International Encyclopedia of Ergonomics and Human Factors, 2006, Second Edition, Volume 3
Edited by Waldemar Karwowski, Boca Raton, FL: CRC Press

Determining Usability Test Sample Size

Carl W. Turner[*], *James R. Lewis*[†], *and Jakob Nielsen*[‡]
[*]State Farm Insurance Cos., Bloomington, IL 61791, USA
[†]IBM Corp., Boca Raton, FL 33487 USA
[‡]Nielsen Norman Group, Fremont, CA 94539, USA

1 INTRODUCTION

Virzi (1992), Nielsen and Landauer (1993), and Lewis (1994) have published influential articles on the topic of sample size in usability testing. In these articles, the authors presented a mathematical model of problem discovery rates in usability testing. Using the problem discovery rate model, they showed that it was possible to determine the sample size needed to uncover a given proportion of problems in an interface during one test. The authors presented empirical evidence for the models and made several important claims:

- Most usability problems are detected with the first three to five subjects.
- Running additional subjects during the same test is unlikely to reveal new information.
- Return on investment (ROI) in usability testing is maximized when testing with small groups using an iterative test-and-design methodology.

Nielsen and Landauer (1993) extended Virzi's (1992) original findings and reported case studies that supported their claims for needing only small samples for usability tests. They and Lewis (1994) identified important assumptions about the use of the formula for estimating problem discovery rates. The problem discovery rate model was recently re-examined by Lewis (2001).

2 THE ORIGINAL FORMULAE

Virzi (1992) published empirical data supporting the use of the cumulative binomial probability formula to estimate problem discovery rates. He reported three experiments in which he measured the rate at which usability experts and trained student assistants identified problems as a function of the number of naive participants they observed. Problem discovery rates were computed for each participant by dividing the number of problems uncovered during an individual test session by the total number of unique problems found during testing. The average likelihood of problem detection was computed by averaging all participants' individual problem discovery rates.

Virzi (1992) used Monte Carlo simulations to permute participant orders 500 times to obtain the average problem discovery curves for his data. Across three sets of data, the average likelihoods of problem detection (p in the formula above) were 0.32, 0.36, and 0.42. He also had the observers (Experiment 2) and an independent group of usability experts (Experiment 3) provide ratings of problem severity for each problem. Based on the outcomes of these experiments, Virzi made three claims regarding sample size for usability studies: (1) Observing four or five participants allows practitioners to discover 80% of a product's usability problems, (2) observing additional participants reveals fewer and fewer new usability problems, and (3) observers detect the more severe usability problems with the first few participants. Based on these data, he claimed that running tests using small samples in an iterative test-and-design fashion would identify most usability problems and save both time and money.

$$\text{Proportion of unique problems found} = 1 - (1 - p)^n \quad (1)$$

where p is the mean problem discovery rate computed across subjects (or across problems) and n is the number of subjects.

Seeking to quantify the patterns of problem detection observed in several fairly large-sample studies of problem discovery (using either heuristic evaluation or user testing) Nielsen and Landauer (1993) derived the same formula from a Poisson process model (constant probability path independent). They found that it provided a good fit to their problem-discovery data, and provided a basis for predicting the number of problems existing in an interface and performing cost-benefit analyses to determine appropriate sample sizes. Across 11 studies (five user tests and six heuristic evaluations), they found the average value of p to be .33 (ranging from .16 to .60, with associated estimates of p ranging from .12 to .58). Nielsen and Landauer used lambda rather than p, but the two concepts are essentially equivalent. In the literature, ? (lambda), L, and p are commonly used to represent the average likelihood of problem discovery. Throughout this article, we will use p.

$$\text{Number of unique problems found} = N(1 - (1 - p)^n)) \quad (2)$$

where p is the problem discovery rate, N is the total number of problems in the interface, and n is the number of subjects.

