Concise paper

A Critical Comparison of Usability Testing Methodologies

Paul Mason
Bay of Plenty Polytechnic
Tauranga, New Zealand
paul.mason@boppoly.ac.nz

Dr. Beryl Plimmer
University of Auckland
Auckland, New Zealand

ABSTRACT

Usability testing is critical to system success. A system, which is functionally perfect, will fail if it is too difficult to use. A recent New Zealand example is the $26 million Department of Corrections jail database system that has been reported as being so difficult to operate that staff do not use it (Boyes, 2005). There is a plethora of usability testing methodologies, yet little information on the differences between them. This paper describes a case study where we usability tested an application with two common methodologies and compared the findings from each test. From this study, we suggest when each methodology may be more appropriate.

1. INTRODUCTION

Usability testing is defined as "a technique for ensuring that the intended users of a system can carry out the intended tasks efficiently, effectively and satisfactorily" (Gaffney, 1999). It is a critical component in system success, as if functional but unusable, a system will more than likely fail. This is highlighted by a recent New Zealand example: a jail database developed by the New Zealand Government has been under scrutiny as "the system is so difficult to operate, staff often don't use it, or don't know how to use it … the database may have to be dumped" (Berry, 2004). Another classic example is the nuclear accident at Three Mile Island in 1979 where the system alerted the operators but the interface was such that they could not properly diagnose the problem (United States United Regulatory Commission, 2004).

There are a number of usability testing methodologies suggested in the literature. Yet there is little to guide the software engineer as to which methodology is most suitable and the type of information each will provide. This paper critically compares the "Think-Aloud Protocol" and "Focus Groups", two dominant usability methodologies. A case study was undertaken on a Tablet PC with an application titled "Penmarked" (Plimmer & Mason, 2004) that provides a paperless environment for marking students' assignments. This is a demanding system to usability test because pen-based interaction is not yet well understood and the software is quite novel.

The structure of the remainder of this paper is as follows: next, the background provides a brief survey of usability testing methodologies. Section 3 describes the case study and section 4 discusses our findings. Finally, the conclusions suggest situations where each methodology may be more appropriate.

2. BACKGROUND

Usability testing requires *real* users to participate in the studies using the software. There are a number of different usability testing methodologies suggested in the literature: we briefly describe the most popular here and refer readers to Preece *et al.* (2002), or Holzinger (2005) for a more thorough explanation.

First, a number of methodologies are based on observing users verbalising their experience while using the software, for example think-aloud, co-discovery and question-asking protocols (Preece *et al*., 2002; Holzinger, 2004; Hom, 1998). Because these methodologies provide very rich feedback, small samples can give accurate results. Analytical methodologies that examine raw data include task performance measurements, keystroke analysis, eye tracking and post-task questionnaires (Preece *et al*., 2002; Hom, 1998). Focus groups are a method of post-task investigation that provide qualitative data. Finally, field-testing is the observation of users

in their normal environment (Preece *et al*., 2002; Holzinger, 2004).

We chose to compare the think-aloud protocol and focus groups, two of the most popular methodologies in this project; the think-aloud protocol because of its accuracy with small sample sizes and focus groups because of its time and cost efficiency.

2.1 THINK-ALOUD PROTOCOL

The think-aloud protocol is arguably the most popular usability testing methodology. Here, the user performs a series of tasks and he/she is expected to verbalise every thought, feeling or opinion that comes to mind. This enables the tester to understand how the user views the system (Holzinger, 2004; Hom, 1998; Preece *et al*., 2002). Usually during a session, an observer takes notes and uses video recording for further reference.

The literature suggests the advantage of this technique is that the tester is able to discern the user's experience. The type of data that can be expected includes preferences, problems and misconceptions; additionally, performance data can be collected simultaneously.

The disadvantage of this method is that speaking every feeling feels unnatural for the user. In addition, cognitive overload may occur when the task or the interface requires a high level of concentration, and the user may "forget" to verbalise his/her thoughts. It may be difficult to apply rigorous performance measurement because of delays while talking (Holzinger, 2004; Preece *et al*., 2002). Think-aloud is time consuming for the tester who needs to brief the user, observe sessions and transcribe recorded results. (Holzinger, 2004; Preece *et al*., 2002)

2.2 FOCUS GROUPS

A focus group involves the users in a directed small-group discussion after they have used the system. The tester will often guide the discussion to particular aspects of the user experience. The advantage of focus groups is that they are time and cost efficient for the tester and group discussions often "spark" ideas. However, they require the users to get together and may not encapsulate everybody's opinion as some people feel intimidated in a group situation. A group, if poorly run, can be dominated by an individual. There is also a possibility, due to the delayed nature of the focus group, that the users forget the details of the user experience (Preece *et al*., 2002).

