DOI:10.1145/2133806.2133824

Magic numbers are strictly hocus-pocus, so usability studies must test many more subjects than is usually assumed.

BY MARTIN SCHMETTOW

Sample Size in Usability Studies

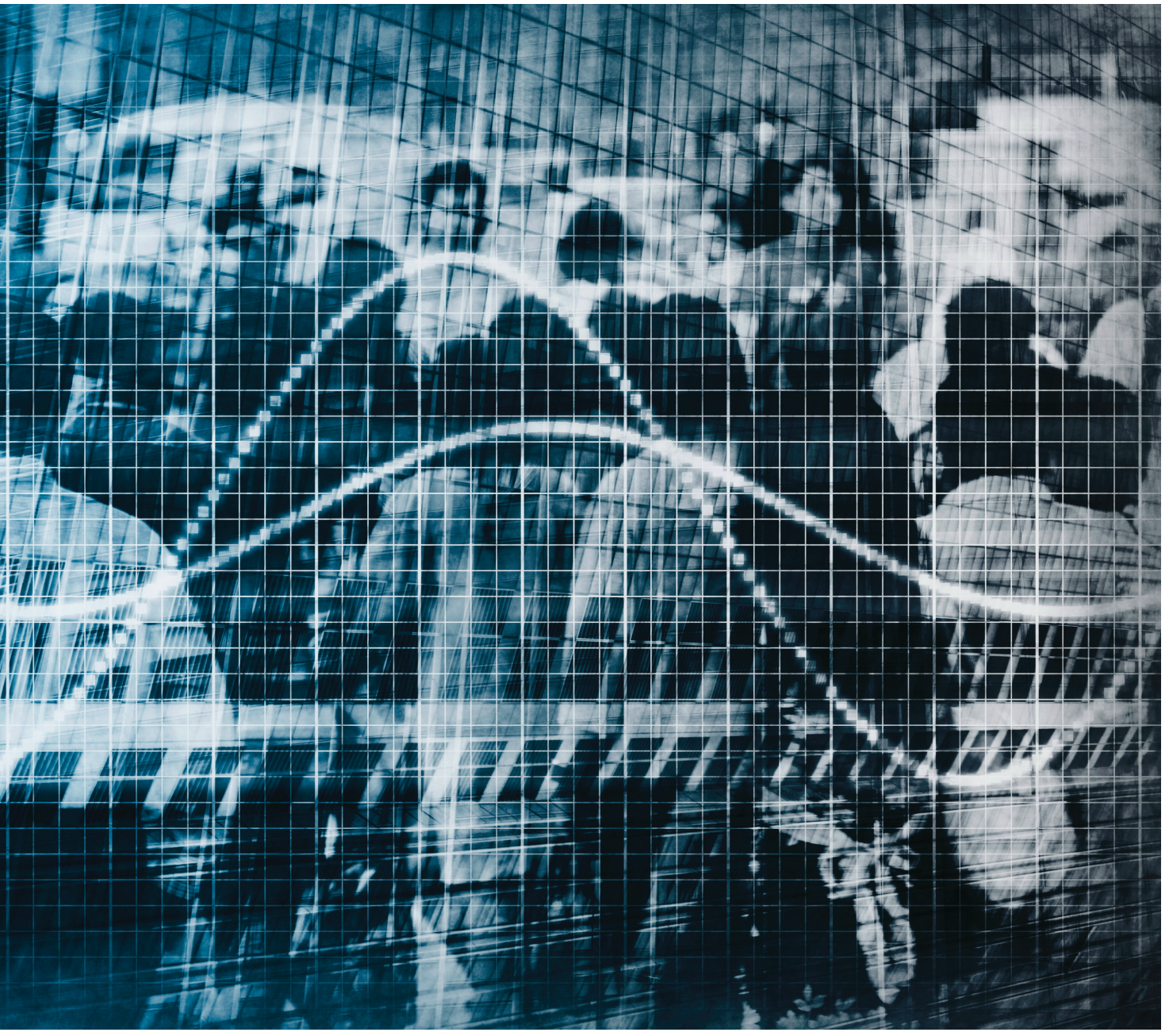
PHOTOGRAPH BY KENTOH / SHUTTERSTOCK.COM

» key insights

- Usability testing is recommended for improving interactive design, but discovery of usability problems depends on the number of users tested.
- For estimating required sample size, usability researchers often resort to either magic numbers or the geometric series formula; inaccurate for making predictions, both underestimate required sample size.
- When usability is critical, an extended statistical model would help estimate the number of undiscovered problems; researchers incrementally add participants to the study until it (almost) discovers all problems.

