DOI: 10.1145/1735223.1735255

BY WONIL HWANG AND GAVRIEL SALVENDY

Number of People Required for Usability Evaluation: The 10±2 Rule

USABILITY EVALUATION IS ESSENTIAL TO MAKE SURE THAT software products newly released are easy to use, efficient, and effective to reach goals, and satisfactory to users. For example, when a software company wants to develop and sell a new product, the company needs to evaluate usability of the new product before launching it at a market to avoid the possibility that the new product may contain usability problems, which span from cosmetic problems to severe functional problems.

Three widely used methods for usability evaluation are Think Aloud (TA), Heuristic Evaluation (HE) and Cognitive Walkthrough (CW). TA method is commonly employed with a lab-based user testing, while there are variants of TA methods, including thinking out aloud at user's workplace instead of at labs. What we discuss here is the TA method that is combined with a lab-based user testing, in which test users use products while simultaneously and continuously thinking out aloud, and experimenters record users' behaviors and verbal protocols in the laboratory. HE is a usability inspection method, in which a small number of evaluators find usability problems in a user interface design by examining an interface and judging its compliance with well-known usability principles, called heuristics. CW is a theory-based method, in which evaluators evaluate every step necessary to perform a scenario-based task, and look for usability problems that would interfere with learning by exploration. These three methods have their own advantages and disadvantages. For instance, TA method provides good qualitative data from a small number of test users, but laboratory environment may influence test user's behaviors. HE is a cheap, fast and easy-to-use method, while it often finds too specific and low-priority usability problems, including even not real problems. CW helps find mismatches between users' and designers' conceptualization of a task, but it needs extensive knowledge of cognitive psychology and technical details to apply.

However, even though these advantages and disadvantages show overall characteristics of three major usability evaluation methods, we cannot compare them quantitatively and see their efficiency clearly. Because one of reasons why so-called discounted methods, such as HE and CW, were developed is to save costs of usability evaluation, cost-related criteria for comparing usability evaluation are meaningful to usability practitioners as well as usability researchers. One of the most disputable issues related to cost of usability evaluation is sample size. That is, how many users or evaluators are needed to achieve a targeted usability evaluation performance, for example, 80% of overall discovery rate? The sample size of usability evaluation is known to depend on an estimate of problem discovery rate across participants.[11] The overall discovery rate is a common quantitative measure that is used to show the effectiveness of a specific usability evaluation method in most of usability evaluation studies. It is also called overall detection rate or thoroughness measure, which is the ratio of ‘the sum of unique usability problems detected by all experiment participants’ against ‘the number of usability problems that exist in the evaluated systems’, ranging between 0 and 1. The overall discovery rates were reported more than any other criterion measure in the usability evaluation experiments and also a key component for projecting required sample size for usability evaluation study.[7] Thus, how many test users or evaluators participate in the usability evaluation is a critical issue, considering its cost-effectiveness.

Table 1. Sources of Data

Sources		Usability evaluation experiments	Experiments (Papers) for data collection1)	Reference papers
Journals	International Journal of Human–Computer Interaction	5	3 (2)	Sears, 1997; Sears & Hess, 1999
	Behaviour & Information Technology	11	1 (1)	Fu, et al. 2002
	International Journal of Human–Computer Studies	5	.	.
	Human Factors	3	1 (1)	Andre, et al. 2003
	Others	4	2 (2)	Cuomo & Bowen, 1994; John & Mashyna, 1997
Proceedings	CHI Conference on Human Factors in Computing Systems	31	8 (4)	Lewis, et al. 1990; Mankoff, et al. 2003; Chattratichart & Brodie, 2004; Law & Hvannberg, 2004
	Human Factors Society Annual Meeting	14	2 (2)	Virzi, et al. 1993; Jacobsen, et al. 1998
	Nordic conference on Human-computer interaction	4	3 (2)	Law & Hvannberg 2002; Kjeldskov, et al. 2004
	IFIP INTERACT Conference on Human–Computer Interaction	7	3 (1)	Kjeldskov & Skov, 2003
	Others	9	.	.
Book Chapters		8	3 (1)	Desurvire, et al. 1992
Technical Report		1	1 (1)	Jacobsen & John, 2000
Total		102	27 (17)	

Notes, 1) Experiments for data collection should employ at least one of three major usability evaluation methods (i.e., TA, HE and CW) and report overall discovery rates.

