales resulting from customer dissatisfaction.

GENERAL DISCUSSION

The law of diminishing returns, based on the cumulative binomial probability formula, applies to problem discovery usability studies. To use this formula to determine an appropriate sample size, practitioners must form an idea about the expected value of p (the average likelihood of problem detection) for the study and the percentage of problems that the study should uncover. Practitioners can use the data in Table 4 or their own ROI formulas to estimate an appropriate goal for the percentage of problems to discover and can examine data from their own or pub-

lished usability studies to estimate p . (The data from this office applications study have shown that p can be as low as 0.16.) With these two estimates, Table 3 (or, more generally, the cumulative binomial probability distribution) can provide the appropriate sample size for the usability study.

Practitioners who wait to see a problem at least twice before giving it serious consideration can see in Table 3 the sample size implications of this strategy. Certainly, all other things being equal, it is more important to correct a problem that occurs frequently than one that occurs infrequently. However, it is unrealistic to assume that the frequency of detection of a problem is the only criterion to consider in the analysis of usability problems. The best strategy is to consider problem frequency and impact simultaneously to determine which problems are most important to correct rather than establishing a cutoff rule such as "fix every problem that appears two or more times."

The results of the present office applications usability study raise a serious question about the relationship between problem frequency and impact (or severity). In this study, problem discovery rates were the same regardless of the problem impact rating. Clearly, the conservative approach for practitioners is to assume independence of frequency and impact until future research resolves the discrepancy in findings between this office applications study and the studies reported by Virzi (1992).

It is important for practitioners to consider the risks as well as the gains when they use small samples in usability studies. Although the diminishing returns for inclusion of additional participants strongly suggest that the most efficient approach is to run a small sample (especially if the expected p is high, if the study will be iterative, and if undiscovered problems will not have dangerous or expensive outcomes), human factors engineers and

other usability practitioners must not become complacent regarding the risk of failing to detect low-frequency but important problems.

The goal of this paper was to address considerations for the selection of a sample size of participants for problem discovery usability studies. However, this is only one element among several that usability practitioners must consider. Another important topic is the selection and construction of the tasks and scenarios that participants encounter in a study. Certainly what an evaluator asks participants to do influences the likelihood of problem discovery. If the likelihood of discovery of a specific problem on a single performance of a task is low, the likelihood of discovery will increase if participants have multiple opportunities to perform the task (or variations on the task). Repeating tasks also allows an evaluator to determine if particular problems that occur early in a participant's experience with a system diminish or persist with practice.

Conversely, repeating tasks increases study time. The decision about whether to have multiple trials depends on the purpose of the study. Concerns about what tasks to ask participants to do is similar to the problem of assessing content validity in psychometrics (Nunnally, 1978). This topic (adequate task coverage in usability studies) deserves more detailed treatment.

ACKNOWLEDGMENTS

I thank my colleagues in the IBM Human Factors group and the *Human Factors* reviewers for their helpful comments concerning this work. In particular, I thank Robert J. Wherry, Jr., who, in reviewing the first draft of this paper, wrote his own BASIC programs to recreate my tables. In doing so, he uncovered an error in my ROI simulation program and prevented me from publishing inaccurate data. This was truly reviewing beyond the call of duty, and I greatly appreciate it.

REFERENCES

- Boehm, B. W. (1981). *Software engineering economics*. Englewood Cliffs, NJ: Prentice-Hall.
- Bradley, J. V. (1976). *Probability; decision; statistics*. Englewood Cliffs, NJ: Prentice-Hall.
- Gould, J. D. (1988). How to design usable systems. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 757-789). New York: North-Holland.
- Grice, R. A., and Ridgway, L. S. (1989). A discussion of modes and motives for usability evaluation. *IEEE Transactions on Professional Communications*, 32, 230-237.
- Hollander, M., and Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.
- Karat, C. M., Campbell, R., and Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Human factors in computing systems: CHI '92 conference proceedings* (pp. 397-404). New York: Association for Computing Machinery.
- Keeler, M. A., and Denning, S. M. (1991). The challenge of interface design for communication theory: From interaction metaphor to contexts of discovery. *Interacting with Computers*, 3, 283-301.
- Kraemer, H. C., and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
- Lewis, J. R. (1982). Testing small system customer set-up. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 718-720). Santa Monica, CA: Human Factors and Ergonomics Society.
- Lewis, J. R., Henry, S. C., and Mack, R. L. (1990). Integrated office software benchmarks: A case study. In *Human-Computer Interaction—INTERACT '90* (pp. 337-343). London: Elsevier.
- Lewis, C., and Norman, D. A. (1986). Designing for error. In D. A. Norman and S. W. Draper (Eds.), *User-centered system design: New perspectives on human-computer interaction* (pp. 411-432). Hillsdale, NJ: Erlbaum.
- Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM*, 4, 254-258.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Parker, S. P. (1984). *Dictionary of scientific and technical terms*. New York: McGraw-Hill.
- Virzi, R. A. (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 and Ergonomics Society.
- Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, 457-468.
- Whitefield, A., and Sutcliffe, A. (1992). Case study in human factors evaluation. *Information and Software Technology*, 34, 443-451.
- Wright, P. C., and Monk, A. F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*, 35, 891-912.

Date received: November 2, 1992

Date accepted: October 7, 1993