USABILITY STUDIES ARE a cornerstone activity for developing usable products. Their effectiveness depends on sample size, and determining sample size has been a research issue in usability engineering for the past 30 years.[10] In 2010, Hwang and Salvendy[6] reported a meta study on the effectiveness of usability evaluation, concluding that a sample size of 10 ± 2 is sufficient for discovering 80% of usability problems (not five, as suggested earlier by Nielsen[13] in 2000). Here, I show the Hwang and Salvendy study ignored fundamental mathematical properties of the problem, severely limiting the validity of the 10 ± 2 rule, then look to reframe the issue of effectiveness and sample-size estimation to the practices and requirements commonly encountered in industrial-scale usability studies.

Usability studies are important for developing usable, enjoyable products, identifying design flaws (usability problems) likely to compromise the user experience. Usability problems take many forms,



possibly slowing users doing tasks, increasing the probability of errors, and making it difficult for users to learn a particular product. Usability studies take two general forms: In empirical usability testing, representative users are observed performing typical tasks with the system under consideration, and in usability inspections, experts examine the system, trying to predict where and how a user might experience problems. Many variants of usability testing and expert inspection have been proposed, but how effective are they at actually discovering usability problems? Moreover, how can HCI researchers increase the effectiveness of usability studies? The answer is simple: Increasing the sample size (number of tested participants or number of experts) means more problems will be found. But how many evaluation sessions should researchers conduct? What is a sufficient sample size to discover a certain proportion of problems, if one wants to find, say, at least 80% of all those that are indeed there to be found?

Attempts to estimate the sample size date to 1982[10]; a distinct line of research emerged about 10 years later when Virzi[20] suggested a mathematical model for the progress of usability studies. The proportion of successfully discovered usability problems D was assumed to depend on the average probability p of finding a problem in a single session and number of independent sessions n (the sample or process size). The progress of discovery D was assumed to follow a geometric series $D=1-(1-p)n$.

In 1993, Nielsen and Landauer[14] reported that the average probability p varies widely among studies.

Based on the average *p=0.31* over several studies, Nielsen later concluded that 15 users is typically enough to find virtually all problems,[13] recommending three smaller studies of five participants each (finding 85% of problems in each group) for driving iterative design cycles. Unfortunately, researchers, students, and usability professionals alike misconstrued Nielsen's recommendations and began to believe a simplified version of the rule: Finding 85% of the problems is enough, and five users usually suffice to reach that target.

This conclusion initiated the "five users is (not) enough" debate, involving proponents and skeptics from research and industry.[a] Spool and Schroeder[18] reviewed an industrial dataset, concluding that complex modern applications require a much larger sample size to reach a target of 80% discovery. In 2001, Caulton[3] said the probability of discovering a particular problem likely differs among subgroups within a user population. Likewise, Woolrych and Cockton[22] presumed that heterogeneity in the sample of either participants or experts could render Virzi's formula biased.

The debate has continued to ponder the mathematical foundation of the geometric series model. In fact, the formula is grounded in another well-known model—binomial distribution—addressing the question of how often an individual problem is discovered through a fixed number of trials (sample size or process size *n*). The binomial model is based on three fundamental assumptions that likewise are relevant for the geometric series model:

Independence. Discovery trials are stochastically independent;

Completeness. Observations are complete, such that the total number of problems is known, including those not yet discovered; and

Homogeneity. The parameter *p* does not vary, such that all problems are equally likely to be discovered within a study; I call the opposite of this assumption "visibility variance."

Observing that the average probability *p* varies across studies[14] is a strong argument against generalized assertions like "*X* test participants suffice to find *y*% of problems." A mathematical solution for dealing with uncertainty regarding *p* devised by Lewis[9] suggested that estimating the mean probability of discovery *p* from the first few sessions of a study is helpful in predicting required sample size. Lewis also realized it is not enough to take only the average rate of successful discovery events as an estimator for *p*. The true total number of existing problems is typically unknown a priori, thus violating the completeness assumption. In incomplete studies, not-yet-discovered problems decrease estimated probability. Ignoring incompleteness results in an optimistic bias for the mean probability *p*. For a small sample size, Lewis suggested a correction term for the number of undiscovered problems, or the Good-Turing (GT) adjustment.

