Scenarios in the Heuristic Evaluation of Mobile Devices: Emphasizing the Context of Use

Jari Varsaluoma

Tampere University of Technology, Human-Centered Technology,
Korkeakoulunkatu 6, P.O. Box 589, FI-33101 Tampere, Finland
jari.varsaluoma@tut.fi

Abstract. Varying contexts of use make the usability studies of mobile devices difficult. The existing evaluation methods, such as Heuristic Evaluation (HE), must be redesigned in order to create more awareness of the mobile context. Through the reworking of existing heuristics and use of written use scenarios, there have already been some promising results. In this study the context of use of mobile devices was examined with written scenarios. The main target was to improve the reliability of HE by increasing the number of right predictions and reducing the number of false positives produced by the evaluators. The results seem to differ from those of a previously conducted study as the scenarios did not improve the HE regarding the numbers of false positives or accurate predictions. There is a need for more research regarding the possible benefits of different scenarios and other factors that affect the outcomes of HE.

Keywords: Heuristic evaluation, scenario, context of use, mobile device, false positive.

1 Introduction

The rapid evolution of mobile devices is constantly bringing new challenges for usability experts. Traditional evaluation methods which were mainly designed for desktop computers cannot take into account all the factors of mobility and varying contexts of use [1]. This includes Heuristic Evaluation (HE) which is a very popular evaluation method in the industry [2]. In HE a usability expert studies the product in order to find usability issues on the basis of a list of usability guidelines. Perhaps the best-known heuristics are those of Nielsen [3]. HE is considered to be quick, cheap and easy to learn [3, 4].

The use of written scenarios during the HE of a mobile device can help evaluators to become aware of the context of use [5]. However, one must note the false positives that may occur during HE. A great number of false positives reduces the reliability of the evaluation method as time and effort are wasted on correcting problems that would not actually affect the end user. The reliability of the evaluation method can be measured by verifying the predicted problems through usability testing (UT) with real users [6].

M. Kurosu (Ed.): Human Centered Design, HCII 2009, LNCS 5619, pp. 332–341, 2009.

In this study scenarios were used in order to provide contextual information for the evaluators. A better knowledge of the context of use of a mobile device is assumed to reduce the number of false positives produced by the evaluators and improve the reliability of the HE.

2 Challenges in the Heuristic Evaluation of Mobile Devices

HE has been criticized for its relatively weak ability to find usability problems and predict their actual scope and severity [5, 6, 7, 8]. Experienced evaluators place themselves in the position of an inexperienced or an expert user. In psychology it is a common opinion that introspection is not an objective method [9]. This suggests that evaluation methods based on introspection are not reliable.

An evaluator using HE, or any other evaluation method, has to bear in mind the multiple factors that can affect the outcome of the evaluation. Table 1 presents factors that have been shown to affect the results of HE.

Table 1. Empirical studies have shown that the factors presented here can affect the outcome of heuristic evaluation. Note that the target of using scenarios or a realistic environment is actually to increase the evaluator's knowledge of the context of use.

Evaluation method	Evaluator characteristics	System evaluated and its context
Heuristics used (heuristics for mobile computing) [1] Number of evaluators [10, 11] Scenarios [5] Time spent on evaluation [7]	Education or job experience in usability area [3, 11] Experience with the application domain [1, 3, 5, 11] Experience with the heuristics used [1] Knowledge of the context of use (user, environment) [5, 6]	Evaluation environment (laboratory vs. real context) [5]

Different heuristics (Nielsen vs. Gerhardt-Powals) and media used for reporting usability problems have been studied, but these did not have significant effects on the results [7]. The importance of evaluators' understanding of the heuristics used and the method itself has been discussed [6, 8]. Still, in order to maintain the fast and inexpensive implementation of the method, it is not yet clear what manner of training would be the most appropriate. The criterion for judging evaluators as experts or novices in the usability area also lacks a definition [12].

The main challenge in the usability evaluation of mobile devices is their dynamic context of use. Important context types include location, identity, time, and activity [13]. As it is not possible to cover all possible use situations during the evaluation, one must choose the ones relevant for the study.

