

# Why and When Five Test Users aren't Enough

*Alan Woolrych*

Informatics Centre  
St. Peters Campus  
University of Sunderland  
Sunderland, SR6 0DD  
alan.woolrych@sunderland.ac.uk

*Gilbert Cockton*

Informatics Centre  
St. Peters Campus  
University of Sunderland  
Sunderland, SR6 0DD  
gilbert.cockton@sunderland.ac.uk

## SUMMARY

Nielsen's claim that "Five Users are Enough" [5] is based on a statistical formula [2] that makes unwarranted assumptions about individual differences in problem discovery, combined with optimistic setting of values for a key variable. We present the initial Landauer-Nielsen formula and recent evidence that it can fail spectacularly to calculate the required number of test users for a realistic web-based test. We explain these recent results by examining the assumptions behind the formula. We then re-examine some of our own data, and find that, while the Landauer-Nielsen formula does hold, this is only the case for simple problem counts. An analysis of problem frequency and severity indicates that highly misleading results could have resulted when the number of required users is almost doubled.

Lastly, we identify structure and components of a more realistic approach to estimating test user requirements.

**KEYWORDS** : Usability testing, user differences.

## INTRODUCTION

User testing is widely recognised in the field of HCI as the most reliable way to achieve usability in a software system. However, user testing is expensive and time consuming. Each extra user potentially adds extra costs for recruitment, reward, the test session, data analysis and report synthesis. A cost/benefit balance must be used to determine how many users should test a system. If more than necessary are used, the cost of extra users will outweigh the benefits of the knowledge gained. Conversely, too few test users may miss key problems that render a system close to unusable. A magic formula is needed to tell us that  $x$  users are needed to find  $y\%$  of problems.

This paper presents Landauer & Nielsen's [2] longstanding formula from 1993 and briefly reviews the grounds for doubt from three sources: statistical theories, Spool & Schroeder's recent web study [3], and our own study of the effectiveness of Heuristic Evaluation [4]. These three objections are integrated into a theoretical model that forms the basis for more accurate cost-benefit calculations for user test planning.

## LANDAUER & NIELSEN'S FORMULA

Landauer & Nielsen [2] argued as few as 5 users can find 85% of usability problems. Their claim is based on an asymptotic formula that relates the likely yield from  $i$  test users to the number of known problems  $N$  and the probability  $\lambda$  of any user finding any problem.

$$\text{Problems found } (i) = N(1-(1-\lambda)^i)$$

Landauer & Nielsen [2] claim a typical value for  $\lambda$  to be 31% (.31). This is calculated for an individual study as the mean of the proportions of problems encountered by each user, i.e., for a total of 20 known problems, a user who encounters 10 will result in a value of 0.5. A user who encountered 16 would result in a value of 0.8. If they were the only two test users, the value of  $\lambda$  would be 0.65.

Plotting a curve for the typical value for  $\lambda$  of .31 and any number of known problems indicates that zero users reveal zero problems, 5 users will reveal 84%, and 15 users will reveal 100% of usability problems. It is on this basis that Nielsen claims that 5 users are enough [5].

This formula has been validated in many studies by Nielsen and colleagues, and it also fits our own data (not surprisingly, since study specific values for  $\lambda$  are computed from actual data!) However, this circularity in determining the key parameter ( $N$  is only significant in so far as it determines  $\lambda$ ) that significantly reduces its utility as a planning tool. Apparently, the only way to accurately establish the value of  $\lambda$  and  $N$  is by user testing. However, even though results here will be wholly dependent on the nature of the test (procedural, user and system variations), there are more fundamental problems with the formula.

## STATISTICAL CONSIDERATIONS

Landauer & Nielsen's formula can be derived from first principles from a very simple search problem. Suppose that there are  $N$  "hidden" items  $i_1 \dots i_n$  to be found by participants  $u_1 \dots u_p$ , then  $u_n$  may or may not find  $i_j$ . It is reasonable to suppose each item has an associated "visibility"  $\lambda_j$ , which is the probability that a "randomly chosen" participant will find it. Under the reasonable assumption that participants are independent, then, the

probability that, after  $n$  users,  $i_j$  remains undetected is  $(1 - \lambda_j)^n$ .

