Sample Size Adequacy for Usability Studies: Current Issues

TR 29.3789
May 10, 2004

James R. Lewis

IBM Pervasive Computing Division

Boca Raton, Florida

Abstract

Application of the binomial probability formula to the problem of estimating sample sizes for usability studies began in the early 1980s. Through the 1990s there were several empirical studies that supported the use of this formula and the use of small-sample usability studies. Around 2000, there was some backlash against the use of small samples, and questions arose regarding the applicability of the simple binomial probability formula. The purpose of this report is to summarize the original research supporting the use of small samples and the binomial probability formula, and to examine the research arguing against their unmodified use.

ITIRC Keywords

Usability problems
Problem discovery rate
Binomial probability formula
Sample size estimation
Sample size adequacy

Contents

Introduction 1
The Original Research 3
Critical Responses to the Original Claims 5
Discussion of Critical Responses 7
  Reliability Issues 7
  Model Issues: Should Practitioners Abandon $1-(1-p)^n$ 7
  Recurring Issues 8
Conclusions 11
References 13

Introduction

Application of the binomial probability formula to the problem of estimating sample sizes for usability studies began in the early 1980s. Through the 1990s there were several empirical studies that supported the use of this formula and the use of small-sample usability studies. Around 2000, there was some backlash against the use of small samples, and questions arose regarding the applicability of the simple binomial probability formula. The purpose of this report is to summarize the original research supporting the use of small samples and the binomial probability formula, and to examine the research arguing against their unmodified use.

The Original Research

The earliest use of the formula $1-(1-p)^n$ for the purpose of justifying a sample size in usability studies was in Lewis (1982). In this paper, the formula was derived from the cumulative binomial probability formula. The claims from this paper were that "the recommended minimum number of subjects depends on the number of times a problem must be observed before it is regarded as a problem and the magnitude of the proportion of the user population for which one wishes to detect problems" (p. 719). Wright and Monk (1991) also offered the formula as a means for estimating minimum sample sizes for usability studies, concluding "even a problem that has only a probability of 0.3 of being detected in one attempt has a very good chance of being detected in four attempts" (p. 903). Neither Lewis (1982) nor Wright and Monk (1991) provided any empirical data to assess how well the formula modeled problem discovery in practice.

Virzi (1990, 1992) was one of the first researchers to provide empirical data supporting the use of the formula. He reported three experiments in which he measured the rate at which trained usability experts identified problems as a function of the number of naive participants they observed. He 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 (1992) 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.

Nielsen and Molich (1990) also used a Monte Carlo procedure to investigate patterns of problem discovery in heuristic evaluation as a function of the number of evaluators. The major claims from this paper were that individual evaluators typically discovered from about 20 to 50% of problems available for discovery, but combined information from individual evaluators into aggregates did much better, even when the aggregates consisted of only three to five evaluators. Nielsen (1992) replicated the findings of Nielsen and Molich, and had results that supported the additional claims that evaluators with expertise in either the product domain or usability had higher problem discovery rates than novice evaluators. They also found data to support the claim that evaluators who were experts both in usability and the product domain had the highest problem discovery rates.

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 lambda ranging from .12 to .58). (Note that 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.) Nielsen (2000) used these results along with other arguments to support the recommendation of five participants as the optimal sample size for usability studies.

Lewis (1992, 1994) replicated the techniques applied by Virzi to data from an independent usability study (Lewis, Henry, & Mack, 1990). 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). He also pointed out that estimating the sample size requirement for the number of participants is only one element among several that usability practitioners must consider. Another key element is the selection and construction of the tasks and scenarios that participants will encounter in a study – concerns that are similar to the problem of assuring content validity in psychometrics.

Research following these lines of investigation led to other, related claims. In two thinking-aloud experiments, Nielsen (1994) found that the value of *p* for experimenters who were not usability experts was about .30, and that after running five test participants the experimenters had discovered 77-85% of the usability problems (replicating the results of Nielsen & Molich, 1990). Dumas, Sorce, and Virzi (1995) investigated the effect of time per evaluator and number of evaluators, and concluded that both additional time and additional evaluators increased problem discovery. They suggested that it was more effective to increase the number of evaluators than to increase time per evaluator.

Critical Responses to the Original Claims

Molich et al. (1998, 1999) conducted two studies in which they had several different usability labs evaluate the same system and prepare reports of the usability problems they discovered. There was significant variance among the labs in the number of problems reported, and there was very little overlap among the labs with regard to the specific problems reported.

Kessner, Wood, Dillon, and West (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 (1999, 2001) have described an 'evaluator effect' – that "multiple evaluators evaluating the same interface with the same usability evaluation method detect markedly different sets of problems" (Hertzum & Jacobsen, 2001, p. 421). Across a review of 11 studies, they found the average agreement between any two evaluators of the same system to ranged from 5 to 65%, with no usability evaluation method (cognitive walkthroughs, heuristic evaluations, or thinking-aloud studies) consistently better than another. They have suggested that one way to reduce the evaluator effect is to involve multiple evaluators in usability evaluations.