3 Scenarios in the Heuristic Evaluation of the Mobile Device

A study by Po's et al. [5] proposed a method called Heuristic Walkthrough (HW) that combined scenarios of use with Heuristic Evaluation (HE). They created five scenarios which were located around a university campus. A handheld pocket PC was evaluated by four usability experts using the HW method and four using the HE method. Scenarios seemed to help evaluators to shift their viewpoint from a technical evaluation towards the user's point of view and also predict more critical usability problems. What was not studied was the validity of the predicted usability problems. In order to truly measure the reliability of the method, the predicted problems should be verified with real users in usability tests (UT) [6, 7].

In this study scenarios were first used in UTs with users and then by usability experts during the HE of a mobile device. The results were compared in order to verify correctly predicted problems and false positives produced by the evaluators. The detailed setting of the study is described below.

3.1 Setting for the Study

The mobile device chosen for the study was a Nokia N95 mobile phone using the Series 60 operating system [14]. Four different scenarios were designed for a university context that was familiar to the author. The scenarios were situated in different locations around the campus and the UT was carried out in these locations. The first scenario happened in a usability lab and concentrated on teaching the basic features of the N95. The second scenario was located outside in a parking lot, the third in a noisy canteen, and the fourth in a walkway along the lakeside. The last scenario involved walking while using the device. Tasks varied from simple phone calls and text messages to tasks requiring multitasking and the use of a camera, map and GPS.

Here is an example from Scenario 3: *You are Johanna, and you are 25 years old. You study at the department of computer science and information systems. It is Friday afternoon and you are in the university canteen. There is a lot of noise as people are having lunch. You are waiting for your sister to come from a lecture because you are both going to go to your friend Kati's housewarming party in Tampere. While waiting, you ponder the next week's program and remember that you have a meeting with your thesis supervisor next Monday. You decide to create an appointment in your phone's calendar.*

The study consisted of a usability test (UT) and two heuristic evaluations (H1 and H2, the latter with scenarios) as shown in Fig.1. The UT was conducted first in order to avoid the knowledge of the HE results guiding the author's observations during the UT. The UT was carried out with the users in realistic use contexts and written scenarios were used at the scene. Ten university students aged 20-28 (M=23.7; SD=2.5) participated of whom six were female. None had previous experience of using the N95, but six participants had used mobile phones with Series 40 or 60 operating systems.

The HE sessions were conducted by two groups of four usability experts. The groups were created so as to be similar on the basis of a preliminary questionnaire. The experience of HE and mobile devices and knowledge of Nielsen's heuristics were investigated. The participants were researchers from universities and usability specialists from IT companies. Everyone had experience of at least three expert evaluations and knew Nielsen's heuristics very well or quite well.

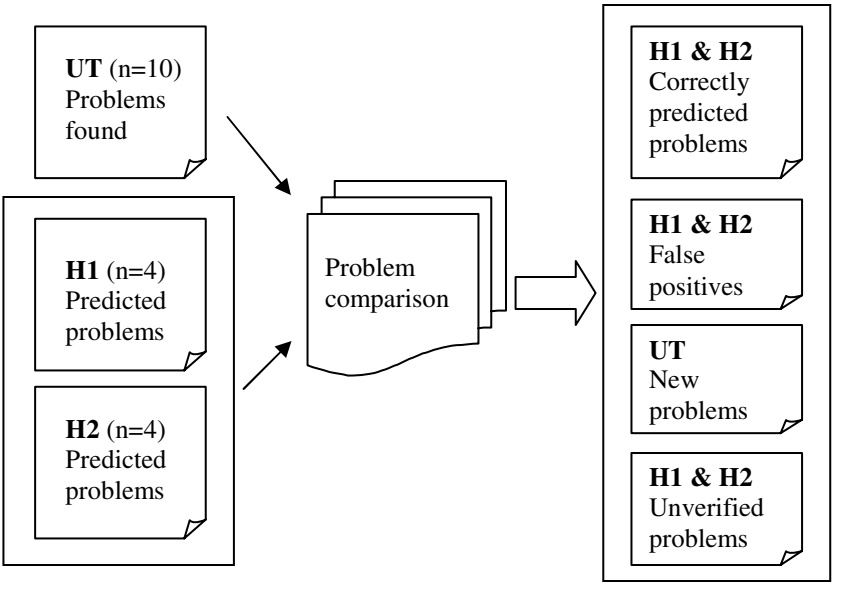

Fig. 1. The research structure. The heuristic evaluation (H1) and the heuristic evaluation with scenarios (H2) were carried out after the usability test (UT) results were analyzed. Problem comparison revealed false positives and usability problems that had been correctly predicted, had not been predicted by the evaluators (UT New problems), or were not in the scope of the tasks in the UT (H1 & H2 Unverified problems).

