Problem Discovery in Usability Studies: A Model Based on the Binomial Probability Formula

James R. Lewis

IBM Human Factors Group, P. O. Box 1328, Boca Raton, FL 33429-1328

Abstract

Product developers want their products to be as easy to use as possible, but must consider constraints such as cost and schedule. The primary goal of many usability studies is to discover design problems. After discovery, designers can take steps to eliminate or minimize problem impact. This paper shows that problem discovery in usability studies is consistent with the binomial probability formula. The problem discovery curves from two recent studies lend empirical support to this problem discovery model. One practical application of the model is to help estimate appropriate sample sizes for problem discovery usability studies. This model can help usability researchers simultaneously consider cost (minimized by running as small a sample as possible) and risk (minimized by running as large a sample as possible) to maximize the efficiency of a study.

1. 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 observer 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 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, p. 290) 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)." Yet, in any setting, not just an industrial one, the appropriate sample size is the one that accomplishes the goals of the study as efficiently as possible (Kraemer and Thiemann, 1987).

2. PROBLEM DISCOVERY AND BINOMIAL PROBABILITIES

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 that one participant experiences do not 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-discovery categories are (1) the participant encountered the problem during the study and (2) the participant did not experience the problem during the study. Finally, the sampled observations do not deplete the source. See Figure 1 for cumulative problem discovery likelihoods for problems with

probabilities that range from .01 to .90 and for sample sizes ranging from 1 to 20. (Also, see Wright and Monk, 1991.)

The probability that a given sample size will lead to at least one instance of problem occurrence in a study is 1 minus the probability of no occurrences, 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 $1-(1-p)^n$. For example, if a set of instructions in a user's manual will, on average, be confusing to 50% of the user population, the likelihood that the instructions will confuse one participant is .5. If two representative users participate, the probability of confusing either one or both participants is .75. If three users participate, the likelihood that at least one of them will be confused is .88 (as shown in Figure 1).

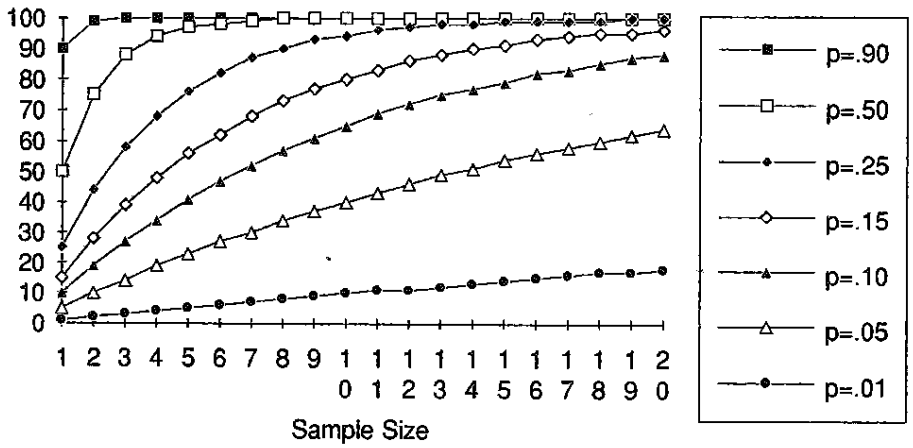

Figure 1. Predicted problem discovery percentage as a function of problem likelihood and sample size.

Given that the cumulative binomial probability formula is a reasonable model for problem discovery, the entries in Table 1 show the minimum sample size required to detect problems of varying probabilities at least once. 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 occurrence of .25 and planned to discover 90% of the problems, the study would require eight participants. If a practitioner planned to discover problems at least once with probabilities as low as .01 and with a cumulative likelihood of discovery of .99, the study

would require 418 participants. A need for a sample size of this magnitude shows that these study requirements are probably not realistic.

Table 1
Sample size requirements as a function of problem probability and the cumulative likelihood of detecting the problem at least once

Problem Probability	Cumulative likelihood of detecting the problem at least once					
	.50	.75	.85	.90	.95	.99
.01	68	136	186	225	289	418
.05	14	27	37	44	57	82
.10	7	13	18	22	28	40
.15	5	9	12	14	18	26
.25	3	5	7	8	11	15
.50	1	2	3	4	5	7
.90	1	1	1	1	2	2

Note: These are the minimum sample sizes that result after rounding cumulative likelihoods to two decimal places.

