Usability Problems Experienced by Different Groups of Skilled Internet Users: Gender, Age, and Background

Jane Billestrup(✉), Anders Bruun, and Jan Stage

Department of Computer Science, Aalborg University,
9220 Aalborg East, Denmark
{jane,bruun,jans}@cs.aau.dk

Abstract. Finding the right test persons to represent the target user group, when conducting a usability evaluation is considered essential by the HCI research community. This paper explores data from a usability evaluation with 41 participants with high IT skills, to examine if age, gender, and job function or educational background, has an impact on the amount and types of usability problems experienced by the users. All usability problems were analysed and categorised through closed coding, to group the test persons differently in relation to gender, age, and job function or educational background. The study found that the usability problems experienced across gender, age group and job function or educational background, are approximately the same. This indicates that the usual characteristics of test persons, might not be as important, and opens up for further research in regards to, if users with different skill levels, in regards to internet usage, might be more applicable.

Keywords: Usability evaluation · Test persons · Demography

1 Introduction

Usability evaluation is a strong tool for identifying areas of an interactive system that need improvement. In practice, one of the key challenges for usability evaluators is to find users that can participate as tests subjects. Recruitment of test subjects is challenging, and the time required for test sessions and the subsequent data analysis is usually dependent on the number of the number of test subjects. Therefore, there have been attempts to determine the minimal number of test users required for a usability evaluation [4, 7, 11].

Combining Other researchers have criticised these attempts to define the minimal number. One of the arguments is that different users experience different usability problems [6, 9]. In these discussions, there has been little evidence as to the actual differences between the usability problems experienced by different groups of users.

For specialised systems that are used by a homogeneous group of users, this issue is not particularly relevant. However, for systems that are aimed at diverse and heterogeneous groups of users, it is highly relevant.

© IFIP International Federation for Information Processing 2016
Published by Springer International Publishing Switzerland 2016. All Rights Reserved
C. Bogdan et al. (Eds.): HCSE 2016/HESSD 2016, LNCS 9856, pp. 45–55, 2016.
DOI: 10.1007/978-3-319-44902-9_4

This paper presents results from an exploratory study of the usability problems experienced by different users. The focus of this study was to what extent different test persons, who are all experienced internet users, experience different types of usability problems, across gender, age, and educational background or job function.

The system we evaluated was a government data dissemination website aimed at a very broad user population. In the following section, the related work is presented, followed by a description of the method used for data collection and analysis. Then the results are presented, and finally, the results are discussed and concluded upon.

2 Related Work

The question about the number of test subjects needed in a usability evaluation has been discussed for many years. Virzi [11] focused on the need exists to reduce the cost of applying good design practices, such as user testing, to the development of user interfaces. He was one of the first to experiment with the number of test subjects needed. Over a series of 3 experiments, he found that 80 % of the usability problems were detected with four or five subjects, additional subjects were less and less likely to reveal new information, and the most severe usability problems were likely to be detected with the first few subjects. In the experiments, he used test subjects who were from the surrounding community or undergraduate students. There is no further description of their demography.

Lewis [7] emphasices that the aim of a usability evaluation is to have representative participants. He reports from an experiment with fifteen employees of a temporary help agency who all had at least three months' experience with a computer system but had no programming training or experience. Five were clerks or secretaries and ten were business professionals. In this study, using five participants uncovered only 55 % of the problems. To uncover 80 % of the problems would require 10 participants. The results show that additional participants discover fewer and fewer problems. The most important result was that problem discovery rates were the same regardless of the problem severity. Again, there is no concern for the demography of the test subjects.

Caulton [2] argues that the results obtained in these early experiments were based on the assumption that all types of users have the same probability of encountering all usability problems, and he denotes this as the homogeneity assumption. If that is violated, more subjects are needed. He argues that with heterogeneous user groups, problem detection with a given number of subjects is reduced. The more subgroups, the lower the proportion of problems expected. If ten unknown user subgroups exist, 50 randomly sampled subjects should yield 80 % of the problems.

