

A Comparative Analysis of Usability Evaluation Methods on Their Versatility in the Face of Diversified User Input Methods

Daiju Ishikawa, Takashi Kato^(✉), and Chigusa Kita

Graduate School of Informatics, Kansai University, 2-1-1 Ryozenjicho,
Takatsuki, Osaka 569-1095, Japan
tkato@kansai-u.ac.jp

Abstract. Every command consists of an action and an object, suggesting that a usability problem can occur whenever the user is unable to identify an appropriate action and/or the object associated with his/her current goal. The recent shift from mouse-based to touch-based interaction demands that any usability evaluation method be sensitive to not only object-related but also action-related usability problems. This study involved a total of 32 participants, four kinds of tasks differing in the difficulty of identifying objects and executing actions, and four qualitative methods of usability evaluation. Analyses of sets of observation data with concurrent and retrospective protocol by the same participant and interpretive protocol by a new participant indicate that while the oral instruction method seems least appropriate, the newly-devised narration method seems to have better prospects than the observation and the think aloud method for the usability evaluation of touch-based interaction.

Keywords: Usability evaluation method · Qualitative method · Touch-based interaction · Mouse-based interaction

1 Introduction

Although a variety of interactive devices and applications are now available, it remains unchanged that almost every command consists of an action and an object, causing usability problems to occur whenever the user is unable to identify an appropriate action and/or the object associated with his/her current goal. The present study was prompted by the recent shift from mouse-based to touch-based interaction, which demands re-focusing on the ease of specifying and executing required actions. In mouse-based interaction with the Web, for example, required actions are so simple (i.e., dragging and clicking) that most usability problems concern the ease of identifying appropriate objects. In touch-based interaction with the Web, however, various actions or gestures are available, some of which may not be so obvious to the user and/or may require rather precise execution. This suggests that any usability evaluation method for touch-based interaction need be sensitive to not only object-related but also action-related usability problems.

Involving a total of 32 participants, and four kinds of tasks that differed in the difficulty of identifying objects and executing actions, the present study examined the

effectiveness of four qualitative methods of usability evaluation. One particular focus was on the ability to not only identify both object-related and action-related user errors but also elicit verbal protocol that can help clarify the reasons or causes of such errors. Another focus was on the ability to control the cognitive load that might be placed on both participants and researchers in running evaluation studies. We believe that reducing the cognitive load is important in order to increase, both quantity- and quality-wise, spontaneous elicitation of verbal protocol.

2 The Usability Evaluation Methods Compared

Four usability evaluation methods compared were all qualitative methods that were designed to yield both observation and verbal protocol data. They were modified versions of the observation method, the think aloud protocol method [1], and the oral instruction protocol method [4], and a newly-devised, narration protocol method. There were two groups of four participants each for each of the four methods.

Except in the observation method, one group of four participants was asked to yield a particular type of verbal protocol specified by the method as they worked on assigned tasks (concurrent protocol, or CP) using a tablet device. The CP procedures of the three protocol methods are described below. One common feature was that the instruction was given at the start of the session and the experimenter basically refrained from intervening the participant's work during the session.

Having completed the tasks, while watching the video recordings of their own performance, the participants were asked to describe their interaction and recall their intentions at that time (retrospective protocol, or RP). They were told not to hesitate to repeat what might have been said in CP. The RP instruction was similar to, or probably less restrictive than, that used in other studies [2, 8].

Without performing any tasks themselves, another group of four participants attempted to describe what the person in the video was trying to accomplish and why (interpretive protocol, or IP). The participants were provided with the same tablet device, however, and were completely free to work on it as they deemed it necessary. IP is similar in its intent to the collegial protocol obtained from professionals describing recorded performance of their colleagues [3]. The videos were played for RP and IP without audios to avoid the effects of CP contents on the elicitation of RP and IP.

- Observation method with RP and IP

The observation method was included as a method that could allow observation of more natural interaction between the user and the system, given the limitation that the usability testing in this study was conducted in an artificial, laboratory room. Participants were first asked to complete four assigned tasks without any additional requirement, such as CP, or any specific instructions on how to work on the tasks. They were later asked to provide RP for their own performance, for which IP was in turn obtained from a different participant.