However, when evaluating the prediction from small-size subsamples via Monte-Carlo sampling, Lewis treated the original studies as if they were complete. Hence, he did not adjust the baseline of total problem counts for potentially undiscovered problems, which is critical at small process size or low effectiveness. For example, in Lewis's MacErr dataset, a usability testing study with 15 participants, about 50% of problems (76 of 145) were discovered only once. This ratio indicates a large number of problems with low visibility, so it is unlikely that all of them would be discovered with a sample of only 15 users. Hence, the dataset may be incomplete.

Moreover, Lewis's approach was still based on Virzi's original formula, including its homogeneity assumption. In 2008, I showed that homoge-

a For a comprehensive view of the debate see Jeff Sauro's Web site http://www.measuringusability.com/blog/five-history.php

Figure 1. Binomial model fit of the Law and Hvannberg study[8] 169×169mm (72×72DPI).

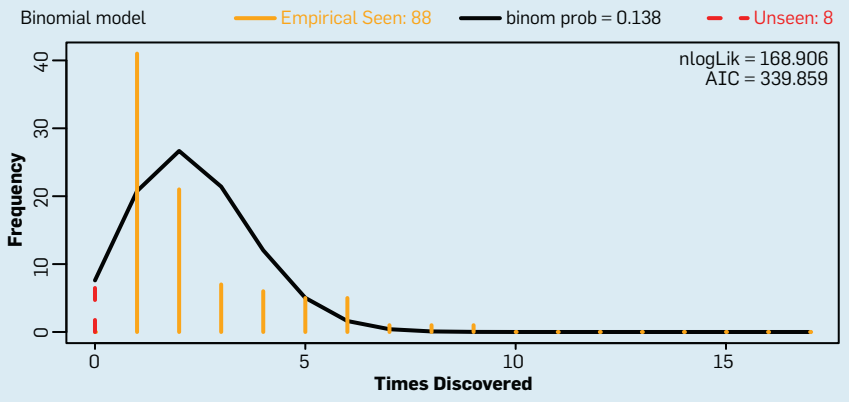
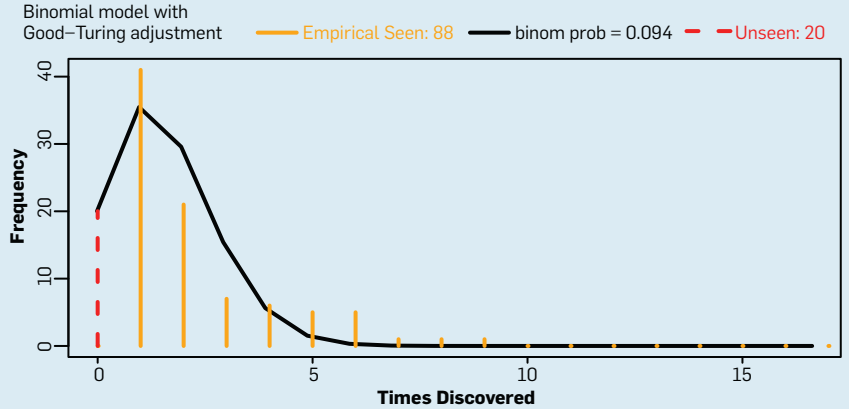


Figure 2. Binomial model fit with Good-Turing adjustment of the Law and Hvannberg study[8] 169×169mm (72×72DPI).



neity cannot be taken for granted.[17] Instead, visibility variance turned out to be the regular case, producing a remarkable effect; progress no longer follows the geometric series, moving instead much more slowly over the long term. The consequence of ignoring visibility variance and not accounting for incompleteness is the same; the progress of a study is overestimated, so the required sample size is underestimated.