The first group (H1) acted as a control group and was given a simple list of features to evaluate, while the scenario group (H2) used the same written scenarios as the students in the UT. The N95 features that were evaluated were the same in both groups.

The resulting lists of predicted usability problems from the HE groups were compared to the observed problems list from the UT. Correctly predicted problems, false positives and problems found only in the UT were counted. Predictions related to N95 features that were not part of the UT could not be verified as true or false and were separated from the results as unverified problems.

3.2 Progress of the Study

Usability test with scenarios. The pretest questionnaire for choosing the participants was conducted via email. The test session started in a lab room with introductions and a questionnaire about previous experience with Series 40 or 60 phones. After a short briefing the first scenario was presented on paper. The participant was asked to think aloud during the test and was not assisted unless helpless for several minutes. An assistant videotaped the session and the author was responsible for all interaction with the user. A video feed from the N95 screen was captured by screen recording software (Fig. 2) and was sent wirelessly to a laptop that the user carried in a backpack (Fig. 3).

Fig. 2. Usability test session in a university canteen (Scenario 3). Video from the camera and the N95 screen shown on the left were merged afterwards to help the analysis of the detected usability problems.

Fig. 3. Scenario 4 included walking by the lakeside. When it was raining, the author held an umbrella for the participant. Cold weather made the usability testing challenging as users had to wear gloves while using the device.

Scenarios 2 and 4 were carried out outdoors and read aloud by the author so that the user's hands remained free to use the device (Fig. 3). At other times participants read the tasks by themselves (Fig. 2). After the last scenario the participant answered a short questionnaire about the usability of the device. The answers were talked through together and the participant was rewarded for her/his contribution.

The test videos were carefully studied by the author and all the mistakes, problems and comments were listed. Usability problems were graded as slips and problems. Slips were made by mistake (e.g. pressing the wrong button while writing) and were quick to fix (a few seconds). Other events were graded as problems. These included getting lost in menus, negative comments, and long times needed for consideration. It would have required each user to watch their own performance afterwards to judge if pressing a wrong button was a matter of the small keyboard or the result of an error of thought. Using any subjective problem ratings would also have required several researchers and discussions to have been more reliable. As this was not possible at the

time, greater emphasis was placed on objectively counting the numbers of similar problems between different users.

Heuristic evaluation with scenarios. A preliminary questionnaire was administered via email to choose usability experts for both study groups. Instructions were sent at least 2 days before the HE session. The material included a list of Nielsen's ten heuristics [3] and an example of a problem reporting form. For each predicted problem the evaluator had to select the heuristics that had been violated and the severity ratings in two different scales (a 1-4 scale and a fourfold table) on the basis of those presented by Nielsen [15].

The HE session started with a questionnaire about previous experience with similar operating systems. The problem reporting form was explained in detail. Next, the scenarios or a list of features were presented to the evaluator, depending on the evaluation group. The evaluators were asked to report any usability problems they would predict that both new and experienced users would have. The author answered any questions about the N95 functions during the evaluation session. Each evaluation finished within the allocated 3 hours.

3.3 Results

The UT revealed 290 observations of problems. These were categorized as 65 different usability problems of which 10 counted as slips and 55 as problems. 36 of the total of 65 were rated as *unique* as they were experienced with only one user.