Law and Hvannberg [6] have worked more on the influence of subgroups on problem detection through an experiment with usability tests conducted in four different European countries. They conclude that the heterogeneity of subgroups in a test will dilute the problem detection rate. Not only for severe problems but also for moderate and minor ones, the diluting effect implied a reduction. The problem detection rate for the severe problems is significantly higher than for the less severe, but the absolute value for the severe problems is not particularly high. Between nine and ten participants were required to uncover 80 % of the severe problems, whereas

15 participants were required to uncover 80 % of the minor problems. In addition, they found no significant correlation between problem detection rate and problem severity level. Based on their results, they reject that so-called “magic five” assumption as 11 participants were required to obtain 80 % of the usability problems.

More recently, there has been another attempt to define a specific “magic” number [4]. This new attempt has been criticised for being flawed [9]. A detailed analysis has been made of the use of the “magic five” assumption. None of these or the previous references in this area have explored in more detail how heterogeneous different subgroups are and how different user groups experience different usability problems.

3 Method

We have conducted an exploratory study of usability problems experienced by different user groups. This section describes how the data was collected and analysed.

3.1 Data Collection

The data was gathered through a usability evaluation of a data dissemination website (dst.dk). This site provides publicly available statistics about the population (e.g. educational level or IT skills), the economy, employment situation, etc.

Test Persons. All test persons were invited through emails distributed across the university. For this study data from 41 usability evaluations were included. The test persons consist of 12 faculty members from Ph.D students to professors, from different departments, 15 students in technical or non-technical educations, and 14 participants from technical and administrative staff from different departments. All participants received a gift with a value of approximately 20 USD for their participation. An overview of the participants can be seen in Table 1 on the following page.

All test persons were placed in one of six groups in regards to gender and age. The test persons varied in age between 21 and 66 years and consisted of 19 males and 22 females. All test persons were asked to assess their own skill level in regards to Internet usage on a scale from 1 to 5, where 1 was the lowest and 5 the highest score. The average for each group is shown in the table, none of the 41 test persons assessed themselves lower than 4. Originally 43 usability evaluations were conducted, but the data from two usability evaluations were excluded from this study, due to these test persons assessed themselves at skill level 3 in regards to Internet usage. All test persons were asked if they were familiar with, and used this website. 19 people answered that they had never used the website, 20 answered that they were familiar with the site and used it approximately once a year, and, two people answered that they use the website approximately once a month.

Usability Evaluations. All tests were conducted as think-aloud evaluations in a usability laboratory. The test monitor and test person were placed in different rooms and communicated through microphone and speakers in order to avoid the possibility of the test moderator’s body language or other visible expressions, influencing each test person. All test persons were asked to fill out a short questionnaire after the test in regards to their participation.

Tasks. Each user solved eight tasks all varying in difficulty. Examples of this were that the first task was to find the total number of people living in Denmark while a more difficult task was to find the number hotels and restaurants with one single employee in a particular area of Denmark.

Data Handling. All usability evaluations were recorded and the collected recordings were analysed by conducting video analysis. All recordings were analysed by two evaluators. Both evaluators had extensive previous experience in analysing video data. The videos were analysed in different random order, to reduce possible bias from learning.

Table 1. Demography for the 41 test persons.

Number of people in each category	Age	Age average	Gender	Backgrounds	Average Internet experience
6	< 27	24	M	5 Computer Science students 1 Computer Science faculty member	5
8	< 27	22	F	5 Computer science students 2 humanities students 1 office trainee	4.6
8	27–44	36	M	4 computer science faculty members 1 social science faculty member 1 technical staff 1 administrative staff 1 engineering student	4.8
8	27–44	38	F	6 administrative staff 1 social science faculty member 1 information science student	4.3
5	44 <	55	M	3 computer science faculty members 1 faculty member medicine 1 technical staff	4.8
6	44 <	50	F	4 administrative staff 1 faculty member computer science 1 faculty member medicine	4.5

The following characteristics were used to determine a usability problem;

(A) Slowed down relative to their normal work speed
(B) Inadequate understanding e.g. does not understand how a specific functionality operates or is activated
(C) Frustration (expressing aggravation)
(D) Test moderator intervention
(E) Error compared to correct approach.

The data handling resulted in a list of 147 usability problems after duplicates had been removed. To determine similarities between problems from each list, the usability problems found by each evaluator were discussed. Across the analysis, the evaluators had an any-two agreement of 0.44 (SD = 0.11), which is relatively high compared to other studies [3]. Further information about the data collection can be found in [1].