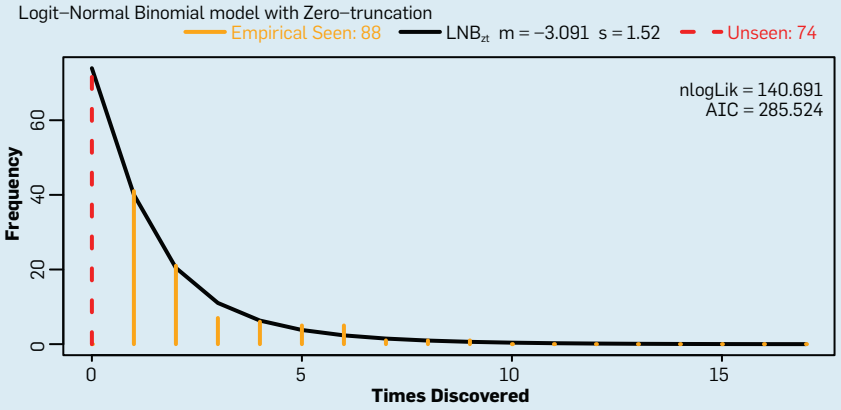
In their 2010 meta study, Hwang and Salvendy[6] analyzed the results of many research papers published since 1990 in order to define a general rule for sample size (replacing Nielsen's magic number five). Hwang's and Salvendy's minimum criterion for inclusion in their study was that a study reported average discovery rates, or number of successful problem discoveries divided by total number of trials (number of problems multiplied by number of sessions). However, this statistic may be inappropriate, as it neither accounts for incompleteness nor for visibility variance. Taking one reference dataset from the meta study as an example, I now aim to show how the 10±2 rule is biased. It turns out that the sample size required for an 80% target is much greater than previously assumed.

Seen and Unseen

In a 2004 study conducted by Law and Hvannberg,[8] 17 independent usability inspection sessions found 88 unique usability problems, reporting on the frequency distribution of the discovery of each problem. A first glance at frequency distribution reveals that nearly half the problems were discovered only once (see Figure 1). This result raises suspicion that the study did not uncover all existing problems, meaning the dataset is most likely incomplete.

In the study, a total of 207 events represented successful discovery of problems. Assuming completeness, the binomial probability is estimated as *p=207/(17*88)=0.138*. Using Virzi's formula, Hwang and Salvendy estimated the 80% target being met through 11 sessions, supporting their 10±2 rule. However, Figure 1 shows the theoretical binomial distribution is far from matching the observed distribution, reflecting three discrepancies:

Figure 3. Fit of the LNB_{zt} model on the Law and Hvannberg study[8] 169×169mm (72×72DPI).



Never-observed problems. The theoretical distribution predicts a considerable number of never-observed problems;

Singletons. More problems are observed in exactly one session than is predicted by the theoretical distribution; and

Frequent occurrences. The number of frequently observed problems (in more than five sessions) is undercounted by the theoretical distribution.

The first discrepancy indicates the study was incomplete, as the binomial model would predict eight unseen problems. The GT estimator Lewis proposed is an adjustment researchers can make for such incomplete datasets, smoothing the data by setting the number of unseen events to the number of singletons, here 41.[b] With the GT adjustment the binomial model obtains an estimate of *p=0.094* (see Figure 2). The GT adjustment lets the binomial model predict the sample size for an 80% discovery target at 16, which is considerably beyond the 10±2 rule.

Variance Matters

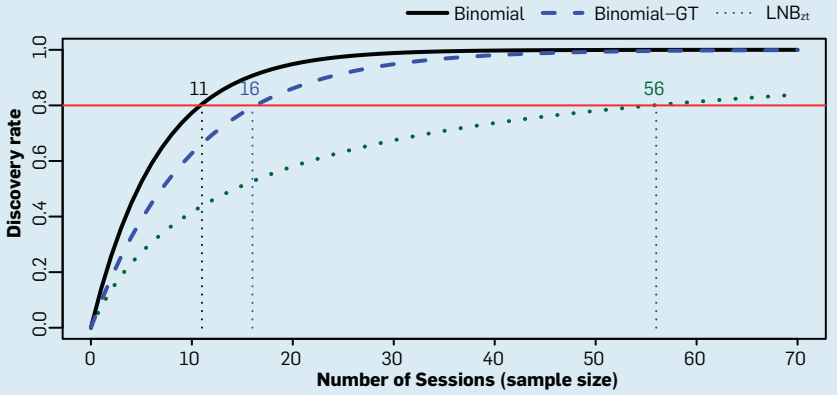
The way many researchers understand variance is likely shaped by the common analysis of variance (ANOVA) and underlying Gaussian distribution. Strong variance in a dataset is interpreted as noise, possibly forcing researchers to increase the sample size; variance is therefore often called a nuisance parameter. Conveniently, the Gaussian distribution has a separate parameter for variance, uncoupling it from the parameter of interest, the mean. That is, more variance makes the estimation less accurate but usually does not introduce bias. Here, I address why variance is not harmless for statistical models rooted in the binomial realm, as when trying to predict the sample size of a usability study.