We can introduce an indicator  $F_j(n)$ , which takes the value 1 if  $i_j$  has been found after  $n$  users and is zero otherwise. The expected value of  $F_j(n)$ , given  $\lambda_j$ , is then  $1 - (1 - \lambda_j)^n$ . We can sum these expectations over all  $i_n$  to find the expected number of problems found after  $n$  participants,

$$\sum_{j=1}^n 1 - (1 - \lambda_j)^n$$

To get Landauer & Nielsen's formula, we must set  $\lambda_j = \lambda$  for all  $j$ . This is only reasonable when all items have an equal probability of being found. The shortest reflection will reveal that cannot be the case with a typical set of usability problems, and may be the reason why inferences made from a typical  $\lambda$  of 0.31 do not hold in many testing contexts. A more reliable model would have to replace the fixed value for  $\lambda$  with a probability density function that recognised variability in the "discoverability" of problems.

#### **SPOOL & SCHROEDER'S STUDY**

Spool & Schroeder [3] found in a recent study that only 35% of problems were found by the first 5 users. A low  $\lambda$  (our calculations indicate a theoretical value of 0.081, or 26% of Landauer and Nielsen's value) resulted from the interaction between a high number of problems (378) and a low chance of any one user finding a specific problem. This can be directly attributed to the test task, an unconstrained web shopping task (for a music CD of the user's choice) where users had real money to spend or keep to ensure ecological validity. One criticism of this approach is that users can choose any appropriate web site to buy from, and thus the task can be achieved in so many ways on so many sites (shops, portals and search engines in combination) that  $N$  must be high and  $\lambda$  must be low. This is true, but it is not an objection to the results, it is the reason for them and it is unavoidable.

From Spool & Schroeder's study, it is clear that Nielsen's claim that "5 users is enough" only holds when  $\lambda$  is 0.31. From the theoretical analysis earlier, it also only holds when the probability of every problem being found is very close to  $\lambda$ , i.e., there is little variance in the "discoverability" of problems. We checked these conditions against one of our own studies to advance the debate arising from Spool & Schroeder's study.

#### **RE-EXAMINING OUR OWN DATA**

Our user test data comes from an extensive study of Heuristic Evaluation [4]. Problems predicted by Heuristic Evaluation were used to derive task sets for

user testing. The aim of the study was to assess the scope and accuracy of HE. The methodology involved the strategic development of task sets that would be used in user testing, to stress potential problem identified by Heuristic Evaluation.

Our data results in a  $\lambda$  of 0.43, so in hindsight using Landauer & Nielsen's formula we need only have used 3 rather than 12 users, which would have identified 81% (13) of the 16 problems. We then further examined our data to identify what risks would have arisen had we only tested three users.

Our high lambda value was not wholly due to the simple application under test (the 1995 PowerPoint drawing editor), since the required drawing tasks introduced a high degree of potential task method variation (and thus high variance in problem discoverability). We would argue that the derivation of the test tasks from problem predictions increased the probability of probable problems being encountered by users.

Three users should have found 13 of our 16 known problems. The first three found 12, which is close. However, other selections of three users would not produce such good results, since the highest individual lambda value  $\lambda_j$  was .625 and the lowest was .25 (overall, the standard deviation is 0.11, which does not look high, but with the low numbers it is high enough). Now, if all our users were as good as the "best" at flushing out problems, just two test users would have found 86% of the known problems. Had they all been as poor as our "worst" problem finders, we would have needed six test users to find 86%. With a  $\lambda_j$  as low as the implied  $\lambda$  of 0.081 for the Spool & Schroeder study, 23 users would be required to achieve this!

In summary, Nielsen's 5 user claim can only be trusted when  $\lambda$  is 0.31 or higher and when the variance for individual users is low, how low we can't say yet, but this isn't the only problem!