The problem discovery rate was approximately .3 when averaged across a large number of independent tests, but the rate for any given usability test will vary depending on several factors (Nielsen & Landauer, 1993). These factors include:

- Properties of the system and interface, including the size of the application.
- Stage in the usability lifecycle the product is tested in, whether early in the design phase or after several iterations of test and re-design.
- Type and quality of the methodology used to conduct the test.
- Specific tasks selected.
- Match between the test and the context of real world usage.
- Representativeness of the test participant.
- Skill of the evaluator.

Research following these lines of investigation led to other, related claims. Nielsen (1994) applied the formula in Equation 2 to a study of problem discovery rate for heuristic evaluations. Eleven usability specialists evaluated a complex prototype system for telephone company employees. The evaluators obtained training on the system and the goals of the evaluation. They then independently documented usability problems in the user interface based on published usability heuristics. The average value of p across 11 evaluators was .29, similar to the rates found during talk-aloud user testing (Nielsen & Landauer, 1993; Virzi, 1992).

Lewis (1994) replicated the techniques applied by Virzi (1992) to data from a usability study of a suite of office software products. The problem discovery rate for this study was .16. The results of this investigation clearly supported Virzi's second claim (additional participants reveal fewer and fewer problems), partially supported the first (observing four or five participants reveals about 80% of a product's usability problems as long as the value of p for a study is in the approximate range of .30 to .40), and failed to support the third (there was no correlation between problem severity and likelihood of discovery). Lewis noted that it is most reasonable to use small-sample problem discovery studies "if the expected p is high, if the study will be iterative, and if undiscovered problems will not have dangerous or expensive outcomes" (1994, p. 377).

3 RECENT CHALLENGES

Recent challenges to the estimation of problem discovery rates appear to take two general forms. The first questions the reliability of problem discovery procedures (user testing, heuristic evaluation, cognitive walkthrough, etc.). If problem discovery is completely unreliable, then how can anyone model it? Furthermore, how can one account for the apparent success of iterative problem-discovery procedures in increasing the usability of the products against which they are applied?

The second questions the validity of modeling the probability of problem discovery with a single value for p. Other issues – such as the fact that claiming high proportions of problem discovery with few participants requires a fairly high value of p, that different task sets lead to different opportunities to discover problems, and the importance of iteration – are addressed at length in earlier papers (Lewis, 1994; Nielsen, 1993).

3.1 Is Usability Problem Discovery Reliable?

Molich *et al*. (1998) conducted a study in which four different usability labs evaluated a calendar system and prepared reports of the usability problems they discovered. An independent team of usability professionals compared the reports produced by the four labs. The number of unique problems identified by each lab ranged from four to 98. Only one usability problem was reported by all four labs. The teams that conducted the studies noted difficulties in conducting the evaluations that included a lack of testing goals, no access to the product development team, a lack of user profile information, and no design goals for the product.

Kessner *et al*. (2001) have also reported data that question the reliability of usability testing. They had six professional usability teams test an early prototype of a dialog box. The total number of usability problems was determined to be 36. None of the problems were identified by every team, and only two were reported by five teams. Twenty of the problems were reported by at least two teams. After comparing their results with those of Molich *et al*. (1999), Kessner *et al*. suggested that more specific and focused requests by a client should lead to more overlap in problem discovery.

Hertzum and Jacobsen (2001) have termed the lack of inter-rater reliability among test observers an 'evaluator effect' – that "multiple evaluators evaluating the same interface with the same usability evaluation method detect markedly different sets of problems" (p. 421). Across a review of 11 studies, they found the average agreement between any two evaluators of the same system ranged from 5% to 65%, with no usability evaluation method (cognitive walkthroughs, heuristic evaluations, or think-aloud user studies) consistently more effective than another. Their review, and the studies of Molich *et al*. (1999) and Kessner *et al*. (2001) point out the importance of setting clear test objectives, running repeatable test procedures, and adopting clear definitions of usability problems. Given that multiple evaluators increase the likelihood of problem detection (Nielsen, 1994), they suggested that one way to reduce the evaluator effect is to involve multiple evaluators in usability tests.