Binomial distribution has a remarkable property: Its variance is tied to the binomial parameters, the sample size n and the probability p, as in $Var = np(1-p)$. If the observed variance exceeds $np(1-p)$ it is called overdispersion, and the data can no longer be taken as binomially distributed. Overdispersion has an interesting interpretation: The probability parameter p varies, meaning, in this case, problems vary in terms of visibility. Indeed, Figures 1 and 2 shows the observed distribution of problem discovery has much fatter left and right tails than the plain binomial and GT-adjusted models; more variance is apparently observed than can be handled by the binomial model.

Regarding sample-size estimation in usability studies, the 2006 edition of the *International Encyclopedia of Ergonomics and Human Factors* says, "There is no compelling evidence that a probability density function would lead to an advantage over a single value for p."[19] However, my own 2008–2009 results call this assertion into question. The regular case seems to be that p varies, strongly affecting the

b Lewis favors an equally weighted combination of normalization procedure and GT adjustment, but its theoretical justification is tenuous, ultimately making only a small difference to prediction (p=0.085).

Figure 4. Comparing process predictors on the Law and Hvannberg study[8] 169×169mm (72×72DPI).



progress of usability studies.[16,17] When problem visibility varies, progress toward finding new problems would be somewhat quicker in early sessions but decelerate compared to the geometric model as sample size increases. The reason is that easy-to-discover problems show up early in the study. When discovered, they are then frequently rediscovered, taking the form of the fat right tail of the frequency distribution. These reoccurrences increase the estimated average probability p but do not contribute to the study, as progress is measured only in terms of finding new problems. Moreover, with increased variance comes more intractable problems (the fat left tail), and revealing them requires much more effort than the geometric series model might predict.[c]

Improved Prediction

Looking to account for variance of problem visibility, as well as unseen events, I proposed, in 2009, a mathematical model I call the "zero-truncated logit-normal binomial distribution," or LNB_{zt}.[16] It views problem visibility as a normally distributed latent property with unknown mean and variance, so the binomial parameter p can vary by a probability density function—exactly what the encyclopedia article by Turner et al.[19] neglected. Moreover, zero-truncation accounts for the unknown number of never-discovered problems.

Figure 3 outlines the LNB_{zt} model fitted to the Law and Hvannberg dataset. Compared to the binomial model, this distribution is more dispersed, smoothly resembling the shape of the observed data across the entire range. It also estimates the number of not-yet-discovered problems at 74, compared to eight with the binomial model and 20 with GT adjustment, suggesting the study is only half complete.

The improved model fit can also be shown with more rigor than through visual inspection alone. Researchers can use a simple Monte-Carlo procedure to test for overdispersion.[d,17] A more sophisticated analysis is based on the method of maximum likelihood (ML) estimation. Several ways are available for comparing models fitted by the ML method; one is the Akaike Information Criterion (AIC).[2] The lower value for the LNB^{zt} model (AIC=286, see Figure 3) compared to the binomial model (AIC=340, see Figure 1) confirms that LNB^{zt} is a better fit with the observed data.[e]

The LNB_{zt} model also helps usability researchers predict the progress of the evaluation process through the derived logit-normal geometric formula.[16] For the Law and Hvannberg study[8] a sample size of n=56 participants is predicted for the 80% discovery target (see Figure 4), taking HCI researchers way beyond the 10±2 rule or any other magic number suggested in the literature.