A further problem arises when we consider the frequency and severity of problems, both critical attributes when prioritising solution recommendation and implementation. Even when the variance between users is "low" (whatever that means), varying distributions of problem frequency and severity could further undermine the Landauer-Nielsen formula. Problem impact (or severity), and frequency cannot be determined by inspection, nor are they independent of the number of test users. The frequency of a problem is dictated by the number of users it affects. Similarly, the severity of a problem can only be ascertained by analysis of the difficulties encountered by individual users with individual problem elements. To explore the impact of these variables, we examined our data further.

We had devised three categories of problem frequency (high medium and low) with their own specific criteria for classification:

- High frequency - problems encountered by more than three users.
- Medium frequency - problems encountered by two or three users.
- Low frequency - problems encountered by a single user.

There are three categories to describe problem impact (severe, nuisance and minor). The criteria devised for the classification of problem impact were:

- Severe impact – problems where the user wasted more than two minutes without progress. Users suffered task failure, or major impact on task quality (at the discretion of the analyst).
- Nuisance impact – problems where the user wasted up to two minutes without progress and/or suffered minor impact on task quality (at the discretion of the analyst).
- Minor impact – problem occurrences were when the user encountered a difficulty but made immediate recovery with no overall effect on task quality.

The most serious problems were those which had the highest aggregate of severe difficulties. For example, Problem 1 was classified as the most severe problem as 9 of the 12 users suffered severe impact determined by the above criteria. Problem 16 was the least severe problem as it was encountered by only a single user who only suffered minor impact as a result.

Figure 1 shows those problems found by specific users. The difference in shading determines the severity of difficulty encountered by each individual user. Black shading indicates severe difficulties, hatched squares indicate nuisance severity, whilst gray squares indicate minor severity.

Figure 1 ranks problems by severity (and within that, by frequency). Problem 1 is thus the most severe and Problem 16 the least severe. The matrix clearly demonstrates individual differences. For example, users who encountered Problems 2, 3 and 4 suffered the whole range of difficulties (from severe to minor), whilst five participants encountered no difficulty with Problem 3.

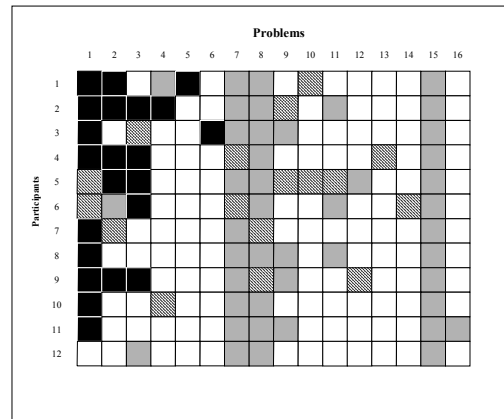


Figure 1: Problems and user matrix

We conduct our analysis of severity/frequency for our study generously, using five rather than the post-hoc optimum of three test users.

Using Figure 1 as a guide and assuming that Users 7, 8, 10, 11 and 12 were our 5 participants, the distribution of of problem severity would be changed. Problems 2 and 3 (2<sup>nd</sup> and 3<sup>rd</sup> most severe problems overall) would be recorded as only having nuisance and minor impact respectively, and of low frequency, and thus losing priority for design change when clearly more users with diverse abilities encountered much more severe problems.

### INTEGRATING THE THREE CRITIQUES

The Landauer-Nielsen formula makes risky assumptions about the exchangeability of test users. Nielsen's "five is enough" makes further assumptions about typical  $\lambda$  values. When the assumptions really are risky, and the complexity of the task is increased by user and/or system factors, then both the value for  $\lambda$  and the underlying formula become increasingly unreliable. The reliability of the Landauer-Nielsen formula rests on the exchangeability of test users for a given system, and thus ignores the many potential impacts of individual differences on user performance, impacts that have been described extensively by Dillon & Watson [1].

Individual differences, and their interaction with the system under test and the tasks chosen for the test will determine the variance across individual  $\lambda_j$  that in turn determines the reliability of a single overall  $\lambda$  value. The understanding of such diversity in users, task complexity, and the tool under test are key variables that are neglected in Landauer & Nielsen's formula, and can only rarely be successfully addressed by setting  $\lambda$  to a more realistic value (i.e., not 0.31!)