Table 2. Results from the heuristic evaluations after the predicted problems were verified by comparing them to the usability test results

	Evaluator	Correctly predicted problems	False positives	Unverified predictions	Total
H1	1	10	1	3	14
Control	2	5	2	3	10
group	3	8	1	2	11
	4	7	4	0	11
Total		30	8	8	46
Mean		7.5	2	2	
SD		2.08	1.41	1.41	
H2	5	5	5	3	13
Scenario	6	7	7	1	15
group	7	9	0	1	10
	8	4	2	2	8
Total		25	14	7	46
Mean		6.25	3.5	1.75	
SD		2.22	3.11	0.96	

The HE control group H1 reported 51 problems (M=12.75; SD=1.71) and the scenario group H2 reported 47 (M=11.75; SD=2.87). The difference was not significant (t=0.599; p>0.05). When the predicted problems were analyzed, it was noticed that a single problem reporting form could actually contain problems from several of the UT problem categories. In these cases the reported problem was divided into several predicted problems with the same severity ratings and violated heuristics. There were also 9 cases where one evaluator had reported several problems that would fit into only one of the predefined UT problem categories. For example, the problems that users had when turning on the main camera had been categorized as one problem. For these predictions the mean value was counted for the severity ratings and all the violated heuristics were included. This process was repeated for each evaluator. After the analysis there were 46 predicted problems in both groups. 15 predictions could not be verified, because they were not within the scope of the UT scenarios. These predictions included N95 features that were not used by any of the UT participants. The results after the verification of the problems are presented in Table 2.

H1 outperformed H2 but the difference between the groups was not significant in terms of correct predictions (t=0.822; p>0.05) or false positives (t=-0.878; p>0.05).

Correct predictions in the HE were studied without the problems that were rated as unique or slips in the UT. The remaining problems occurred more frequently with different users and can be considered more severe. Table 3 shows that H1 made more

Table 3. Correct predictions from the heuristic evaluation, excluding problems that were rated as unique or slips in the usability test

	Evaluator	Correctly predicted without the unique problems	Correctly predicted without the unique problems and slips
H1	1	7	7
Control	2	5	5
group	3	6	5
	4	6	3
Total		23	20
Mean		5.75	5
SD		0.96	1.63
H2	5	3	3
Scenario	6	4	4
group	7	5	4
	8	4	4
Total		16	15
Mean		4	3.75
SD		0.82	0.5

correct predictions. Without the unique problems the difference was significant ($t=2.782$; $p=0.032$). Without the unique problems and slips there was no significant difference according to the Mann-Whitney test ($Z=-1.348$; $p>0.05$).

35 problems from the UT were not predicted in the HE. Only two of these were closely related to the environment: bright sunlight made the phone screen difficult to read and the keypad was troublesome to use in cold weather with gloves on. Perhaps the scenarios did not include enough contextual cues for the evaluators to predict these problems.

Problems with the Reliability of the Heuristic Evaluation. The evaluator effect [10] could be seen in the HE results. The chosen severity ratings varied greatly between the evaluators. Only one problem was rated as critical (4) by one evaluator and another evaluator rated the same problem as small (2). None of the evaluators chose the same set of violated heuristics for the same problem. Three problems were reported without heuristics being chosen, which suggests that the heuristics used did not cover all the predicted problem types.

The control group H1 predicted 10 problems correctly (of which 3 were rated as slips) that were not reported in the scenario group H2. H2 predicted 7 problems correctly that were not reported in H1. This may be due to the evaluator effect or the scenarios might have affected the way the evaluators examined the system.

4 out of 8 false positives in H1 and 9 out of 14 in H2 were somehow related to underestimating the user's skills. It seems that the evaluators had difficulties in estimating how much the user's previous experience affects the use situation.

4 Discussion

The results of this study suggest that using written scenarios during HE does not increase the reliability of the method. The scenarios did not increase the number of accurate predictions or reduce the number of false positives. The scenarios might have had some effect on the way the evaluators examined the device, as different problems were found between the groups. It would have required more evaluators to do the evaluation to be certain that the difference was not just because of the evaluator effect [10].

During the HE some evaluators would have liked to know more about the users' previous experience with the mobile device than the scenarios described. In this study the evaluators had to consider both new and experienced users, which made the evaluation more demanding. Concentrating either on inexperienced or experienced users and providing more information about the users' skills and experience with the device being evaluated might have reduced the evaluators' mental load and led to fewer false positives.