Not So Magical

Using the LNBzt model since 2008 to examine many usability studies, I can affirm that visibility variance is a fact and that strong incompleteness usually occurs for datasets smaller than n=30 participants. Indeed, most studies I am aware of are much smaller, with only a few after 2001 adjusting for unseen events and not one accounting for visibility variance. The meta study by Hwang and Salvendy[6] carries both biasing factors—incompleteness and visibility variance—thus most likely greatly understating required sample size.

Having seen data from usability studies take a variety of shapes, I hesitate to say the LNB_{zt} model is the last word in sample-size estimation. My concern is that the LNB_{zt} model still makes assumptions, and it is unclear how they are satisfied for typical datasets "in the wild." Proposing a single number as the one-and-only solution is even less justified, whether five, 10, or 56.

Problem Population

Besides accounting for variance, the LNB_{zt} approach has one remarkable advantage over Lewis's predictor for required sample size: It allows for estimating the number of not-yet-discovered problems. The difference between the two approaches—LNB_{zt} vs. Lewis's adjustment—is that whereas Lewis's GT estimation first smooths the data by adding virtual data points for undiscovered problems, then estimates p, the LNB_{zt} method first estimates the parameters on the unmodified data, then determines the most likely number of unobserved problems.[16]

Recasting the goal from predicting sample size to estimating the number of remaining problems is not a wholly new idea. In software inspection, the so-called capture-recapture (CR) mod-

c Rephrasing this in terms of reliability engineering, the geometric series model becomes the discrete version of the exponential probability function, resulting in a stable hazard function for a problem's likelihood of being discovered. With visibility variance, the hazard function decreases over an increasing number of sessions.

d For a program and tutorial on the Monte-Carlo test for overdispersion see http://schmettow.info/Heterogeneity/

e The GT adjustment adds virtual data points so cannot be compared through AIC.

els have been investigated for managing defect-discovery processes; see, for example, Walia and Carver.[21] CR models are derived from biology, serving to estimate the size of animal populations, as in, for example, Dorazio and Royle.[4] Field researchers capture and mark animals on several occasions, recording each animal's capture history and using it to estimate the total number of animals in the area. Several CR models notably allow for the heterogeneous catchability of animals, usually referred to as Mh models. In inspection research, Mh models allow for visibility variance of defects, frequently helping predict the number of remaining defects better than models with a homogeneity assumption; see, for example, Briand et al.[1]

Also worth noting is that most studies in inspection research focus on a single main question: Have all or most defects been discovered or are additional inspections required? Sample-size prediction is rarely considered. In addition, the number of inspectors is often below the magic numbers of usability-evaluation research. One may speculate that software defects are easier to discover and possibly vary less in terms of visibility compared to usability problems. A detailed comparison of sample size issues in usability studies and management of software inspections has not yet been attempted.

The Timing of Control

The LNB_{zt} model promises to bridge these parallel lines of research, as it supports both goals: predicting sample size and controlling the process. Generally, three strategies are available for managing sample size:

Magic number control. Claims existence of a universally valid number for required sample size;

Early control. Denotes estimating sample size from the first few sessions, as introduced by Lewis[9]; and

Late control. Abstains from presetting the sample size, deciding instead on study continuation or termination by estimating the number of remaining problems; a decision to terminate is made when the estimate reaches a preset target, when, say, less than 20% of problems are still undiscovered.

An approach based on a magic number is inappropriate for prediction because usability studies differ so much in terms of effectiveness. Early control might seem compelling, because it helps make a prediction at an early stage of a particular study when exact planning of project resources is still beneficial; for example, a usability professional may run a small pilot study before negotiating the required resources with the customer. Unlike the late-control strategy, early control is conducted on rather small sample sizes. Hence, the crucial question for planning usability studies is: Do early sample-size predictors have sufficient predictive power?

Confidence of Prediction

The predictive power of any statistical estimator depends on its reliability, typically expressed as an interval of confidence. For the LNB_{zt} model the confidence intervals tend to be large, even at moderate sample size, and are too large to be useful for the early planning of resources; for example, the 90% confidence interval in the full Law and Hvannberg[8] dataset ranges from 37 to 165, for an 80% target. This low reliability renders the early-control strategy problematic, as it promises to deliver an estimate after as few as two to four sessions.[9]

Worth noting is that confidence intervals for the binomial model are typically much tighter.[16] However, tight confidence intervals are never an advantage if the estimator p is biased. There can be no confidence without validity. Fortunately, confidence intervals get tighter when the process approaches completeness and can serve as, say, a late-control strategy.