Spool and Schroeder (2001) observed the same task performed by 49 users on four production web sites and tracked the rates of problem discovery of new usability problems on each site. They found that five users would not provide sufficient information to find 85% of the total number of usability problems. They attributed this to relatively low values of p for the four sites they evaluated (all p less than 0.16). For this type of situation, they recommended a strategy of ongoing usability testing with a user or two every week rather than a test with six to eight users every six months. Perfetti and Landesman (2001) described similar results and conclusions in usability evaluations conducted on the task of purchasing an audio CD from a web site.

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. Finally, they pointed out that the problems available for discovery depend on the task sets used to evaluate an interface, and that there are issues in the extraction of true usability problems from observational data.

Caulton (2001) also challenged the assumption that a simple estimate of *p* is sufficient for the accurate prediction of problem discovery, especially when the user population is composed of distinctly heterogeneous groups. Caulton further 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*.

Lewis (2001), responding to the 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*. 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. "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).

Discussion of Critical Responses

The critical responses to the original claims 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 isn't reliable, then how can anyone model it? The second questions the validity of modeling the probability of problem discovery with a single value for p (or lambda, or L, depending on the different conventions used in different papers). Other criticisms, 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 present in the earlier work (for example, see Lewis, 1994 and Nielsen, 2000).

Reliability Issues

Results from Hertzum and Jacobsen (1999, 2001), Molich et al. (1998, 1999), and Kessner et al. (2001) are disturbing. Are usability practitioners engaging in a massive self-deception similar to that of clinicians using projective tests (Lilienfeld, Wood, & Garb, 2000)? Usability practitioners need to come to grips with the results from these studies, making this an important topic for continued research.

In practice, practitioners often only have a single opportunity to evaluate an interface, so they can't tell if their interventions have really improved an interface. In my experience, however, when I have conducted standard scenario-based problem-discovery studies with multiple participants and one observer, and have done so iteratively, the measurements across iterations consistently indicate substantial and statistically reliable improvements in usability (for example, see Lewis, 1996).

Despite the apparent reality of lack of correspondence of problem reports among groups of investigators and a probable evaluator effect, experience suggests that these usability evaluation methods can work. An important task for future research is to reconcile the apparent lack of reliability with the apparent reality of usability improvement achieved through the iterative application of usability evaluation methods.

In usability studies, observer and participant performance is not perfectly reliable (in other words, there is an 'evaluator' effect) – otherwise there would be no reason to cumulate data across observers and participants. Furthermore, no single usability method can detect all possible usability problems – models of problem discovery apply only to the specific usability evaluation setting (method, tasks, types of participants, etc.) Finally, another potential cause of the reported low reliability of problem discovery across practitioner groups is the vague definition of what constitutes a usability problem.

Model Issues: Should Practitioners Abandon $1-(1-p)^n$

Some recent papers (Caulton, 2001; Woolrych & Cockton, 2001) have claimed that using a single value for *p* in the standard formula based on the binomial probability formula is too restrictive. Woolrych and Cockton (2001) have recommended replacing the single value of *p* with a probability density function. Caulton (2001) claimed that it was necessary to replace the single value of *p* with *p*s for heterogeneous groups of participants and problems.

In the light of new investigations addressing the problem of small-sample inflation of estimates of p (Lewis, 2001), such adjustments do not seem to be necessary. The data from Lewis (2001) indicate good prediction of usability problem discovery given adjustment of the initial inflated estimates of *p*, and prediction using adjusted values of *p* are likely to address many of the problems that these authors (Caulton, 2001; Woolrych & Cockton, 2001) have observed.

The binomial probability formula does not have any underlying assumption of homogeneity. Data from Lewis (1994) indicates good modeling for individual problems or groups of problems (using mean *p* across problems in a group to model that group). At the lowest possible level of definition, each participant is the sole member of a 'group'. At the highest level of definition, all participants are members of one large group (the group of humans). It can be informative to classify participants into groups and to model problem discovery by group, but this does not mean that sample size projections based on the standard formula are without value.

Recurring Issues

It is important not to overstate the power of small-sample usability problem-discovery studies. The value of *p* directly affects the sample size required to achieve any specific problem-discovery percentage goal. The five-participant rule of thumb (observing five participants leads to the discovery of 80% or more of problems) operates only when the value of *p* is relatively high (specifically, higher than .275). While this appears to generally be the case for many domains in which practitioners evaluate usability (Lewis, 1994; Nielsen & Landauer, 1993), it is certainly not always true (Lewis, 1994; Nielsen & Landauer, 1993; Perfetti & Landesman, 2001; Spool & Schroeder, 2001). Whether web site evaluation results in generally lower values of *p* than the evaluation of other types of software (as suggested by Perfetti & Landesman, 2001 and Spool & Schroeder, 2001) remains an open question.