- Think aloud method with CP, RP, and IP

Participants were asked to verbalize what they were thinking as they performed the assigned tasks. Prompts to encourage verbalization when the participants remained

silent were intentionally kept less frequent than in a standard think aloud method [1] to avoid otherwise increased stress and anxiety on the part of the participants. RP and IP were obtained in the same way as above.

- Narration method with CP, RP, and IP

The narration method involved a pair of a participant and an observer sitting next to the participant. The participant was asked to describe to the observer what he or she was thinking about the task, focusing particularly on the evaluation of the current state and the specification of the next goal or intention. The narration method, obviously not entirely new, was devised with the same intent of the question-asking protocol method [5] or more broadly the coaching method [6] as an alternative to the think aloud method. The main purpose was to alleviate the task demands placed on the participant by the requirement of monologue-type, real-time verbalization. The expectation was that participants would find it easier to talk to someone actually there rather than to engage in continuous, overt monologue [7]. It would also be easier for them to verbalize intentions and execute intended actions in sequence rather than to simultaneously verbalize their thinking and execute actions. The narrations elicited by the participants were treated as CP, and RP and IP were obtained in the same way as in the other methods.

- Oral instruction method with CP, RP, and IP

The oral instruction method involved a pair of a participant and an operator. The participant was to give requests or instructions orally to the operator regarding what and how he or she would like the operator to perform on his/her behalf [4]. The participant was asked to provide as much clear and detailed instructions as possible and such oral instructions were treated as CP. The operator was actually a member of the research team and tried to be a “faithful” operator, who neither inferred the participant’s intention nor performed anything unspecified in the instruction. The operator was to ask for clarification whenever the participant’s instruction was not clear or specific enough. RP was provided by the participant and IP by a new participant in the same way as in the other methods.

3 Method

3.1 Participants

A total of 32 participants, 31 undergraduate students (14 males and 17 females) and one recent graduate (male), were recruited on the conditions that they were smartphone users but that they had no or little experience of using tablet devices. They were assigned to one of the eight conditions with four participants each.

3.2 Tasks and Equipment

Four kinds of tasks were devised that differed in the difficulty of identifying objects associated with goals and in the variety of actions available for participants to apply to goal-related objects. All tasks were performed using a tablet device (iPad 2 with iOS

7.0.3) in which Safari and Sketches were installed for the Web navigation and sketching tasks described below.

- **Simple Objects/Single Action**
Using the Web browser, the participant was to find their university library regulation regarding the maximum number of books that can be loaned. The action needed was only that of tapping the target link and the sequence of links to be followed was short with each link being easily identifiable on each page.
- **Complex Objects/Single Action**
Using the Web browser, the participant was to find the opening hours of one of their university cafeterias. The action needed was again only that of tapping the target link. However, the to-be-followed sequence of links was more complex and the correct links were more difficult to identify on the pages, due partly to the less straightforward mapping between the link names and the target information.
- **Simple Objects/Multiple Actions**
The task was to group application icons on the home screen into one folder and vice versa. The objects for this task were application icons and the home button, which should not be difficult to identify. However, multiple and various actions were needed to complete the task.
- **Complex Objects/Multiple Actions**
The tasks were to draw a map and save it in the photo library using Sketches and to close all the applications that remained open in the background. While these tasks demanded precise execution of a variety of actions, identifying target objects seemed more difficult, partly because explicit cues were not available for some parts of the tasks.

3.3 Procedure

Using the iPad 2, a group of 16 participants carried out the four tasks described above under one of the four evaluation method conditions. There were four participants in each method condition. Using the Latin square method, the order of the tasks was counterbalanced among four participants in each condition. Having completed the tasks, the participants engaged in the RP task. A different group of 16 participants performed the IP task with each participant randomly assigned a video of a particular participant in the other group. All sessions were video recorded, which captured the entire tablet and touchscreen operations along with all utterances made by the participant and the experimenter.

4 Results and Discussion

We first compiled usability problems encountered by any one of the 16 participants in any one of the four tasks. For each identified usability problem, we then checked to see whether or not CP (except in the observation method), RP, and/or IP were provided by any participant. Based on the compiled data, we constructed a problem-by-participant

table to obtain an overview showing which evaluation method was relatively successful in identifying usability problems and obtaining related verbal protocol data. Although the number of individual cases in each method condition was small, some interesting patterns are still visible in the table, which we discuss in terms of the strengths and weaknesses of the four evaluation methods.

