Empirical Comparison of Task Completion Time between Mobile Phone Models with Matched Interaction Sequences

Shunsuke Suzuki[1], Yusuke Nakao[2], Toshiyuki Asahi[1], Victoria Bellotti[3], Nick Yee[3], and Shin'ichi Fukuzumi[2]

[1] NEC Corporation, Common Platform Software Research Laboratories, 8916-47, Takayama-Cho, Ikoma, Nara 630-0101, Japan
{s-suzuki@cb,t-asahi@bx}.jp.nec.com
[2] NEC Corporation, Common Platform Software Research Laboratories, 2-11-5, Shibaura, Minato-ku, Tokyo 108-8557, Japan
{y-nakao@bp,s-fukuzumi@aj}.jp.nec.com
[3] Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA
{bellotti,nyee}@parc.com

Abstract. CogTool is a predictive evaluation tool for user interfaces. We wanted to apply CogTool to an evaluation of two mobile phones, but, at the time of writing, CogTool lacks the necessary (modeling baseline) observed human performance data to allow it to make accurate predictions about mobile phone use. To address this problem, we needed to collect performance data from both novice users' and expert users' interactions to plug into CogTool. Whilst novice users for a phone are easy to recruit, in order to obtain observed data on expert users' performance, we had to recruit owners of our two target mobile phone models as participants. Unfortunately, it proved to be hard to find enough owners of each target phone model. Therefore we asked if multiple similar models that had matched interaction sequences could be treated as the same model from the point of view of expert performance characteristics. In this paper, we report an empirical experimental exercise to answer this question. We compared identical target task completion time for experts across two groups of similar models. Because we found significant differences in some of the task completion times within one group of models, we would argue that it is not generally advisable to consider multiple phone models as equivalent for the purpose of obtaining observed data for predictive modeling.

Keywords: Cognitive Model, CogTool, Evaluation, Human Centered Design, Human Interface, Mobile Phone, Systematization, Usability Test.

1 Introduction

Usability evaluation should be performed in the early phase of a product development process [1]. In addition, commercial enterprises demand that evaluations do not incur high costs. To satisfy these requirements, some of the authors of this paper have been

J.A. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2009, LNCS 5612, pp. 114–122, 2009.

developing systematized evaluation methods and tools that can be applied early and economically [2].

CogTool [3] is a user interface evaluation tool for predicting task execution process and task completion time, using a given interface. In CogTool, a user model based on ACT-R cognitive architecture [4] mimics execution of a task by using graphical specification data extracted from frames of a storyboard for the task, which is input into CogTool in advance.

CogTool offers a low-cost evaluation approach for the early part of a product development process. Just a sketch as the storyboard, which need not be functionally implemented, is enough for evaluation with CogTool. This small requirement allows us to evaluate the user interface early, cutting the cost of developing the system in which the user interface works. As a computational user model, not an actual human, executes tasks in CogTool, costs such as recruiting, organizing and paying participants and use of a usability lab are avoided.

To apply CogTool to an evaluation for a new system, it is necessary to refine the user model to improve the accuracy of its predictions, using observed data of actual experts' and novice's task execution (observed performance data). In this refinement, we planned to incorporate observed performance data into CogTool's user model and then compare its predictions with additional observed data [5]. The user model in CogTool can represent a novice who explores how they should interact with the target system, or an expert who can quickly execute the most efficient interaction sequences.

In order to collect enough observed data to both incorporate in CogTool and to compare with its predictions, we needed to recruit a considerable number (approximately twenty) of experts, who had owned the specific model of the product to be evaluated for longer than two months. This research was part of an effort that also included comparing CogTool's predictions in mobile phone evaluation to subjective user impressions in order to see if there were any correlations between these two different evaluation approaches.

2 Challenge of Recruiting Owners of Specific Models

Recruiting owners of specific mobile phone models is very hard because the numbers of owners of a given model are low, due to the fact that new models are released frequently. Also, recruiting is expensive because "Owner of a specific mobile phone model" is a stricter qualifying condition for a recruiter than general conditions such as age and gender. Even if cost is not an obstacle, the recruiting may take a long time. Of course, an alternative way to make the recruiting easier would have been to reduce the number of participants. However we wanted to keep the required number (twenty) because we would like to modify the user model in high accuracy and to analyze the correlation between completion time and subjective impression statistically.

Thus, it was clear that it would be quicker and cheaper to find owners of several similar related models that have matched interaction sequences for target tasks than only owners of one specific model. We defined a "matched interaction sequence" as the same sequence of key presses required to complete a given task.

As mentioned above, our main objective in this research was a planned comparison between CogTool model predictions and observed user performance data. Our planned method was to capture the duration of each interaction event on the mobile phone by analyzing video frames and recording real user key presses [5] as they perform task execution steps in the same order as is specified for the CogTool model (thus excluding idiosyncratic user performance). This protocol was what drove the demand that all mobile phone models for the observed data have to have matched interaction sequences.