More Research Needed

The late-control strategy continuously monitors whether a study has met a certain target. Continuous monitoring may eventually enable usability practitioners to offer highly reliable usability studies to their paying customers. However, to serve projects with such strict requirements means any estimation procedure needs further evidence to produce accurate estimates under realistic conditions. The gold standard for assessing the accuracy of estimators is Monte-Carlo sampling, as it makes no assumptions about the shape of the probability distribution. Unfortunately, Monte-Carlo sampling requires complete datasets, implying huge sample sizes. Moreover, such studies must also cover a range of conditions. It cannot be expected that a study involving a complex commercial Web site has the same properties as a study testing, say, a medical infusion pump.

Several studies involving software inspection have validated CR models by purposely seeding defects in the artifacts being considered. This is another way to establish completeness, as the total number of seeded defects is known in advance. However, I doubt it is viable for usability studies. Usability problems are likely too complex and manifold, and designing user interfaces with seeded usability problems requires a substantial development effort and financial budget.

A conclusive approach, despite being lightweight, is to compare goodness-of-fit among various models, as I have tried to show here. A model that better fits the data is probably also superior at predicting a study's future progress. As another advantage, researchers may approach the task of picking a solid predictive model by reexamining existing datasets. However, such an examination requires access to the frequency distribution of problem discovery. Few studies report on that distribution, so, another meta study would require the cooperation of the original authors.

Industrial Applications?

To my knowledge, adoption of quantitative management is marginal in industrial usability studies. Objections seem to reflect two general themes: supporting different goals in the development process and interpreting raw observational data from the studies.

Reacting to Hwang and Salvendy,[6] Molich[11] said that rigid quality assurance is rarely the sole purpose of a usability study; such studies are often done as a kind of screening test to justify another redesign cycle. Accordingly, Nørgaard and Hornbæk found that industrial usability studies are often used to confirm problems that are already known.[15]

Molich[11] also advocated for a series of smaller studies driving an iterative design cycle, reflecting a

broad consensus among usability engineers. However, this approach barely benefits from quantitative control, as such small-scale studies do not strive for completeness. This view is also indirectly supported by John and Marks[7] showing that fixing usability problems is often ineffective and might even introduce new problems. Iterative design mitigates this issue by putting each redesign back into the loop. In the literature, the same study is often cited when the so-called downstream utility of usability evaluation is addressed. Downstream utility carries the effectiveness of usability studies beyond basic discovery of problems by focusing on effective communication and proper redesign guidance. However, such issues are admittedly of higher priority compared to the quantitative control of usability studies.

While the importance of sample size management depends on the context of the study, data quality is a precondition for a prediction to be of value. The models for estimating evaluation processes are based primarily on observing the reoccurrence of problems. Hence, for any observation to be counted it must first be clear to the researchers whether it is novel or a reoccurrence of a previously discovered problem. Several studies have shown only weak consensus on what constitutes a usability problem. Molich's comparative usability evaluation (CUE) series of studies (1998–2011) repeatedly found that any two professional teams running a usability study typically report results that differ in many respects; see, for example, Molich and Dumas.[12] Furthermore, the pattern of reoccurrence depends on the exact procedure to map raw observations onto defined usability problems.[5] All this means that estimations of sample size or remaining problems may lack objectivity because they depend on the often idiosyncratic procedures of data preparation.

Conclusion

Predicting the progress of a usability study is less straightforward than has been assumed in the HCI literature. Incompleteness and visibility variance mean the geometric series formula grossly understates required sample size. Most reports in the literature on usability evaluation effectiveness reflect this optimistic bias, as does the 10±2 rule of Hwang and Salvendy.[6] Consequently, I doubt that 80% of problems can be discovered with only 10 users or even with 10 experts. This limitation should also concern usability practitioners who test only a few participants in iterative design cycles. Most problems are likely to remain undiscovered through such studies.