Data Analysis. We also uncovered which types of usability problems that were experienced by the different groups of participants. We did this through closed coding [10] where each problem was categorised according to the 12 types listed in Nielsen et al. [8]. Two of the authors conducted this coding and did so independently of each other. It was decided in advance that the raters would code all and only use the data from the codings where the authors agreed on the category independently of each other. An interrater reliability analysis using the Fleiss Kappa statistic was performed to validate the result. This determines the level of consistency among the two raters. The result of was a moderate level of agreement (Kappa = 0.44, $p < 0.001$, 95 % CI =0.37, 0.52) [5]. The 12 categorised used for this study are described next.

Affordance relates to issues on the user's perception versus the actual properties of an object or interface.

Cognitive load regards the cognitive efforts necessary to use the system.

Consistency concerns the consistency in labels, icons, layout, wording, commands etc. on the different screens.

Ergonomics covers issues related to the physical properties of interaction.

Feedback regards the manner in which the interface relays information back to the user on an action that has been done and notifications about system events.

Information covers the understandability and amount of information presented by the interface at a given moment.

Interaction styles concern the design strategy and determine the structure of interactive resources in the interface.

Mapping is about the way in which controls and displays correlate to natural mappings and should ideally mimic physical analogies and cultural standards.

Navigation regards the way in which the users navigate from screen to screen in the interface.

Task flow relates to the order of steps in which tasks ought to be conducted.

User's mental model covers problems where the interactive model, developed by the user during system use, does not correlate with the actual model applied to the interface.

Visibility regards the ease with which users are able to perceive the available interactive resources at a given time.

The coding and analysis by two raters resolved in a list of 83 coded usability problems, out of originally 147 usability problems. This reduction happened as all usability problems where the raters did not agree on the category was removed from the study.

These categorisations were used to distinguish if test persons experienced the same type of usability problems, or if there were deviations across gender, age, job function or educational background. The results of this analysis are presented in the following section.

4 Results

In this section, we present the results from conducting this study. The results are presented from four different perspectives. First, the test persons are divided into males and females, then into the three age groups without taking the gender into perspective, then, the test persons are divided into groups both in regards to age and gender, and finally, the test persons are divided into groups in regards to education or work function. This was conducted to show if gender, age or background plays a role in regards to differences in the perceiving of usability problems. The numbers shown in the tables in the result section represent an average number of usability problems found per test person in each category. This was conducted to be able to compare groups containing different numbers of test persons, and still make the numbers comparable.

The results show that problems were found in regards to five of the twelve closed codings. Affordance, Cognitive Load, Feedback, Information, and Visibility, respectively. As problems were not found relating to Consistency, Ergonomics, Interaction Styles, Mapping, Navigation, User's Mental Model, and Task Flow, these categorisations will not be mentioned further.

Note that all results are based on the number of problems to which the two raters agreed on the categorisations, e.g. if the two raters did not agree on the code of a particular problem, this was excluded from the result. Out of the total 147 problems the raters agreed on 83.

4.1 Gender

We analysed whether males and females with similar skills in regards to internet usage experienced the same amount and type of usability problems. The results are presented in Table 2.

An independent samples t-test revealed no significant differences in the total number of experienced between the genders ($t = -0.9$, $df = 39$, $p > 0.2$). We did, however find significant differences when considering the problem types related to feedback ($t = -1.2$, $df = 10$, $p < 0.01$) and information ($t = -1.8$, $df = 39$, $p < 0.01$).

Table 2. The average number of usability problems experienced when dividing the test persons by gender.

Group statistics					
Gender		N	Mean	Std. deviation	Std. error mean
Affordance	M	10	1,40	0,516	0,163
	F	11	1,36	0,674	0,203
Cognitive load	M	19	2,32	2,126	0,490
	F	22	3,77	2,159	0,460
Feedback	M	7	1,00	0,000	0,000
	F	5	1,20	0,447	0,200
Information	M	19	3,58	1,610	0,369
	F	22	4,95	2,952	0,629
Visibility	M	17	2,00	1,225	0,297
	F	19	1,58	0,769	0,176
Total	M	19	9,79	3,896	0,894
	F	22	11,05	4,904	1,045

4.2 Age

We also analysed if age had an impact on the experienced amount of usability problems. The results are presented in Table 3 on the following page.