In this paper, we report on a preliminary procedure that was conducted prior to our main observed user performance data collection effort. This was an empirical validation to clarify whether we could treat multiple mobile phone models with matched interaction sequences as equivalent for the purposes of predictive modeling.

3 Experiment

This section describes our experiment in which we collected user task completion times (not individual key presses as is planned for our future study) across phone models in two groups (A and B), each defined by its members being a target phone that we wished to evaluate with CogTool or having matched interaction sequences to the target phone. This meant that within each group, the user interfaces were similar and tasks could be executed using exactly the same steps across models. The main difference between the models was simply their physical form factor (they had equivalent but differently sized and spaced keys).

Participants executed a set of the same tasks with the same key press sequences across all the models in the group. After collecting data on participant performance, we compared the mean of the completion time for each task between the phone models. We explain this method in more detail below.

3.1 Mobile Phone Models

We defined 2 mobile phone model groups. In Group A, there was the N905i, which was a target model for CogTool, and the N905iμ, which has the same interaction sequences as the N905i (Fig.1). In Group B, there was W61CA, another target model for CogTool, and the W61H and W53H, which had same interaction sequences as the W61CA (Fig.2). The models in each group have a matched key layout. For example, both N905i and N905iμ have a mail key above a main menu key. However, size, form, depth to press, and distance of keys varies by model. Because Fitts' law [6] [7], which was used in CogTool, was the logarithm of key distance / key size, we supposed that a small gap of the distance or the size between the models would not affect the time to reach from a key to another.

In this experiment, we selected only tasks with matched interaction sequences on the phone model within the group. For instance, in the case where a user has to select a target item to go to the next frame in a task, if the number of items above the target is different between the models, the number of key presses is also different with them. This difference means that their interaction sequences are not matched. Therefore, we

Fig. 1. N905i (left) and N905iμ (right) in Group A

Fig. 2. W61CA (left), W61H (center), and W53H (right) in Group B

did not use either these models or tasks in the experiment. If the difference in number of items did not affect the interaction sequence (e.g., the items which are below the target item), the models and the task were usable in this experiment. Also, in cases where displayed labels of the items except the target one differed between the models, we used the models and the tasks in this experiment, because the participants trained to be experts, who already knew the interaction sequences for the task, could find which item they should select without comparing its label with other items'.

3.2 Participants

20 participants (16 males and 4 females, age: 20-40s) for Group A and 24 participants (18 males and 6 females, age: 20-40s) for Group B took part in this experiment. We did not select participants based on prior experience with specific mobile phone models. Instead, we provided all participants with extensive time to learn specific target tasks on specific phone models as described in 3.6 Learning.

Table 1. Task list for Group A

	Content	The number of key presses
Task 1	After inputting 000-3454-1111, store the number in a phonebook. Enter the name “Nihondenki” in Kana letter.	39
Task 2	Set a schedule named “Nomikai” in Kana letter from 19:00 to 23:00 tomorrow.	43
Task 3	Turn on/off “the auto keypad lock after folded”	24
Task 4	Check the newest mail sent to Aoki-san in the “Friend” folder.	8
Task 5	Take a picture soon after launching a camera. Then save the picture in the “Camera” folder in the “My Picture” folder.	11

Table 2. Task list for Group B

	Content	The number of key presses
Task 1	After inputting 000-3454-1111, store the number in a phonebook. Enter the name “Nihondenki” in Kana letter.	39
Task 2	Set a schedule named “Nomikai” in Kana letter from 19:00 to 23:00 tomorrow.	40
Task 3	Check the newest mail sent to Aoki-san in the “Friend” folder.	6
Task 4	Take a picture soon after launching a camera. Then save the picture as an idle screen, with a sub menu.	13

3.3 Tasks

We used 5 tasks for Group A and 4 tasks for Group B. The tasks are listed in Table 1 and Table 2. They are common mobile phone functions. At the same time, we selected tasks with various numbers of key presses. Although there were alternative interaction sequences for each task, we instructed all participants to use the same sequence for each task in this experiment.

3.4 Task Assignment to the Participants

For Group A, we assigned two of the five tasks (see Table 1) to each participant. Thus, for each task in Group A, we had data from eight participants. For Group B, we

again assigned two of the four tasks (see Table 2) to each participant. Thus, for each task in Group B, we had data from twelve participants. Each participant completed these two tasks across the phone models within the group.

3.5 The Number of Trials for Data Collection

During the study, each participant repeated each task five times. We will refer to this portion of each task as the "main trial". In addition, when participants switched from one phone model to the next, they performed two practice trials before the main trials for each task. These practice trials are not included in the data analysis.