The oral instruction method, previously shown to be effective in identifying action-related as well as object-related usability problems [4], was least successful particularly in detecting usability problems related to more complex actions. One might think that such usability problems did not surface simply because all the actions were carried out by the experimenter on the participant's behalf. Further analysis based on one participant's CP, however, reveals a more interesting picture. Although the experimenter was careful not to infer the participant's intention or proceed beyond what was verbally requested, when it came to executing the requested action, he somehow did it right. That is, the oral instruction method may be less effective in detecting potential difficulty associated with execution of a correct action. This drawback may not be compensated for by additional verbal protocols such as RP and IP. Unless a given problem is initially pointed out in CP such that a correctly performed action by an experimenter is deviated from a participant's intention, it is next to impossible for the participant or a new participant to realize the presence of the problem afterwards. Another interesting observation was that the participants were more successful than those in the other method conditions in identifying correct links in the Web tasks, probably because the requirement of giving explicit instructions made them more attentive to the overall information on the page before giving a specific instruction. It seems that the oral instruction method is likely to underestimate usability problems concerning the execution aspect of complex actions and those caused by less careful but more natural interaction behavior on the part of the user.

Contrary to our expectation that RP and IP could supplement the lack of CP in the observation method, few IP data were obtained across the tasks, which was also the case in the other three methods. Evidently, interpreting someone's interaction behavior without one's own experience is much harder than expected for ordinary users. The amount of RP was not satisfactory either, implying that possible causes of usability problems basically would have to be inferred from observed, overt interaction behavior. However, our hunch is that RP could be increased with the experimenter's directive prompts pointing to not only overt but also potential usability problems. The observation method had one advantage over the other three methods such that the participants tended to proceed with the tasks further than those participants in the other methods, probably because they were able to better concentrate on the tasks in the absence of mandate verbalization and/or interaction with the experimenter. One might want to use the observation method to explore potential usability problems as far as possible within a given time and then to seek RP, using directive prompts, to clarify the reasons or causes behind those usability problems.

There were not major differences between the think aloud and the narration method with respect to the ability to identify usability problems. One notable difference, however, was in the variability of the amount of CP among the participants. The individual differences were much greater in the think aloud than in the narration method, partly because we did not prompt participants to verbalize as frequently as in a

standard think aloud procedure, and partly because verbalization in the narration method was perceived more natural or less artificial than that in the think aloud method. While the observation method may need to be supplemented with RP, which could double the time and cost of usability testing, the narration method can be effective without RP and may be less susceptible to variability in verbalization among prospective participants.

Acknowledgments. This research was supported by JSPS KAKENHI Grant Number 25510018. Daiju Ishikawa is currently at Marubeni Information Systems Co., Ltd.

References

1. Carroll, J.M., Mack, R.L.: Learning to use a word processor: by doing, by thinking, and by knowing. In: Thomas, J.C., Schneider, M.L. (eds.) *Human Factors in Computer Systems*, pp. 13–51. Ablex, Norwood (1984)
2. Elling, S., Lentz, L., de Jong, M.: Retrospective think-aloud method: using eye movements as an extra cue for participants' verbalizations. In: *CHI 2011 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1161–1170. ACM, New York (2011)
3. Erlandsson, M., Jansson, A.: Verbal reports and domain-specific knowledge: a comparison between collegial and retrospective verbalisation. *Cogn. Technol. Work* **15**, 239–254 (2013)
4. Hori, M., Kihara, Y., Kato, T.: Investigation of indirect oral operation method for think aloud usability testing. In: Kurosu, M. (ed.) *HCD 2011. LNCS*, vol. 6776, pp. 38–46. Springer, Heidelberg (2011)
5. Kato, T.: What “question-asking protocols” can say about the user interface. *Int. J. Man Mach. Stud.* **25**, 659–673 (1986)
6. Mack, R.L., Robinson, J.B.: When novices elicit knowledge: question asking in designing, evaluating, and learning to use software. In: Hoffman, R.R. (ed.) *The Psychology of Expertise: Cognitive Research and Empirical AI*, pp. 245–268. Springer, New York (1992)
7. Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S., Ashenfelter, K.T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: *CHI 2010 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2381–2390. ACM, New York (2010)
8. Van den Haak, M.J., de Jong, M.D.T., Schellens, P.J.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behav. Inf. Technol.* **22**, 339–351 (2003)