A one-way ANOVA test revealed no significant differences in number of experienced problems between the three age groups (F = 1.02, df = 40, p > 0.3).

Table 3. Usability problems experienced by different age groups.

		N	Mean	Std. deviation
Affordance	<27	5	1,40	0,548
	27–44	9	1,56	0,726
	>44	7	1,14	0,378
	Total	21	1,38	0,590
Cognitive load	<27	14	3,79	2,326
	27–44	16	3,81	2,257
	>44	11	2,91	1,700
	Total	41	3,56	2,134
Feedback	<27	5	1,00	0,000
	27–44	6	1,17	0,408
	>44	1	1,00	
	Total	12	1,08	0,289
Information	<27	14	5,29	2,555
	27–44	16	4,19	2,562
	>44	11	3,27	2,005
	Total	41	4,32	2,494

(*Continued*)

Table 3. (*Continued*)

		N	Mean	Std. deviation
Visibility	<27	13	1,62	0,961
	27–44	13	1,69	0,855
	>44	10	2,10	1,287
	Total	36	1,78	1,017
Total	<27	14	11,43	4,767
	27–44	16	10,69	4,771
	>44	11	8,91	3,419
	Total	41	10,46	4,456

4.3 Job Function and Educational Background

Finally, we analysed if a large number of test persons with a background in computer science had an impact in regards to the amount of usability problems experienced. The results are presented in Table 4.

The table shows, that when dividing the test persons into job function or educational background, students which are not in computer science, experience more problems related to cognitive load and information. A one-way ANOVA test revealed no significant differences in the total number of problems experienced across job function or educational background ($F = 0.6$, $df = 40$, $p > 0.6$).

Table 4. The average amount of usability problems experienced when dividing the test persons in regards to job function or educational background.

		N	Mean	Std. deviation
Affordance	Other students	2	1,00	0,000
	CS students	3	1,67	0,577
	TAP	10	1,50	0,707
	CS faculty	3	1,00	0,000
	Other faculty	3	1,33	0,577
	Total	21	1,38	0,590
Cognitive load	Other students	4	5,75	0,500
	CS students	11	3,27	2,195
	TAP	15	3,47	1,846
	CS faculty	7	4,00	2,887
	Other faculty	4	1,75	0,500
	Total	41	3,56	2,134
Feedback	Other students	0		
	CS students	3	1,00	0,000
	TAP	3	1,00	0,000
	CS faculty	4	1,00	0,000

(*Continued*)

Table 4. (*Continued*)

		N	Mean	Std. deviation
	Other faculty	2	1,50	0,707
	Total	12	1,08	0,289
Information	Other students	4	5,00	3,559
	CS students	11	4,55	2,067
	TAP	15	4,87	3,021
	CS faculty	7	3,14	1,069
	Other faculty	4	3,00	1,826
	Total	41	4,32	2,494
Visibility	Other students	3	1,33	0,577
	CS students	10	1,60	1,075
	TAP	14	1,79	0,802
	CS faculty	6	2,33	1,506
	Other faculty	3	1,67	1,155
	Total	36	1,78	1,017
Total	Other students	4	12,25	4,031
	CS students	11	10,00	4,123
	TAP	15	11,20	5,003
	CS faculty	7	10,14	4,140
	Other faculty	4	7,75	4,787
	Total	41	10,46	4,456

5 Discussion

This study has focused on comparing the amount of usability problems found when grouping the test persons in regards to gender, age, and job function or educational background. This was conducted as all test persons assessed themselves as experienced internet users, as each rated themselves as either 4 or 5 on a scale from 1 to 5, where five was the highest score. This way, it could be explored if test persons of a high degree of internet skills experienced different types of usability problems, or if they could be considered a homogeneous group, where neither age, gender, and job function or educational background made a difference in regards to the average amount of usability problems.

5.1 Comparison with Related Work

Related work has shown that the amount of needed test persons varies [7, 11]. As demographical data was not included in these studies it is not possible for us to draw any conclusions in relation to the results from this study, though it raises the question of, if the test persons chosen by Virzi [11] were more homogeneous than the test persons chosen by Lewis [7] in regards to the skills of Internet usage or IT in general.