The results of these studies are in stark contrast to earlier studies in which usability problem discovery was reported to be reliable (Lewis, 1996; Marshall, Brendon, & Prail, 1990). The widespread use of usability problem discovery methods indicates that practitioners believe they are reliable. Despite this widespread belief, an important area of future research will be to reconcile the studies that have challenged the reliability of problem discovery with the apparent reality of usability improvement achieved through iterative application of usability problem discovery methods. For example, there might be value in exploring the application of signal detection theory (Swets, Dawes, & Monahan, 2000) to the detection of usability problems.

3.2 Issues in the Estimation of *p*

Woolrych and Cockton (2001) challenged the assumption that a simple estimate of *p* is sufficient for the purpose of estimating the sample size required for the discovery of a specified percentage of usability problems in an interface. Specifically, they criticized the formula for failing to take into account individual differences in problem discoverability and also claimed that the typical values used for *p* (around .30) are overly optimistic. They also pointed out that the circularity in estimating the key parameter of *p* from the study for which you want to estimate the sample size reduces its utility as a planning tool. Following close examination of data from a previous study of heuristic evaluation, they found combinations of five participants which, if they had been the only five participants studied, would have dramatically changed the resulting problems lists, both for frequency and severity. They recommended the development of a formula that replaces a single value for *p* with a probability density function.

Caulton (2001) claimed that the simple estimate of *p* only applies given a strict homogeneity assumption – that all types of users have the same probability of encountering all usability problems. To address this, Caulton added to the standard cumulative binomial probability formula a parameter for the number of heterogeneous groups. He also introduced and modeled the concept of problems that heterogeneous groups share and those that are unique to a particular subgroup. His primary claims were (1) the more subgroups, the lower will be the expected value of *p* and (2) the more distinct the subgroups are, the lower will be the expected value of *p*.

Most of the arguments of Woolrych and Cockton (2001) were either addressed in previous literature or do not stand up against the empirical findings reported in previous literature. It is true that estimates of *p* can vary widely from study to study. This characteristic of usability testing can be addressed by estimating *p* for a study after running two subjects and adjusting the estimate as the study proceeds (Lewis, 2001). There are problems with the estimation of *p* from the study to which you want to apply it, but recent research (discussed below) provides a way to overcome these problems. Of course, it is possible to select different subsets of participants who experienced problems in a way that leads to an overestimate of *p* (or an underestimate of *p*, or any value of *p* that the person selecting the data wishes). Test administrators should follow accepted practice and select evaluators who represent the range of knowledge and skills found in the population of end users. There is no compelling evidence that a probability density function would lead to an advantage over a single value for *p*, although there might be value in computing confidence intervals for single values of *p*.

Caulton's (2001) refinement of the model is consistent with the observation that different user groups expose different types of usability problems (Nielsen, 1993). It is good practice to include participants from significant user groups in each test; three or four per group for two groups and three participants for more than two groups. If there is a concern that different user groups will uncover different sets of usability problems then the data for each group can be analyzed separately, and a separate *p* computed for each user group. However, Caulton's claim that problem discovery estimates are always inflated when averaged across heterogeneous groups and problems with different values of *p* is inconsistent with the empirical data presented in Lewis (1994). Lewis demonstrated that *p* is robust, showing that the mean value of *p* worked very well for modeling problem discovery in a set of problems that had widely varying values of *p*.

4 IMPROVING SMALL-SAMPLE ESTIMATION OF *p*

Lewis (2001), responding to an observation by Hertzum and Jacobsen (2001) that small-sample estimates of *p* are almost always inflated, investigated a variety of methods for adjusting these small-sample estimates to enable accurate assessment of sample size requirements and true proportions of discovered problems. Using data from a series of Monte Carlo studies applied against four published sets of problem discovery databases, he found that a technique based on combining information from a normalization procedure and a discounting method borrowed from statistical language modeling produced very accurate adjustments for small-sample estimates of *p*. The Good-Turing (GT) discounting procedure reduced, but did not completely eliminate, the overestimate of problem discovery rates produced by small-sample *p* estimates. The GT adjustment, shown in Equation 3, was:

$$p_{GT\text{-}adj} = \frac{p_{est}}{\left(1 + \frac{E(N_1)}{N}\right)} \qquad (3)$$

where p_{est} is the initial estimate computed from the raw data of a usability study, $E(N_1)$ was the number of usability

problems detected by only one user, and N was that total number of unique usability problems detected by all users.