3. EMPIRICAL EVIDENCE SUPPORTING THE BINOMIAL MODEL

The problem discovery curves from two recent studies (Lewis, Henry, and Mack, 1990; Virzi, 1990) lend empirical support to this model of problem discovery based on the binomial probability formula (see Figure 2). Virzi (1990) conducted a study in which 20 participants used a computer-based appointment calendar to complete 21 different tasks. Virzi and his associates identified 40 separate usability problems with an average problem occurrence likelihood of 0.36. Lewis et al. (1990) observed fifteen participants complete a set of eleven tasks with an office application system. They identified 145 usability problems with an average problem occurrence likelihood of 0.16. Placing the value of the average problem occurrence likelihood in the cumulative binomial probability formula produced predictive cumulative problem discovery curves that closely matched Monte Carlo problem discovery simulations based on the data provided by Virzi (1990) (Komolgorov-Smirnov $\underline{J}'3 = 0.79$, $\underline{p}=0.56$) and Lewis et al. (1990) (Komolgorov-Smirnov J'3=0.73, p=0.66). This provides empirical evidence that a model of problem discovery based on the binomial probability formula is reasonably accurate. With knowledge about the range of average problem occurrence likelihoods in their past usability studies, usability researchers can use this formula (for example, as embodied in Table 1) to help them select an efficient sample size for future studies.

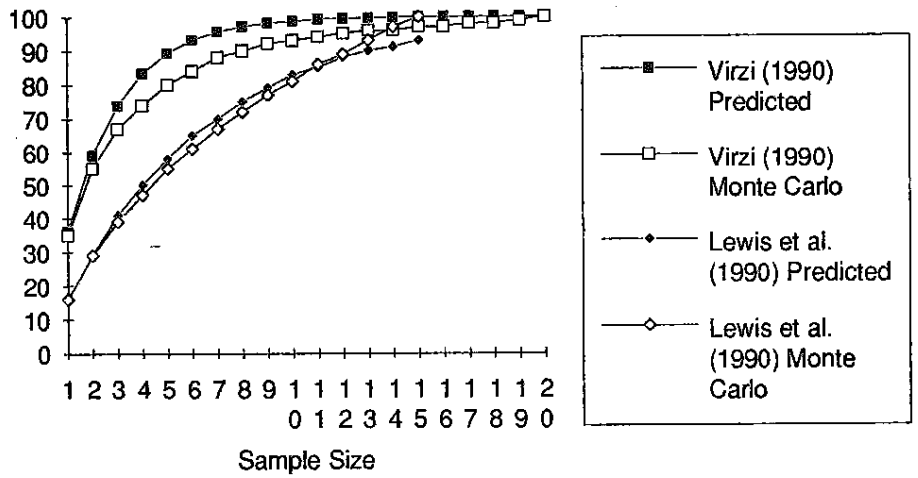

Figure 2. Comparison of predicted (by the cumulative binomial probability formula) and Monte Carlo simulated problem discovery curves for two empirical studies.

Despite these results, researchers should be cautious because there is no way to know how many low-frequency but severe problems exist outside the bounds of a study. This is a critical consideration for usability researchers. In many cases, the diminishing returns expected from running additional participants strongly suggest that the most efficient approach is to run a small sample in problem discovery usability studies, especially if the study will be iterative and undiscovered problems will not lead to dangerous situations (Lewis, 1991). On the other hand, researchers must not become complacent regarding the risk, associated with small sample studies, that they may fail to detect some low-frequency but important problems.

4. REFERENCES

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, NY: North-Holland.

Grice, R. A. and Ridgway, L. S. (1988). A discussion of modes and motives for usability evaluation. IEEE Transactions on Professional Communications, 32, 230-237.

Karat, C. M., Campbell, R., and Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In CHI ' 92 Conference Proceedings (pp. 397-404). Monterey, CA: 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 Thieman, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage Publications.

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: Lawrence Erlbaum.

Lewis, J. R. (1991). Legitimate use of small samples in usability studies: Three examples (Tech. Report 54.594). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R., Henry, S., and Mack, R. L. (1990). Integrated office software benchmarks: A case study. In Human-Computer Interaction -- INTERACT '90 (D. Diaper et al. eds., pp. 337-343). New York, NY: Elsevier.

Norman, D. A. (1983). Design rules based on analyses of human error. Communications of the ACM, 4, 254-258.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. In Proceedings of the Human Factors Society 34th Annual Meeting (pp. 291-294). Orlando, FL: Human Factors Society.

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.