3. CASE STUDY

The application chosen for usability testing is a system developed by the authors, Penmarked (2004). It is assignment-marking software developed for the Tablet PC. It bridges the gap between traditional and electronic marking systems by allowing digital ink annotation onto electronic documents whilst retaining the benefits of paperless electronic systems where marks are only recorded once. This software is an early prototype of a new class of application and utilises new hardware. This newness provides an ideal environment to compare usability testing methodologies as there is minimal previous research to guide the development.

The hardware used was a standard Tablet PC that supports pen input affording a 'piece of paper' paradigm. Penmarked consists of three main frames: an assignment display and annotation frame, a mark schedule, and a student list. With this software a marker can open an assignment, annotate it with digital ink, and record scores with digital ink or the keyboard. Workflow support provides: loading of a list of assignments, rapid transition between students and export of annotated assignments as PDF files and export of grades in an xml format.

Markers grading computer programs for a computer science course at the University of Auckland were the subjects for this study. Five markers were available to participate; each marked approximately 30 students' assignments.

3.1 THINK-ALOUD PROTOCOL

The think-aloud study followed the standard protocol that required the markers to verbalise their thoughts whilst working with the software. One of the authors observed each session, noting relevant interactions. The sessions were also video taped. We use a statistically small sample size, but this should be sufficient given the richness of think-aloud data (Holzinger, 2004).

We found that persuading the markers to talk was extremely difficult; much worse than anticipated. They appeared to be concentrating on the task and seemed to find talking a distraction. Often when they did speak, it was about the program they were marking, not Penmarked.

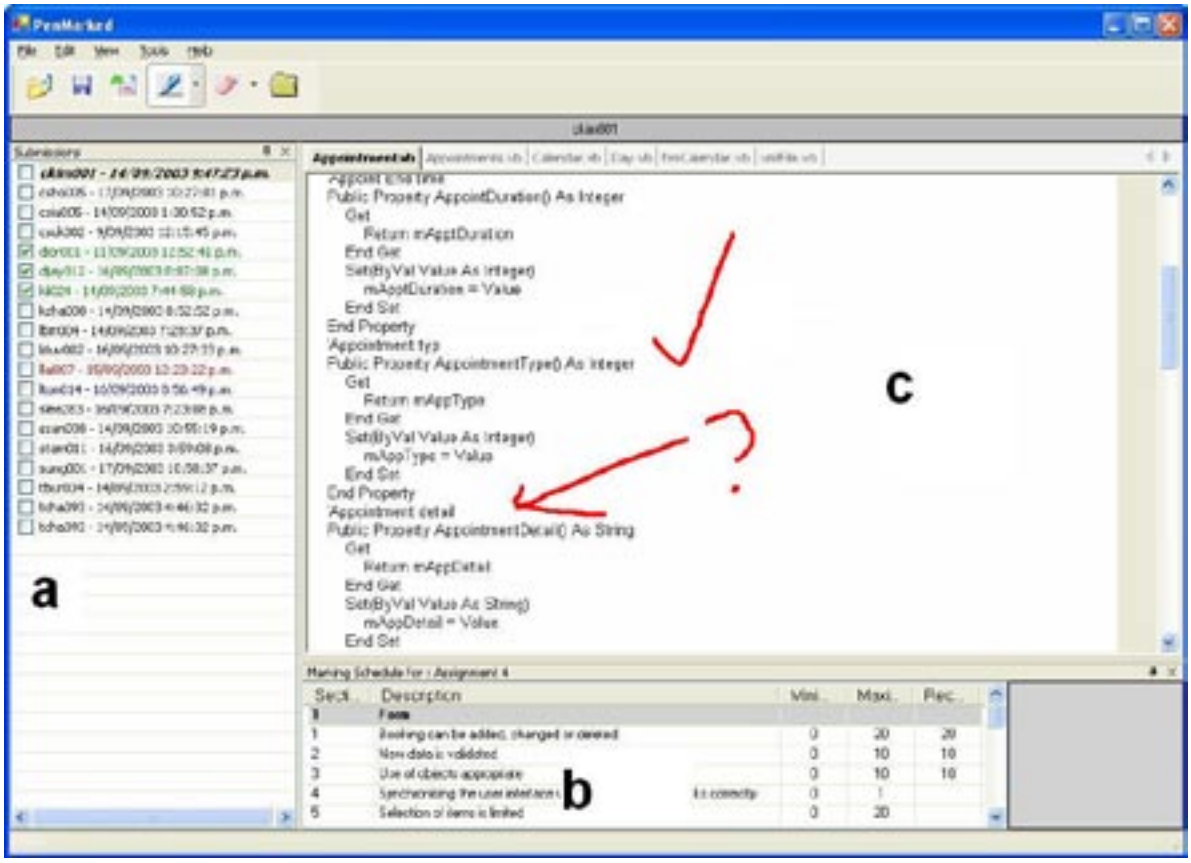

Figure 1. A screenshot of Penmarked showing the student list (a), the mark schedule (b) and the annotation frame (c).

We attribute this to cognitive overload due to the cognitive demands of program review (Robins *et al*., 2003).