There is No Consensus on Sample Sizes

There have been many discussions regarding optimal sample size of usability evaluation. Nielsen and Molich[9] reported that five evaluators found about $^{2}/_{3}$ of usability problems using HE, and Virzi[11] indicated that only four or five users are needed to detect 80% of usability problems when TA method is used for usability testing. With these empirical studies, so-called ‘4±1’[8] or ‘magic number five’ rule for detecting 80% of usability problems has been spread in usability evaluation community.

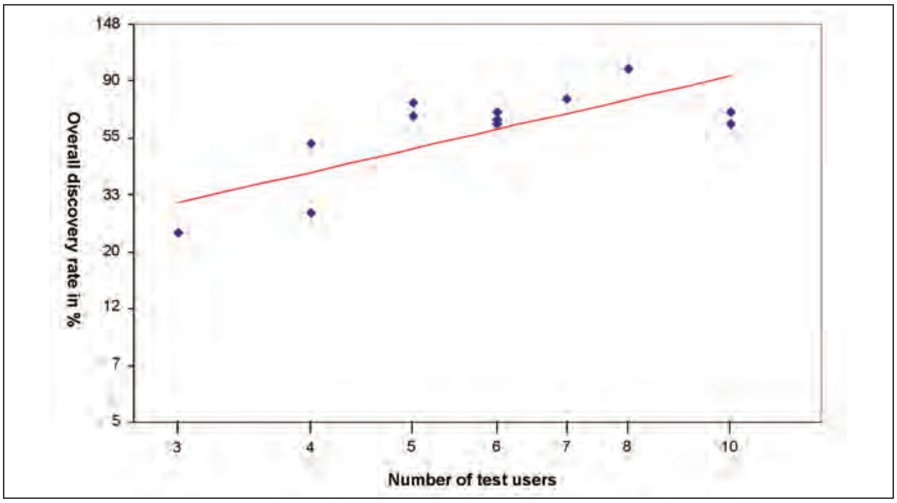
In the meanwhile, there have been other empirical results that do not or partially support Nielsen[8] and Virzi’s[11] conclusions. Lewis[6] partially supported their conclusions in that four or five subjects detected 80% of usability problems as long as the mean probability of detecting a problem by a subject (p) existed between 0.32 and 0.42. Law and Hvannberg[5] reported that 11 subjects were needed to reach 80% overall discovery rate when TA method was employed. Slavkovic and Cross[10] indicated that 5 ~ 10 evaluators as a sample size advocated by Nielsen and Molich[9] does not generalize to assessing complex interfaces when HE was conducted by novice evaluators. Caulton[1] argued that Virzi’s[11] five subjects is based on the assumption that all types of users have the same probability of encountering all usability problems (that is, homogeneity assumption). Thus, even though ‘4±1’ or ‘magic number five’ rule is very attractive when considering the cost of usability evaluation, it cannot be a general rule for optimal number of users or evaluators.

Meta-Analysis for Sample Size Issue

The experimental data for determining optimal sample sizes were collected from usability evaluation studies that satisfied the following two criteria: (a) empirical usability evaluation research was conducted using TA, HE and/or CW and (b) the number of participants as test users or evaluators and overall discovery rates were reported. Online academic databases, including ACM Digital Library, IEEE Xplore, and ScienceDirect, and offline sources were used to search for the relevant studies since 1990. In addition, all the references of the papers found as relevant studies were also checked to make sure that no relevant studies were missed. As a result of such extended search efforts without pre-selected sources of studies, most of major HCI-related journals, such as *International Journal of Human–Computer Interaction, Behaviour & Information Technology, International Journal of Human–Computer Studies*, and *Human Factors*, and the proceedings of major HCI-related conferences, such as *CHI Conference on Human Factors in Computing Systems* and *Human Factors Society Annual Meeting*, were included as the sources of relevant studies.