This study has found indications that a user group can be homogeneous though a variety in age and background. Our results indicated that the test persons from this study experience around the same amount of usability problems in regard to each categorization (Affordance, Cognitive Load, Feedback, Information, Visibility), across gender, age, and background. This corresponds with Caultons' conclusions about homogeneous user groups experiencing the same usability problems [2].

This study shows no greater difference in regards to the types of usability problems experienced by the test persons. This does not correspond with the findings of Law and Hvannberg who concluded that the heterogeneity of subgroups in a test will dilute the problem detection rate [6].

5.2 Implications for Usability Practitioners

Though further research is needed, this study indicates that recruiting test persons across gender, and age might not be necessary, as these findings show that users with approximately the same level of skills in regards to Internet usage, experience the same amount of usability problems. If, the indication that skill level is key, when recruiting test persons for usability evaluations, this means that the most important is to recruit test persons of all skill levels of the target user group for the website or application, and, that variety in age or gender is not important when recruiting test persons. The implications might especially be of interest, when developing websites or applications for large heterogeneous user groups e.g. public websites or self-service applications, as these types of sites are targeted for all citizens in a country. This will make it challenging to represent all types of users when conducting usability evaluations, as a lot of test persons would need to be recruited, and it would be costly to conduct this amount of usability evaluations. On the contrary, if test persons only need to be recruited in regards to their skill level of Internet usage and IT in general, this would reduce the cost considerably.

6 Conclusion

This paper presents a study of to what extent different test persons, who are all experienced internet users, experience different types of usability problems. This has been presented across age, gender, and educational background or job function. The results are interesting as it is indicated that the usability problems experienced by users with a high level of internet experience do not vary significantly, across gender, age or background. This means that finding test persons might not have to be balanced in regards to neither gender or age, but that is more important to find test persons on all levels of internet experience in the target user group. Our results also indicate that people with an education in Computer Science do not experience significantly fewer usability problems, than other experienced internet users.

6.1 Limitations

We do recognise that further studies need to be conducted to be able to actually draw conclusions across user groups at different levels of Internet experience and that these results do not provide enough evidence to definitively rejecting the previously mentioned criticism of the "homogeneity assumption" by Law and Hvannberg [6]. This means that further research should be conducted with more homogeneous user groups with different levels of internet skills, and not just one group of experienced users. As it needs to be investigated further if these results also are valid for other user groups with lower skill levels in regards to Internet usage.

We also recognise the limitations of our test persons having a higher educational background and a self-reported high expertise in internet usage. Also the fact that a lot of the found usability problems were discarded at the coding phase and therefore not included in the data analysis.

References

1. Bruun, A., Stage, J.: An empirical study of the effects of three think-aloud protocols on identification of usability problems. In: Abascal, J., et al. (eds.) INTERACT 2015. LNCS, vol. 9297, pp. 159–176. Springer, Heidelberg (2015)
2. Caulton, D.A.: Relaxing the homogeneity assumption in usability testing. Behav. Inf. Technol. **20**(1), 1–7 (2001)
3. Hertzum, M., Jacobsen, N.E.: The evaluator effect: a chilling fact about usability evaluation methods. Int. J. Hum. Comput. Interac. **15**, 183–204 (2003)
4. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: The 10 ± 2 rule. Commun. ACM **53**(5), 130–133 (2010)
5. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
6. Law, E.L-C., Hvannberg, E.: Analysis of combinatorial user effect in international usability test. In: Proceedings of CHI (2004)
7. Lewis, J.R.: Sample sizes for usability studies: Additional considerations. Hum. Factors **36**, 368–378 (1994)
8. Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J., Stenild, S.: It's worth the hassle! the added value of evaluating the usability of mobile systems in the field. In: Proceedings of NordiCHI. ACM Press (2006)
9. Schmettow, M.: Sample size in usability studies. Commun. ACM **55**(4), 64–70 (2012)
10. Strauss, A., Corbin, J.: Grounded theory methodology. Handbook of qualitative research (1994)
11. Virzi, R.A.: Refining the test phase of usability evaluation: how many subjects is enough? Hum. Factors **34**, 457–468 (1992)