3.6 Learning

We set aside time for participants to learn assigned interaction sequences. In the learning phase, they executed the tasks assigned to each of them, with the assigned interaction sequences. In this experiment, we had to compare data generated by experts, because the observed data required for refinement of CogTool also needed to come from experts (as well as novice users). Another purpose of the practice was to reduce variance in completion time, because the more trials a person does, the smaller the gap of the completion time between the previous trial and the next one, along a general learning curve [8]. In these purposes, we set as many practice trials as possible so that participants were able to learn the interaction and develop as much expertise as possible.

The practice for both models took place before the main trial part, making the participant's learning level for each model more similar since we would expect transfer effects. If the order of the experiment had been "*1. practice with N905i, 2. main trial with N905i, 3. practice with N905iμ, 4. main trial with N905iμ*", the participant's learning levels for N905iμ could have been higher than for N905i because the participants benefit from experience with the phone they use first so that they would have far more experience during *4. main trial with N905iμ* than when they did *2. main trial with N905i*. Therefore, we set the order as, "*1. practice with N905i, 2. practice with N905iμ, 3. main trial with N905i, 4. main trial with N905iμ*". By setting the order alternately for each task, we avoided a large gap between the learning levels for each model.

The numbers of practice for each task were 46 times (23 times per a phone model) in Group A and 36 times (13 times per a phone model) in Group B. The numbers were dictated by a practical concern that the entire session for each participant should be completed in 90 minutes to avoid participant's fatigue. For example, in Group A, there are all 120 trials; 20 main trials for 2 tasks (5 trials × 2 models × 2 tasks), 8 trials to get used to the model when the model switching (2 trials × 2 models × 2 tasks), and 92 practice trials (46 trials × 2 models × 2 tasks). If it takes maximum 45 seconds to execute 1 task, it takes 90 minutes to execute all 120 trials (45 seconds × 120 trials = 5600 seconds).

4 Results

We conducted a one-way repeated-measures ANOVA for each task with phone model as the factor. In Group A, there was no effect of phone model in any of the tasks (p's $> .16$), see Figure 3. In Group B, we found two significant differences. There was

a significant effect of phone model in task 2 ($F[2,22] = 5.86$, $p < .01$) and task 4 ($F[2,22] = 18.72$, $p < .01$), see Figure 4. The other two tasks in Group B were not significant (p's > .11). Post-hoc comparisons showed that in task 2, W53H was significantly different from the other two models (p's < .05). And in task 4, W61CA was significantly different from the other two models (p's < .05).