By contrast, the normalization procedure (Norm) slightly underestimated problem discovery rates. The equation was:

$$p_{\text{Norm-adj}} = \left(p_{\text{est}} - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right) \quad (4)$$

where p_{est} is the initial estimate computed from the raw data of a usability study and n was the number of test participants. He concluded that the overestimation of p from small-sample usability studies is a real problem with potentially troubling consequences for usability practitioners, but that it is possible to apply these procedures (normalization and Good-Turing discounting) to compensate for the overestimation bias. Applying each procedure to the initial estimate of p, then averaging the results, produces a highly accurate estimate of the problem discovery rate. Equation 5 shows the formula for an adjusted p estimate based on averaging Good-Turing and normalization adjustments.

$$p_{\text{adj}} = \frac{1}{2}\left(p_{\text{GT-adj}} + p_{\text{Norm-adj}}\right) \quad (5)$$

"Practitioners can obtain accurate sample size estimates for problem-discovery goals ranging from 70% to 95% by making an initial estimate of the required sample size after running two participants, then adjusting the estimate after obtaining data from another two (total of four) participants" (Lewis, 2001, p.474).

The results of a return-on-investment (ROI) model for usability studies (Lewis, 1994) indicated that the magnitude of p affected the point at which the percentage of problems discovered maximized ROI. For values of p ranging from .10 to .5, the appropriate problem discovery goal ranged from .86 to .98, with lower values of p associated with lower problem discovery goals.

5 AN APPLICATION OF THE ADJUSTMENT PROCEDURES

In the example shown in Table 1, a usability test with eight participants has led to the discovery of four unique usability problems. The problem discovery rates (p) for individual participants ranged from 0.0 to .75. The problem discovery rates for specific problems ranged from .125 to .875. The average problem discovery rate (averaged either across problems or participants), p_{est}, was .375. Note that Problems 2 and 4 were detected by only one participant (Participants 2 and 7, respectively). Applying the Good-Turing estimating procedure from Equation 3 gives

$$p_{\text{GT-adj}} = \frac{0.375}{\left(1 + \frac{2}{4}\right)} = 0.25$$

TABLE 1
Data from a Hypothetical Usability Test with Eight Subjects, p_{est} = .375

	Problem Number					
Subject	**1**	**2**	**3**	**4**	**Count**	***p***
1	1	0	1	0	2	0.500
2	1	0	1	1	3	0.750
3	1	0	0	0	1	0.250
4	0	0	0	0	0	0.000
5	1	0	1	0	2	0.500
6	1	0	0	0	1	0.250
7	1	1	0	0	2	0.500
8	1	0	0	0	1	0.250
Count	7	1	3	1		
P	0.875	0.125	0.375	0.125		0.375

Applying normalization as shown in Equation 4 gives

$$p_{\text{Norm-adj}} = \left(0.375 - \frac{1}{8}\right)\left(1 - \frac{1}{8}\right) = 0.22$$

The adjusted problem discovery rate is obtained by averaging the two estimates as shown in Equation 5 gives

$$p_{\text{adj}} = \frac{1}{2}\left(0.25 + 0.22\right) = 0.235$$

With this adjusted value of p and the known sample size, it is possible to estimate the sample size adequacy of this study using the cumulative binomial probability formula: $1 - (1 - .25)^8 = .90$. If the problem discovery goal for this study had been 90%, then the sample size was adequate. If the discovery goal had been lower, the sample size would be excessive, and if the discovery goal had been higher, the sample size would be inadequate. The discovery of only four problems (one problem for every two participants) suggests that the discovery of additional problems would be difficult. If four problems constitute 90% of the problems available for discovery given the specifics of this usability study, then 100% of the problems available for discovery should be about 4/.9, or 4.44. In non-numerical terms, there probably aren't a lot of additional problems to extract from this problem discovery space.