A simple list of features to evaluate seemed to produce more reliable results than the longer descriptions of the use situations. The scenarios probably made the evaluation sessions more complex and as a result the differences between the evaluators and the number of false positives were greater in the scenario group. It is surprising that when the scenarios limited the user group to university students, the evaluators in the control group still performed better. Perhaps the chosen user group was too general and scenarios would provide more help when evaluating for more specialized user groups, such as children or aged people.

The scenarios did not improve the HE in this setting in the way they did in the earlier study [5]. This may be because of the differences in scenarios, test devices, and analytical methods. Perhaps the scenarios used in this study failed to give enough contextual cues for the evaluators.

What would be the best way to provide contextual cues for evaluators and how much information is needed? If scenarios are used, should they be in the form of text, pictures, cartoons, video, or a combination of some of these? How can the relevant information for mobile devices be chosen and how can it be made sure that this information is valid? If planning realistic scenarios requires a great deal of information to be gathered, then does this vitiate the use of HE as a rapid and inexpensive method?

Another study with a greater number of evaluators should bring more information about the merits of using scenarios with HE. Analyzing and discussing the collected UT and HE data with other usability experts would also provide more reliable results. Information is also needed on the ways that evaluators actually use the given scenarios during the evaluation session and what the best way is to present contextual information.

For HE there are still some factors whose effects on the results would be interesting to study. Does the evaluator's motivation or spryness during the evaluation affect the results? How about evaluator's personal knowledge of the user group? What about the data that are stored in the mobile device during the evaluation? Should these represent the same data as the users have when using the device? Does it matter if the system being evaluated is used for leisure or some dangerous work where lives could depend on its usability? And what if the system being evaluated is still a low-quality prototype or one that is ready for the market?

There are still plenty of questions waiting to be answered about the reliability of evaluation methods. As HE is one of the most widely used evaluation methods, improvements to its reliability are worth pursuing.

References

1. Bertini, E., Gabrielli, S., Kimani, S.: Appropriating and Assessing Heuristics for Mobile Computing. In: Proceedings of the working conference on Advanced visual interfaces, pp. 119–126. ACM Press, New York (2006)
2. Rosenbaum, S., Rohn, A.J., Humburg, J.: A Toolkit for Strategic Usability: Results from Workshops, Panels, and Surveys. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, pp. 337–344. ACM Press, New York (2000)
3. Nielsen, J., Mack, R.L.: Usability inspection methods. John Wiley & Sons, Inc., Chichester (1994)
4. Law, E.L., Hvannberg, E.T.: Complementarity and convergence of heuristic evaluation and usability test: A case study of UNIVERSAL brokerage platform. In: Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 71–80. ACM Press, New York (2002)
5. Po, S., Howard, S., Vetere, F., Skov, B.M.: Heuristic Evaluation and Mobile Usability: Bridging the Realism Gap. In: MobileHCI 2004, pp. 49–60. Springer, Heidelberg (2004)

6. Cockton, G., Woolrych, A.: Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation. In: Joint Proceedings of HCI 2001 and IHM 2001: People and Computers XV, pp. 171–192 (2001)
7. Hvannberg, E.T., Law, E.L., Lárusdóttir, M.K.: Heuristic evaluation: Comparing ways of finding and reporting usability problems. Interacting with Computers 19(2), 225–240 (2007)
8. Law, E.L., Hvannberg, E.T.: Analysis of Strategies for Improving and Estimating the Effectiveness of Heuristic Evaluation. In: Proceedings of the third Nordic conference on Human-computer interaction. ACM International Conference Proceeding Series, vol. 82, pp. 241–250. ACM Press, New York (2004)
9. Tavris, C., Wade, C.: Psychology in perspective. Prentice-Hall, Inc., Englewood Cliffs (2001)
10. Hertzum, M., Jacobsen, N.E.: The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction 13(4), 421–443 (2001)
11. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 373–380. ACM Press, New York (1992)
12. Chattratichart, J., Lindgaard, G.: A comparative evaluation of heuristic-based usability inspection methods. In: CHI 2008 extended abstracts on Human factors in computing systems, pp. 2213–2220. ACM Press, New York (2008)
13. Dey, A.K.: Providing Architectural Support for Building Context-Aware Applications. Georgia Institute of Technology (2000)
14. S60 website, `http://www.s60.com/life`
15. Nielsen, J.: Usability Engineering. Academic Press, London (1993)