This study exposed a number of usability problems. However, taking notes and analysing the recorded material was time consuming and necessitated observing repetitive behaviour. Most significant points were noted during the observation rather than analysis, therefore further studies could rely on observation. We found that the majority of problems were recognised during the first half hour of each session. A total of 20 hours were spent on this study.

The results gained from the think-aloud study were predominately identification of practical bugs and interaction inconsistencies rather than opinions on the system, or suggestions for improvements.

3.2 FOCUS GROUPS

The same group of markers were brought together in a focus group after marking was finished. The discussion centred on problems encountered and suggestions for improvements. Particular questions were put to the group on: interface navigation; the tablet hardware; software bugs; general difficulties; general likes.

Finding a good time for everyone to meet was a bigger problem than expected. During the first five minutes the group was quiet, however, the discussion quickly developed and most appeared to be comfortable talking and conveying their opinions and feelings. One person participated little, voicing his opinion only a couple of times. This lack of involvement was later identified as disinterest as opposed to shyness.

Many interesting comments were made during the 45 minutes the group spent together. Some, but not all, of the problems identified during the think-aloud test were cited. However most of the discussion was on higher-level aspects of the interface such as improving the workflow support.

4. DISCUSSION

The case study exposed strengths and weaknesses of both the think-aloud protocol and focus groups. The think-aloud protocol was reasonably successful at identifying practical problems, with bugs and interaction problems being the major

findings. Most of the results for the think-aloud protocol were found during the first half-hour of each observation. The primary reason for this was the highly repetitive nature of marking assignments. However, towards the end of a session the user seemed more comfortable to offer suggestions and general comments regarding the program.

The main problem with the think-aloud protocol was getting users to verbalise. Our task (marking programs) is cognitively demanding and this could explain the difficulty. Yet, in addition to this, the participants said that they felt intimidated knowing that they were being observed and videoed. Constant reminders were required for the user to speak and their responses tended to be shallow and often about the program being marked as opposed to Penmarked. The think-aloud protocol was time-consuming for the testers. The video camera was essentially a duplication of the written notes and apart from providing a reference for rechecking, was unneeded for analysis.

The focus group provided different information consisting largely of suggestions, appraisals, and feelings. Most comments were suggestions for enhancements to particular features. A disadvantage of this method is that the users adapt quickly to the idiosyncrasies of an interface and problems can be forgotten before the discussion. Organising a time for the focus group was more difficult than anticipated.

The focus group provided rich results in about 45 minutes. This is a minor commitment for the results obtained. Users' actively participated during the focus group and appeared to be comfortable to voice their opinions in the small group environment.

The actual findings obtained from each method were quite different, with minimal overlap. This suggests that using an observation technique as well as a "reflection" technique may be essential to gain a full spectrum of results.

5. CONCLUSION

This case study uncovered many important features of both the think-aloud protocol and focus group methodologies. The focus group took significantly less time. In addition, the focus group provided better suggestions for improvements; the think-aloud protocol tended to uncover practical details. For usability testing involving cognitively demanding activities, the think-aloud protocol may be inappropriate. Testing methods such as observation and post task inquiry may be a better alternative.

On the surface, it appears that focus groups may be better value, however, in terms of results, both contributed to an interesting analysis of the system. We suggest that the choice of methodology should depend on the goals of the particular study. Perhaps when a new, innovative system is under development early focus groups will guide the development towards the most important features for the user. Later observations and think-aloud provided feedback on important minutiae of the interaction experience.

Overall, using more than one methodology is highly recommended. One should not rely solely on an observation technique, nor only on a reflective technique. It would be wise to integrate both techniques, as the results found from each will most likely vary significantly.

ACKNOWLEDGEMENTS

The authors would like to extend a thank you to the students and markers that participated in this study.

REFERENCES

Berry, R. (2004, 11/09/2004). $30m jail database faces axe. The New Zealand Herald.

Boyes, N. (2005, 12/03/2005). Corrections computer system could absorb extra $40m. The New Zealand Herald.

Gaffney, G. (1999). Usability testing. Retrieved 5 March, 2005, from http://www.infodesign.com.au/ftp/UsabilityTesting.pdf

Holzinger, A. (2005). Usability engineering methods for software developers. Communications of the ACM, 48(1), 71.

Hom, J. (1998). The usability methods toolbox. Retrieved 5 March, 2005, from http://jthom.best.vwh.net/usability/

Nielsen, J., & Mack, R. L. (1994). Usability inspection methods. New York: Wiley.

Plimmer, B., & Mason, P. (2004). Designing an environment for annotating and grading student assignments. Paper presented at the OZCHI, Wollongong, Australia.

Preece, J., Rogers, Y., & Sharp, H. (2002). Interaction design: Beyond human-computer interaction. New York: J Wiley & Sons.

Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. Computer Science Education, 13(2), 137-172.

United States Nuclear Regulatory Commission (2004). The Accident at Three Mile Island. Retrieved 30 September, 2004, from http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.pdf