As an example of sample size estimation, suppose you had data from the first four participants and wanted to estimate the number of participants you'd need to run to achieve 90% problem discovery. After running the fourth participant, there were three discovered problems (because Problem 2 did not occur until Participant 7), as shown in Table 2. One of those problems (Problem 4) occurred only once.

TABLE 2
Data from a Hypothetical Usability Test; First Four Subjects, p_{est} = .500

	Problem Number				
Subject	**1**	**3**	**4**	**Count**	**p**
1	1	1	0	2	0.667
2	1	1	1	3	1.000
3	1	0	0	1	0.333
4	0	0	0	0	0.000
Count	3	2	1		
P	0.750	0.500	0.250		0.500

Applying the Good-Turing estimating procedure from Equation 3 gives

$$p_{GT\text{-}adj} = \frac{0.500}{\left(1+\frac{1}{3}\right)} = 0.375$$

Applying normalization as shown in Equation 4 gives

$$p_{Norm\text{-}adj} = \left(0.500-\frac{1}{4}\right)\left(1-\frac{1}{4}\right) = 0.188$$

The average of the two estimates is

$$p_{adj} = \frac{1}{2}\left(0.375+0.188\right) = 0.28$$

Given p = .28, the estimated proportion of discovered problems would be $1 - (1 - .28)^4$, or .73. Doing the same computation with n = 7 gives .90, indicating that the appropriate sample size for the study would be 7. Note that in the matrix for this hypothetical study, running the eighth participant did not reveal any new problems.

6 CONCLUSIONS

The cumulative binomial probability formula (given appropriate adjustment of p when estimated from small samples) provides a quick and robust means of estimating problem discovery rates (p). This estimate can be used to estimate usability test sample size requirements (for studies that are underway) and to evaluate usability test sample size adequacy (for studies that have already been conducted).

Further research is needed to answer remaining questions about when usability testing is reliable, valid, and useful.

REFERENCES

CAULTON, D.A., 2001, Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, 20, 1-7.

HERTZUM, M. and JACOBSEN, N.E., 2001, The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13, 421-443.

KESSNER, M., WOOD, J., DILLON, R.F. and WEST, R.L., 2001, On the reliability of usability testing. In Jacko, J. and Sears, A., (eds), *Conference on Human Factors in Computing Systems: CHI 2001 Extended Abstracts* (Seattle, WA: ACM Press), pp. 97-98.

LEWIS, J.R., 1994, Sample sizes for usability studies: Additional considerations. *Human Factors*, 36, 368-378.

LEWIS, J.R., 1996, Reaping the benefits of modern usability evaluation: The Simon story. In Salvendy, G. and Ozok, A., (eds), *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics – ICAE '96* (Istanbul, Turkey: USA Publishing), pp. 752-757.

LEWIS, J.R., 2001, Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction*, 13, 445-479.

MARSHALL, C., BRENDAN, M. and PRAIL, A., 1990, Usability of product X – lessons from a real product. *Behaviour & Information Technology*, 9, 243-253.

MOLICH, R., BEVAN, N., CURSON, I., BUTLER, S., KINDLUND, E., MILLER, D. and KIRAKOWSKI, J., 1998, Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association Conference* (Washington, DC: UPA), pp. 83-84.

NIELSEN, J., 1993, *Usability engineering* (San Diego, CA: Academic Press).

NIELSEN, J., 1994, Heuristic evaluation. In Nielsen, J. and Mack, R.L. (eds), *Usability Inspection Methods* (New York: John Wiley), pp. 25-61.

NIELSEN, J. and LANDAUER, T.K., 1993, A mathematical model of the finding of usability problems. In *Proceedings of ACM INTERCHI'93 Conference* (Amsterdam, Netherlands: ACM Press), pp. 206-213.

SWETS, J.A., DAWES, R.M. and MONAHAN, J., 2000, Better decisions through science. *Scientific American*, 283(4), 82-87.

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

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