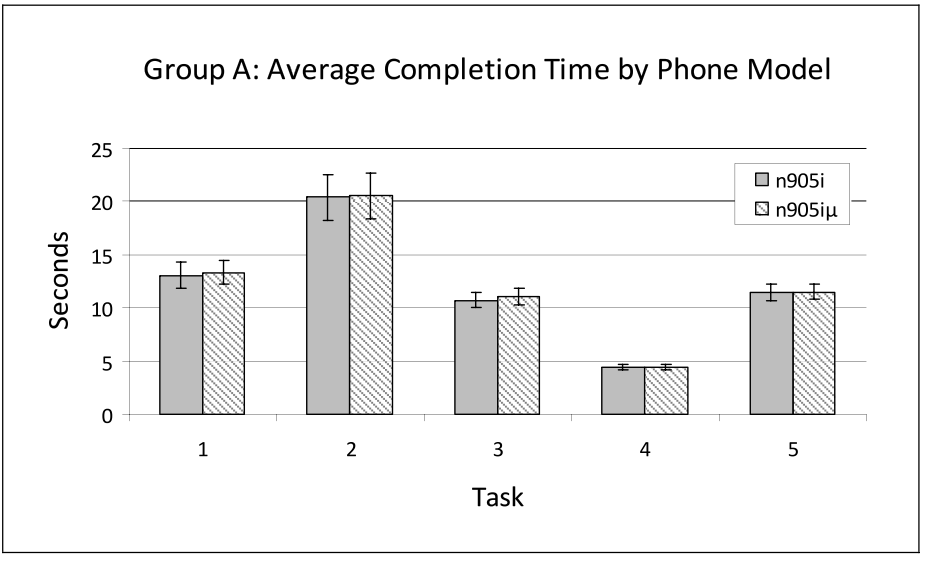

Fig. 3. Average completion time for each task with each model in Group A

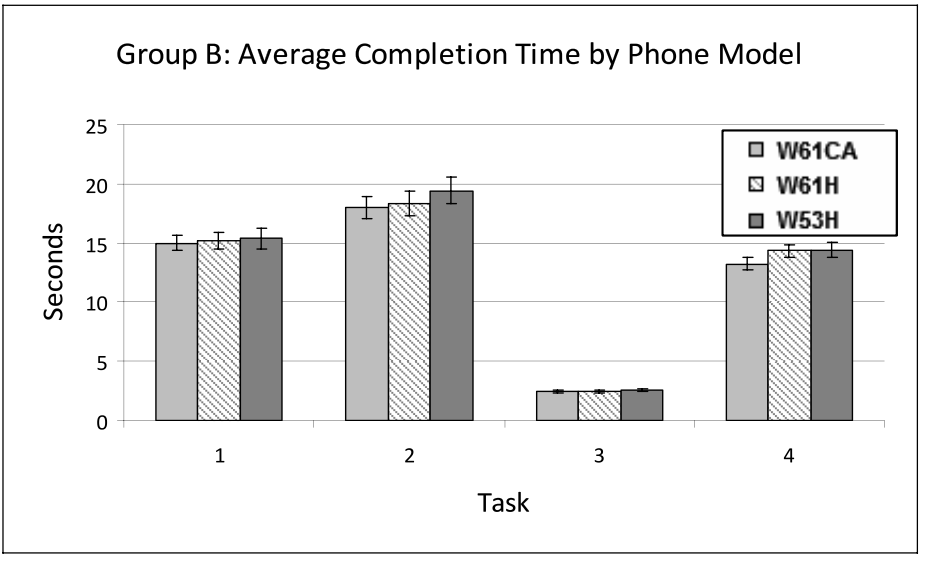

Fig. 4. Average completion time for each task with each model in Group B

5 Discussion

Based on the results discussed above, for our planned data collection exercise to gather mobile phone interaction performance data to incorporate into CogTool, we will use only one model as the target model. Thus, even though it is likely to be more time consuming and difficult, we should recruit only owners of, and do our evaluations on, only the specific target phone model that we plan to model with CogTool, even though the recruiting cost is more expensive.

One possible concern with the study design was that we had 12 participants for each task in Group B, but only 8 participants for each task in Group A. Thus, it may be the case that we only found significant differences in Group B because we had more statistical power from the larger sample size. To examine this concern, we reanalyzed the data from Group B with 4 participants randomly removed from each task. We found that both tasks were still significant at $p < .05$. This suggests that the difference in sample size alone is not why we found significant differences in Group B but not Group A.

One of possible reasons of the significant difference between phones is the physical characteristics of the keys because many of the participants commented that these characteristics had affected their subjective performance. For example, some of the participants commented that flat keys had been difficult to distinguish from adjacent keys because of the lack of tactile cues. Others commented that keys with deeper key press feel made it easier to distinguish multiple repeated key presses using the tactile sense. Actually, W61H and W53H have flatter and shallower keys than W61CA has.

6 Conclusion

The study suggests that we should not consider multiple mobile phone models with matched interaction sequences as equivalent to the same model, because we found significant differences in the mean task completion time between the models in Group B. Even though we found no significant differences between the two models in Group A, the findings from Group B suggest that a more conservative approach overall in using only one model may be warranted for developing cognitive models to minimize potential noise from usage variations across phone models.

In Group B, there were one or two tasks with a significant difference in completion time between the models even though only four tasks out of a total of 10 target tasks were executed in this experiment. Based on the differences found in this preliminary study, we expected that it would be hard to find 10 tasks that did have matched interaction sequences but did not exhibit significant differences in completion time that would be suitable for our planned main objective to collect valid observed data on which to base modeling of mobile phone interaction. With more tasks, more participants and more trials in the main study, we would expect the number of significant differences between models to increase and make our observed data less reliable.

As mentioned in the Discussion section based on the participants' comments, we expect one of possible reasons of the significant difference between phones is the

difference in tactile key press sensation due to hardware differences between different phone models.

References

1. Nielsen, J.: The Usability Engineering Life Cycle. Computer 25(3), 12–22 (1992)
2. Bellotti, V., Fukuzumi, S., Asahi, T., Suzuki, S.: User-Centered Design and Evaluation - The Big Picture. In: Proceedings of Human Computer Interaction International. Springer, Heidelberg (to appear, 2009)
3. John, B.E., Prevas, K., Salvucci, D.D., Koedinger, K.: Predictive Human Performance Modeling Made Easy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004, pp. 455–462. ACM, New York (2004)
4. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036–1060 (2004)
5. Teo, L., John, B.E.: Comparisons of Keystroke-Level Model predictions to observed data. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 1421–1426. ACM, New York (2006)
6. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47, 381–391 (1954)
7. Fitts, P.M., Peterson, J.R.: Information capacity of discrete motor responses. Journal of Experimental Psychology 67, 103–112 (1964)
8. Newell, A., Rosenbloom, P.S.: Mechanisms of skill acquisition and the law of practice. In: Rosenbloom, P.S., Laird, J.E., Newell, A. (eds.) The Soar papers. Research on integrated intelligence, vol. 1, pp. 81–135. MIT Press, Cambridge (1993)