To improve on the Landauer-Nielsen formula, we need to replace  $\lambda$  with a probability density function

parameterised by values that represent beliefs about the likely impact of user, task and tool-under-test differences. It is the shape of the resulting distribution and not the scalar value of  $\lambda$  that will determine whether 5 users really are enough. We intend to examine possible functions with statistical colleagues. Such an approach however would have to become even more sophisticated to cope with the requirements of reducing error in determining problem frequency and severity.

## TWO FINAL TWISTS

In our study, after the first user tests, 98 further analysts in groups generated new predictions from Heuristic Evaluation. As a result, new task sets were designed and further users carried out these new tasks. As a result, two further usability problems emerged, confirming the importance of procedural variables within the user testing itself as key determinants of problem yield. This reinforces the need for a more sophisticated approach to the “visibility” of problems. More complex modelling approaches than simple scalar variables are thus essential. Not only are five users not enough, nor is one user test protocol!

Secondly, Nielsen ignores errors in usability problem extraction [6], which further complicates the relationship between the number of problems that actually get found (by the usability professional, since the evidence is clear that it is they and not users who ultimately find and keep problems). A more sophisticated model could introduce variables to reflect the risk of problem extraction failure. However, adding extra analysts may not be the solution, but instead more rigorous extraction procedures are more likely to achieve results [6].

## CONCLUSIONS

The chances of getting the right (or wrong) 5 users to accommodate Landauer & Nielsen’s formula depends on the distribution of the individual  $\lambda_j$  values. This, in turn, depends on individual differences between test users [1], the tool under test and the tasks performed during testing. The  $\lambda$  value appears to drop below the average established by Landauer & Nielsen [2] with both increased tool diversity and task complexity [3]. However, the use of usability inspection methods to provide a task focus for user testing can reduce the impact of task complexity or problem discoverability.

Only rarely can a simple overall  $\lambda$  value be reliable. What is really needed is a model that uses probability density functions and a more realistic set of variables, but not to predict how many problems some number of users will find. Instead, a more useful model would predict the probability that at least one problem remains

undetected, rather than an expectation of the number detected. This clearly depends on the variation among the  $\lambda$  values. For example, if an individual  $\lambda_j$  value is close to zero (which must be so for the Spool & Schroeder study given the very low implied  $\lambda$ , and even more so given their maximum  $\lambda_j$  of 0.16) then it is likely to remain undetected after many test users.

Lastly, even fixing the formula cannot reduce the inherent risks of counting problems without regard to their actual frequency and severity. A truly safe approach to estimating the need for test users must be sensitive to likely severity and frequency distributions, and not just to individual differences between test users. It is not enough to simply find problems — they must be understood and prioritised, and neither is possible without good frequency and severity data.

## ACKNOWLEDGEMENTS

We thank our colleague Malcolm Farrow for his many timely and extensive insights into the statistical issues associated with the user selection problem.

## BIBLIOGRAPHY

1. Dillon, A., and Watson, A. User Analysis in HCI : the historical lesson from individual differences research. In *International Journal of Human-Computer Studies* (1996) 45, 6, 619-637.
2. Landauer, T. K. and Nielsen, J. A Mathematical Model of the Finding of Usability Problems. In *Proc INTERCHI '93* 206-213
3. Spool, J. and Schroeder, W. “Testing Websites : Five Users is Nowhere Near Enough. In *Proc. CHI 2001, Extended Abstracts*, ACM 285-286
4. Cockton, G. and Woolrych, A., "Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation," to appear in *People and Computers XV*, Eds. A. Blandford and J. Vanderdonckt, Springer-Verlag, 2001
5. Nielsen, J., “Why You Only Need to Test With 5 Users”, Alertbox March 19, 2000, at <http://www.useit.com/alertbox/20000319.html>, accessed on 30/4/01
6. Cockton, G. and Lavery, D. "A Framework for Usability Problem Extraction,” in *INTERACT 99 Proceedings*, eds. A. Sasse and C. Johnson, 347-355, 1999.