As much as usability professionals and HCI researchers want a magic number, the very idea of identifying it is doomed to failure, as usability studies differ so much at identifying usability problems. Estimating a particular study's effectiveness from only a few early sessions is possible in theory, but such predictions are too unreliable to be practical. The late-control approach reflects potential for application domains where safety, economic, or political expectations make usability critical. Expensive, quantitatively managed studies can help develop high-quality interactive systems, reflecting that quality assurance was adequate. Most usability practitioners will likely continue to use strategies of iterative low-budget evaluation where quantitative statements are unreliable but also unnecessary. C

References

1. Briand, L.C., El Emam, K., Freimut, B.G., and Laitenberger, O. A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Transactions on Software Engineering 26*, 6 (June 2000), 518–540.
2. Burnham, K.P. and Anderson, D.R. Multimodel Inference. Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*, 2 (Nov. 2004), 261–304.
3. Caulton, D.A. Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology 20*, 1 (2001), 1–7.
4. Dorazio, R.M. and Royle, J.A. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics 59*, 2 (June 2003), 351–64.
5. Hornbæk, K. and Frøkjær, E. Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers 20*, 6 (Dec. 2008), 505–514.
6. Hwang, W. and Salvendy, G. Number of people required for usability evaluation: The 10±2 rule. *Commun. ACM 53*, 5 (May 2010), 130–133.
7. John, B. and Marks, S. Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology 16*, 4 (1997), 188–202.
8. Law, E.L.-C. and Hvannberg, E.T. Analysis of combinatorial user effect in international usability tests. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria, Apr. 24–29). ACM Press, New York, 2004, 9–16.
9. Lewis, J.R. Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction 13*, 4 (2001), 445–479.
10. Lewis, J.R. Testing small system customer set-up. In *Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society*. Human Factors Society, Santa Monica, CA, 1982, 718–720.
11. Molich, R. How many participants needed to test usability? *Commun. ACM 53*, 8 (Aug. 2010), 7.
12. Molich, R. and Dumas, J. Comparative usability evaluation (CUE-4). *Behaviour & Information Technology 27*, 3 (2008), 263–281.
13. Nielsen, J. *Why You Only Need to Test with 5 Users.* Jakob Nielsen's Alertbox (Mar. 19, 2000); http://www.useit.com/alertbox/20000319.html
14. Nielsen, J. and Landauer, T.K. A mathematical model of the finding of usability problems. In *Proceedings of INTERCHI 1993* (Amsterdam, the Netherlands, Apr. 24–29). ACM Press, New York, 1993, 206–213.
15. Nørgaard, M. and Hornbæk, K. What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the Sixth Conference on Designing Interactive Systems* (University Park, PA, June 26–28). ACM Press, New York, 2006, 209–218.
16. Schmettow, M. Controlling the usability evaluation process under varying defect visibility. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (Cambridge, U.K., Sept. 1–5). British Computer Society, Swinton, U.K., 2009, 188–197.
17. Schmettow, M. Heterogeneity in the usability evaluation process. In *Proceedings of the 22nd British HCI Group Annual Conference on Human-Computer Interaction* (Liverpool, U.K., Sept. 1–5). British Computer Society, Swinton, U.K., 2008, 89–98.
18. Spool, J. and Schroeder, W. Testing Web sites: Five users is nowhere near enough. *CHI Extended Abstracts on Human Factors in Computing Systems* (Seattle, Mar. 31–Apr. 5), ACM Press, New York, 2001, 285–286.
19. Turner, C.W., Lewis, J.R., and Nielsen, J. Determining usability test sample size. In *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed. CRC Press, Boca Raton, FL, 2006, 3084–3088.
20. Virzi, R.A. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society 34*, 4 (1992), 457–468.
21. Walia, G.S. and Carver, J.C. Evaluation of capture-recapture models for estimating the abundance of naturally occurring defects. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (Kaiserslautern, Germany, Oct. 9–10). ACM Press, New York, 2008, 158–167.
22. Woolrych, A. and Cockton, G. Why and when five test users aren't enough. In *Proceedings of the IHM-HCI Conference*, J. Vanderdonckt, A. Blandford, and A. Derycke, Eds. (Lille, France, Sept. 10–14). Cépadèus Éditions, Toulouse, France, 2001, 105–108.

Martin Schmettow (m.schmettow@utwente.nl) is an assistant professor in the Department Cognitive Psychology and Ergonomics of the University of Twente, Enschede, The Netherlands.