# Current Issues in the Determination of Usability Test Sample Size: How Many Users is Enough?

Carl W. Turner (State Farm Insurance, carl.turner.hxyf@statefarm.com)
Jakob Nielsen (Nielsen Norman Group, nielsen@nngroup.com)
James R. Lewis (IBM Corp., jimlewis@us.ibm.com)

## Abstract
The topic of "how many users" is of great interest to usability specialists who need to balance project concerns over ROI and timelines with their own concerns about designing usable interfaces. The panel will review the current controversies over usability test sample size, test validity, and reliability.

## Introduction
Virzi (1992), Nielsen and Landauer (1993), and Lewis (1994) published influential articles on the topic of sample size in usability testing. In these articles, the authors presented a mathematical model for determining the sample size for usability tests. The authors presented empirical evidence for the models and made several important claims:

- Most usability problems are detected with three to five subjects
- Running additional subjects during the same test is unlikely to reveal new information
- Most severe usability problems are detected by the first few subjects (claim supported by Virzi's data – not supported by Lewis' data)

Virzi's stated goal of determining an appropriate number of test subjects was to improve return on investment (ROI) in product development by reducing the time and cost involved in product design. Nielsen and Landauer (1993), building on earlier work by Nielsen (1988; 1989) and Nielsen et al. (1990), replicated and extended Virzi's (1992) original findings and reported case studies that supported their claims for needing only small samples for usability tests. The "small sample" claims and their impact on usability methodology have been popularized in Nielsen's (2000) widely read "useit.com" online column.

Since that time, a number of authors have challenged Virzi's and Nielsen's "small sample" findings on methodological and empirical grounds (Bailey, 2001; Caulton, 2001; Spool & Schroeder, 2001; Woolrych & Cockton, 2001). Additionally, two large-scale experiments on usability test methods have been conducted that bear directly on Virzi and Nielsen's claims (Molich et al, 1998; Molich et al, 1999).

The topic of "how many users" is of great interest to usability specialists who need to balance project concerns over ROI and timelines with their own concerns about designing usable interfaces. The goals of this panel discussion are to review the current controversies over usability test sample size and lead a discussion of the topic with the audience:

- Examine the original "small sample" claims, including Nielsen's (1990), Nielsen and Landauer's (1993), Virzi's (1992), and Lewis' (1992, 1994) articles
- Review the responses, including studies that deal with reliability of usability testing
- Make suggestions and recommendations for follow up studies
- Each panelist will present their perspective on the topic of usability sample size (15 minutes). The panelists will then lead discussion of the topic with attendees.

## The Original Claims
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.)

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 (1994a) 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.

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 (2001a), 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, 2001a, p.474).

## Discussion

The new 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 papers (for example, see Lewis, 1994).

Some of the issues that will likely arise in the presentation of position papers and during audience participation are:

- What is the role of usability testing in the user-centered design process?
- Are usability results reliable and valid?
- Are problem frequency and impact (severity) correlated or independent?
- To what extent is it import to continue to focus on ROI for cost justification of usability testing?
- What is the role of and importance of the experience of usability tester in detecting usability problems?
- What are the consequences of defining and testing with different user groups?
- Do we really need more complex formulas for modeling problem discovery, or will adjustment of inflated estimates of *p* solve the modeling problems?
- What are attendees' lessons learned based on the impact of selecting a sample for usability testing?
- Case studies: who has them? We need more data!

## References (Both Cited and Background)


Bailey, R. (2001). How reliable is usability performance testing? *UI design update newsletter, Sept. 2001*. http://www.humanfactors.com/library/sep01.asp#bobbailey

Bias, R. (1992). Top 10 ways to muck up an interface project. *IEEE Software*, Nov., 95-96.

Bias, R., & Mayhew, D. (Eds.) (1994). *Cost-justifying usability*. New York, NY: Academic Press.

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.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Conference on Human Factors in Computing Systems: CHI 2000 Conference Proceedings* (pp. 345-352). The Hague, the Netherlands: ACM Press.

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.

ISO 9241-11 (1998). *Ergonomic requirements for office work with visual display terminals (VDT's) – Part 11: Guidance on usability*. Geneva, Switzerland: Author.

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. (1993). Problem discovery in usability studies: A model based on the binomial probability formula. In *Proceedings of the Fifth International Conference on Human-Computer Interaction* (pp. 666-671). Orlando, FL: Elsevier.

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

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

Lewis, J. R. (2001b). Introduction: current issues in usability evaluation. *International Journal of Human-Computer Interaction*, *13*, 343-349.

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, May). 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. (1988): Evaluating the thinking aloud technique for use by computer scientists. In *Proceedings of the IFIP Working Group 8.1. International Workshop on Human Factors of Information Systems Analysis and Design*. London, UK: IFIP.

Nielsen, J. (1989). Usability engineering at a discount. In Salvendy, G. and Smith, M.J. (Eds.), Designing and Using Human-Computer Interfaces and Knowledge Based Systems (pp. 394-401). Amsterdam, Netherlands: Elsevier Science Publishers.

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

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

Nielsen, J. (1994b). Heuristic evaluation. In J. Nielsen and R. L. Mack (Eds.), *Usability inspection methods* (pp. 25-61). New York, NY: John Wiley.

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., Dray, S. M., Foley, J. D., Walsh, P., and Wright, P. (1990). Usability engineering on a budget. Proceedings of the IFIP INTERACT'90 Conference (pp. 1067-1070). Cambridge, U.K.: IFIP.

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.

Spool, J. M. (2001). Eight Is Not Enough -- *UIEtips* 06/05/01. Electronic newsletter.

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. (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.