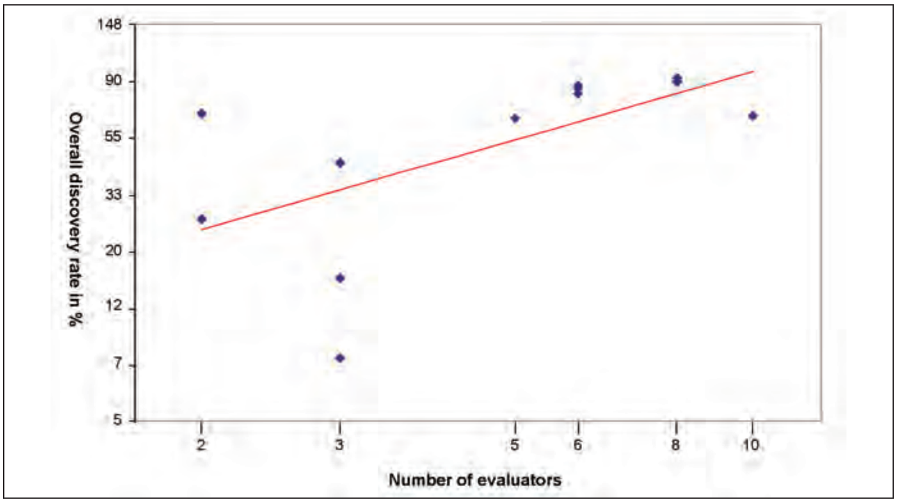
We found 102 usability evaluation experiments that reported quantitative results such as overall discovery rate, severity of problems, and other statistics, but usability evaluation experiments that satisfies the above two criteria were 27 experiments out of 102 experiments. As seen in Table 1, the collected usability evaluation experiments may represent the accumulated empirical results of usability evaluation research, because these sources of experiments included most of major HCI-related journals and conference proceedings. In addition, the collected experiments were fairly evenly distributed from 1990 to 2004 by their published years, meaning that the collected experiments represent most of

Figure 1. Regression Line of Overall Discovery Rates based on TA data (log scales)



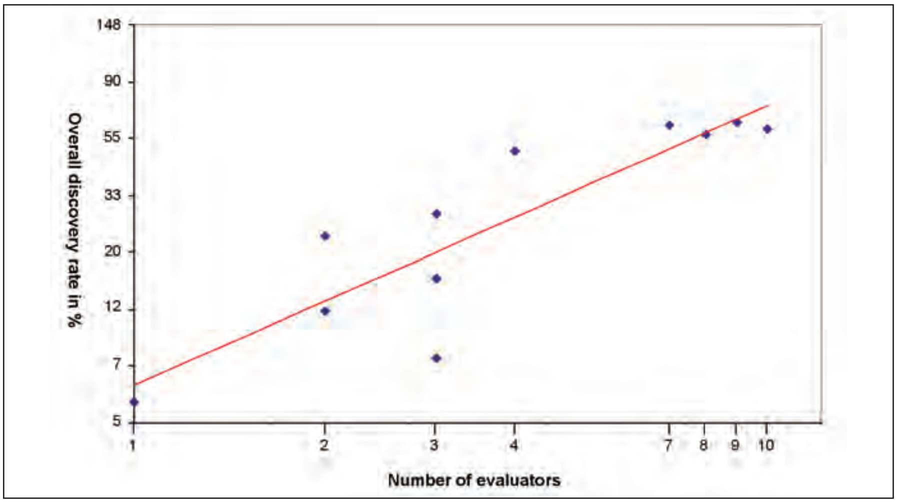
(Solid line: regression line, ln(overall discovery rates in %) = 2.42009 + 0.92425 * ln(number of test users), R^2 = 0.6267; and dots: observations in natural log scale)

Figure 2. Regression Line of Overall Discovery Rates based on HE data (log scales)



(Solid line: regression line, ln(overall discovery rates in %) = 2.60263 + 0.86183 * ln(number of evaluators), R^2 = 0.3656; and dots: observations in natural log scale)

Figure 3. Regression Line of Overall Discovery Rates based on CW data (log scales)



(Solid line: regression line, ln(overall discovery rates in %) = 1.83495 + 1.06617 * ln(number of evaluators), R^2 = 0.7751; and dots: observations in natural log scale)

empirical results of usability evaluation research over the past 15 years.