It is also important to consider iteration an important component of an overall usability testing scheme (Lewis, 1994; Nielsen, 2000). The primary value of running a small sample of participants in a usability study is to use the results (the list of discovered problems) to improve the product before running another study. There is little value in watching participant after participant encounter the same problems. Other iteration schemes might focus on changing the types of participants observed or the tasks that participants undertake.

Despite the evidence to the contrary reported in Lewis (1994), many practitioners continue to believe that they will discover the most important problems early in an evaluation. Virzi (1993) reported this phenomenon, but this effect was not present in the data reported by Lewis (1994).

Resolution of this discrepancy in research results has not yet occurred. There is no component in the binomial probability formula for problem importance, so the conservative (in my opinion, prudent) approach is to assume that problems of differing importance will be discovered according to their probability of discovery (which is probably a combination of their likelihood of occurrence and salience to observers). In other words, there is no guarantee of the discovery of more important problems before less important problems.

Conclusions

Little has changed since 1994 when I pointed out that the most reasonable use of small-sample problem discovery studies is when the expected value of *p* is high, the study will be iterative, and undiscovered problems will not have dangerous or expensive outcomes.

It is undoubtedly true that small-sample estimates of *p* are inflated (Hertzum & Jacobsen, 2001), but it is possible to adjust *p* to reduce the bias (Lewis, 2001). Practitioners should routinely use the adjustment methods of Lewis (2001) to estimate the value of *p* early in their usability studies (using data from the first two participants, then revising the estimate with data from the first four participants) and to use that value as a guide to the number of participants to observe. At the end of a usability study, practitioners can use the adjusted value of *p* to assess the adequacy of the study's sample size (see Lewis, 2002 for procedural details).

For most practical purposes, there does not appear to be any need to complicate the use of $1-(1-p)^n$. A unitary value of *p* will suffice.

Measurements from iterative evaluation and modification of user interfaces indicate that the procedures of usability testing can be reliable, but there is a need for more research to investigate lack of reliability across independent evaluations. For example, does fixing a set of reported problems tend to enhance usability even when the sets of reported problems are different? It will be interesting to see how future research addresses this complex issue.

References

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

Dumas, J., Sorce, J., & Virzi, R. (1995). Expert reviews: how many experts is enough? In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 309-312). Santa Monica, CA: HFES.

Hertzum, M., & Jacobsen, N. E. (1999). The evaluator effect during first time use of the cognitive walkthrough technique. In *Proceedings of HCI International '99 8th International Conference on Human-Computer Interaction* (pp. 1063-1067). Munich, Germany: Lawrence Erlbaum.

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

Lewis, J. R. (1982). Testing small system customer set-up. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 718-720). Dayton, OH: Human Factors Society.

Lewis, J. R. (1992). *Sample sizes for usability studies: additional considerations* (Tech. Report 54.711). Boca Raton, FL: International Business Machines Corp.

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 G. Salvendy and A. Ozok (eds.), *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics -- ICAE '96* (pp. 752-757). Istanbul, Turkey: USA Publishing.

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.

Lewis, J. R. (2002). *Effect of level of problem description on problem discovery rate: Two case studies* (Tech. Report 29.3604). Raleigh, NC: International Business Machines Corp.

Lewis, J. R., Henry, S. C., and Mack, R. L. (1990). Integrated office scenario benchmarks: A case study. In *INTERACT '90* (pp. 337-343). London: North-Holland.

Lilienfeld, S. C., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27-66.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association* (pp. 189-200). Washington, DC: UPA.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., & Arcuri, M. (1999). Comparative evaluation of usability tests. In *Conference on Human Factors in Computing Systems: CHI 1999 Extended Abstracts* (pp. 83-84). Pittsburgh, PA: ACM Press.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the ACM CHI '92 Conference* (pp. 373-380). New York: ACM.

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, *41*, 385-397.

Nielsen, J. (2000, March). Why you only need to test with 5 users. *Alertbox*, March 19, 2000. http://www.useit.com/alertbox/20000319.html

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

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of CHI 90* (pp. 249-256). New York, NY: ACM.

Perfetti, C., & Landesman, L. (2001). Eight Is Not Enough -- *UIEtips* 06/05/01. Electronic newsletter. http://www.uie.com/Articles/eight_is_not_enough.htm

Spool, J. M., & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. In J. Jacko and A. Sears (Eds.), *Conference on Human Factors in Computing Systems: CHI 2001 Extended Abstracts* (pp. 285-286). Seattle, WA: ACM Press.

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: HFES.

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

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

Wright, P. C., & Monk, A. F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, 35*, 891-912.