Using overall discovery rates and number of test users or evaluators as main variables, 36 data points were extracted for analysis from 27 experiments because an experiment could report more than one overall discovery rate, corresponding to more than one usability evaluation method that were employed in the experiment. In order to find a more general rule regarding number of test users or evaluators needed to reach 80% overall discovery rate, linear regression analysis was conducted based on the 36 data points from 27 experiments. Because overall discovery rates has a non-linearly increasing tendency when the number of test users or evaluators (that is, sample size) increase, natural logarithm transformation was utilized for overall discovery rate and number of test users or evaluators. As seen in Figure 1 through Figure 3, linear regression analyses were conducted with three different data sets, such as TA data, HE data, and CW data.

Prediction of Optimal Sample Sizes. The regression lines are well fitted on the data sets except HE data (R^2 = 0.63 for TA data; R^2 = 0.37 for HE data; and R^2 = 0.78 for CW data). Based on the regression equations, the natural logarithmic values of overall discovery rates are predicted with TA, HE and CW data set each. The predictions of number of test users or evaluators to reach 80% overall discovery rate are different from the results of previous individual experiments (see Table 2). When TA is used, 9 users are needed to detect 80% of usability problems, whereas 8 evaluators are needed when HE is employed. These results are far from the results of Nielsen,[8] and Nielsen and Molich,[9] but similar to Law and Hvannberg's[5] result. When CW is used, the prediction from linear regression indicates 11 evaluators for detecting 80% of usability problems. This result seems to be due to the characteristic of CW, which tends to detect severe usability problems but less number of usability problems. The estimation based on Hertzum & Jacobsen[3] also shows that 13 evaluators needed to detect 80% of usability problems when CW is used, which are more than those from TA or HE.

In addition to prediction of optimal sample size to reach the target performance of overall discovery rate (80%) based on linear regression, the investigation of outliers that are far from the fitted regression line may give other implications. There are three outliers, whose x (number of test users or evaluators) and y (overall discovery rate in %) coordinates are (2, 68.3) and (3, 8) in HE data (see Figure 2), and (3, 8) in CW data (see Figure 3). The first outlier in HE data has too high overall discovery rate (68.3%) compared to number of evaluators (*n*=2).[4] It came from the use of HE for six hours by a computer science and usability specialist and a cognitive psychologist specialized in HCI, who evaluated experimental version of interfaces. The second outlier in HE data has too low overall discovery rate (8%) compared to number of evaluators (*n*=3)[2]. HE was used by three non-experts like participants in lab studies, and three hours were given for doing six tasks and reporting usability problems in specific forms. The outlier in CW data came from the same source paper as the second outlier in HE data, and it has too low overall discovery rate (8%) compared to number of evaluators (*n*=3). CW was conducted by three non-experts like participants in lab studies, and three hours were given for doing three tasks and usability problems were reported through automated CW reports.

From the investigation of outliers' experimental conditions, it is concluded that evaluator's expertise, duration of evaluation and report format may affect the overall discovery rate. The first outlier of too high overall discovery rate in HE data indicates that usability or HCI specialists evaluated relatively simple interfaces for enough duration of evaluation, whereas the second in HE data and the outlier in CW data that show too low overall discovery rate indicate that non-experts evaluated interfaces for relatively short duration and reported usability problems in special formats, such as structured forms, automated report, and diary. As seen in Figure 2, HE showed relatively bigger variance, including two outliers, than TA and CW. It is partly because of different heuristics. Even though most of HE experiments tended to use Nielsen's 10 heuristics basically, some of experiments modified existing heuristics or added additional heuristics to reflect the characteristics of evaluated systems and to improve the performance of HE. Thus, different set of heuristics may results in different performance of HE.

Table 2. Number of Test Users or Evaluators needed in Usability Evaluation

References	Usability evaluation methods	*p* [1)]	Number of test users or evaluators needed to reach 80% overall discovery rate
Nielsen[8]	TA	0.282	5
Law and Hvannberg[5]	TA	0.140	11
Nielsen and Molich[9] - Mantel case	HE	0.380	4[2)]
Hertzum & Jacobsen[3]	CW	0.121	13[2)]
Predictions from regression analysis in this study	TA	-	9
	HE	-	8
	CW	-	11

Notes, 1) p is mean probability of detecting a problem by a test user or an evaluator
2) Number of evaluators were not reported from the original studies, and thus estimated based on p in this study.

Conclusion

Based on the predictions using the observed data with a variety of experimental conditions, a general rule for optimal sample size (such as, test users for TA and evaluators for HE and CW) would be '10±2' instead of '4±1.' That is, TA, HE and CW methods require nine test users, eight evaluators and 11 evaluators, respectively, to reach 80% overall discovery rate, rather than 3 ~ 5 evaluators. Thus, the use of TA and HE is recommended to detect 80% of usability problems if small number of test users or evaluators is desirable, whereas we can expect that CW detect severer problems with a larger number (for example, 11 evaluators) of evaluators than TA and HE. In the meanwhile, the general rule of '10±2' means that optimal sample sizes of '10±2' can be applied to a general or basic evaluation situation, for example, just basic training provided to evaluators and a limited evaluation time allowed. As shown in outlier analysis, if we want to reduce sample size while still reaching a targeted overall discovery rate in usability evaluation, evaluation conditions, such as evaluators' expertise, evaluation duration, task type, report type and so on, other than sample size should be improved. In addition, we can consider mixed methods, such as HE and TA. Even though there is a tendency that HE is used in the early phase and TA is utilized in the later phase of design development process, the combination of HE and TA seems to require smaller number of evaluators and test users than individual use of HE and TA to reach target performance because of their complementary characteristics. We expect its empirical evidence in the future studies.

References

1. Caulton, D. A. Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology 20*, 1 (2001), 1-7.
2. Desurvire, H.W., Kondziela, J. M. and Atwood, M. E. What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, and M.D. Harrison (Eds.), *People and Computers Volume VII*. Cambridge, England: Cambridge University Press, 1992, 89–102.
3. Hertzum, M. and Jacobsen, N.E. The evaluator effect during first-time use of the CW technique. In H.-J. Bullinger&J. Ziegler (Eds.), *Human–Computer Interaction: Ergonomics and User Interfaces* (Vol. 1.), London: Lawrence Erlbaum Associates, Inc., 1999, 1063–1067.
4. Law, L.-C. and Hvannberg, E.T. Complementarity and convergence of HE and usability test: A case study of universal brokerage platform. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, ACM, 2002, 71-80.
5. Law, L. -C. and Hvannberg, E.T. Analysis of combinatorial user effect in international usability tests. In *CHI Conference on Human Factors in Computing Systems*, ACM, 2004, 9–16.
6. Lewis, J.R. Sample sizes for usability studies: Additional considerations. *Human Factors 36*, 2 (1994), 368-378.
7. Lewis, J.R. Evaluation procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction 13*, 4 (2001), 445-479.
8. Nielsen, J. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human–Computer Studies 41*, (1994), 385–397.
9. Nielsen, J. and Molich, R. HE of user interface. In *CHI '90 Conference Proceedings*. ACM, 1990, 249-256.
10. Slavkovic, A. and Cross, K. Novice HEs of a complex interface. In *CHI '99 extended abstracts on Human Factors in Computing Systems*, ACM, 1999, 304–305.
11. Virzi, R.A. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors 34*, (1992), 457-468.

Wonil Hwang (wonil@ssu.ac.kr) is an assistant professor at Department of Industrial and Information Systems Engineering at Soongsil University in Seoul, Korea.

Gavriel Salvendy (salvendy@purdue.edu) is a professor at School of Industrial Engineering at Purdue University in West Lafayette, Indiana, and chair professor and head of Department of Industrial Engineering at Tsinghua University in Beijing, P.R. China.